and it's due to climate change. Flood is predicted in several locations using some advanced technologies which just helps the people to be prepared for upcoming disasters. It is very difficult to create a predictive model using machine learning Machine learning gives computers the capability to learn without being explicitly programmed. Machine learning has a role in preventing many natural disasters like earthquakes, floods and many more. Machine learning make decisions using past data and these data are fed into the algorithms and the output is predicted. Machine learning (ML) can be classified into three categories Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. The Supervised learning can be again further divided into two types Regression and Classification. Regression algorithms are used if there is a relationship between the input variable and output variable. It is used for the prediction of continuous variables. Classification algorithms are used when the output variable is categorical which means there are two classes such as Yes- No, Male-Female, True-False. Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead models itself find the hidden patterns and insights from the given data. The Unsupervised learning algorithm can be further categorised into two types: Clustering and Association. Clustering is a method of grouping objects into clusters such that objects with most similarities remain into a group and have less or no similarities with the objects of another group. Association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database

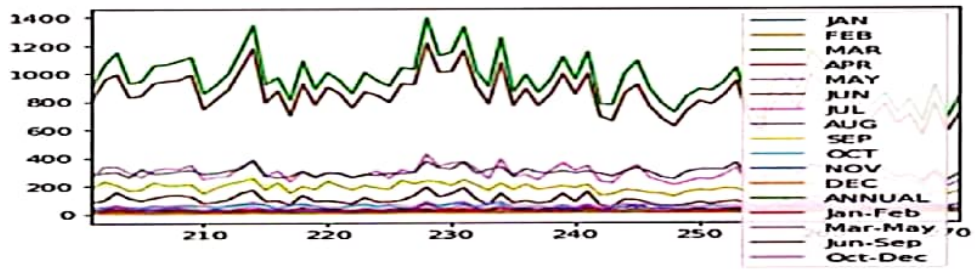


FIG 1.1 FLOOD PREDICTION OF EACH MONTH

A reason for the popularity of ML models is that they can numerically formulate the flood nonlinearity, solely based on historical data without requiring knowledge about the underlying physical processes. Data-driven prediction models using ML are promising tools as they are quicker to develop with minimal inputs. ML is a field of artificial intelligence (AI) used to induce regularities and patterns, providing easier implementation with low computation cost, as well as fast training, validation, testing, and evaluation, with high performance compared to physical models, and relatively less complexity. The continuous advancement of ML methods over the last two decades demonstrated their suitability for flood forecasting with an acceptable rate of outperforming conventional approaches. In comparison to traditional statistical models, ML models were used for prediction with greater accuracy. Many ML algorithms, e.g., artificial neural networks (ANNs) , neuro-fuzzy, support vector machine (SVM), and support vector regression (SVR), were reported as effective for both short-term and long-term flood forecasts. In addition, it was shown that the performance of ML could be improved through hybridization with other ML methods, soft computing techniques, numerical simulations, and/or physical models. Such applications provided more robust and efficient models that can effectively learn complex flood systems in an adaptive manner. Nonetheless, ML algorithms have important characteristics that need to be carefully taken into consideration. The first is that they are as good as their training, whereby the system learns the target task based on past data. If the data is scarce or does not cover varieties of the task, their learning falls short, and hence, they cannot perform well when they are put into work. The second aspect is the capability of each ML algorithm, which may vary across different types of tasks. This can also be called a "generalization problem", which indicates how well the trained system can predict cases it was not trained for, i.e., whether it can predict beyond the range of the training dataset. For example, some algorithms may perform well for short-term predictions, but not for long-term predictions. These characteristics of the algorithms need to be clarified with respect to the type and amount of available training data, and the type of prediction task, e.g., water level and streamflow. Here, we should note that surveys of ML models used for predictions of floods on sites where rain gauges or intelligent sensing systems are used.

1.2 INTODUCTION TO FLOOD PREDICTION TECHNIQUES OF ML

Floods are among the most destructive natural disasters and it causes lots of damage to property and human life. The yearly data shows that the amount of rainfall is increasing

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Among the natural disasters, floods are the most destructive, causing massive damage to human life, infrastructure, agriculture, and the socioeconomic system. Governments, therefore, are under pressure to develop reliable and accurate maps of flood risk areas and further plan for sustainable flood risk management focusing on prevention, protection, and preparedness. Flood prediction models are of significant importance for hazard assessment and extreme event management. Robust and accurate prediction contribute highly to water resource management strategies, policy suggestions and analysis, and further evacuation modeling. Thus, the importance of advanced systems for short-term and long-term prediction for flood and other hydrological events is strongly emphasized to alleviate damage. However, the prediction of flood lead time and occurrence location is fundamentally complex due to the dynamic nature of climate condition. Therefore, today's major flood prediction models are mainly data-specific and involve various simplified assumptions. Physical models showed great capabilities for predicting a diverse range of flooding scenarios, they often require various types of hydro-geomorphological monitoring datasets, requiring intensive computation, which prohibits short-term prediction .

FIG 1.0 FLOOD

multipurpose dam area in Magway region, Myanmar. It is based on the Markov Chain model and estimated flood input as reservoir level (RL) get from sensor combining with weather data from Google API's weather data.

2.9 PREDICTION OF FLOOD BY RAINF ALL USING MLP CLASSIFIER OF NEURAL NETWORK MODEL

Floods are powerful devastating natural hazards and upgrading them is a risky task. The system development of predicting the flood helps in risk reduction, policy recommendations, a reduction in human life loss, and a reduction in property harm, all of which are correlated with floods. During the last two decades, neural network approaches have made significant contributions to the enhancement of predicting the flood by including the dynamic mathematical equations of physical flood processes. This has resulted in improved performance and cost-effective solutions. To overcome these issues and forecast the floods based on rainfall neural networks technique are used. Dealing with variable creation, missing value treatments, data cleaning/preparation, exploratory review, and assessment was all part of the analysis process. Many algorithms are used for predicting the flood such as k-means clustering, support vector machine, MLP classifier. Among them MLP produces the good accuracy and displays output in Graphical user Interface.

2.10 PREDICTION OF FLOOD IN BANGLADESH USING K-NEAREST NEIGHBORS ALGORITHM

Bangladesh is a flood-prone country. With limited resources and a major portion of the population living below the poverty line, flood impacts are severe. Deaths, malnutrition, widespread diseases, damage to infrastructure, disruption in the economy are some of the after-effects of this cataclysm. In order to put a flood management system into effect, it is essential to predict flooding events ahead of time. In this work, we applied different correlation coefficients for feature selection and k-nearest neighbors (k-NN) algorithm for the prediction of flood. The detailed result analysis shows that we achieved a high testing accuracy of 94.91%, average precision of 92.00% and an average recall of 91.00% using the k-NN machine learning model.

2.6 BIG DATA ANALYTICS FOR FLOOD INFORMATION MANAGEMENT IN KELANTAN, MALAYSIA

Big data analytics is expected to be of useful approach to many aspects of problem solving. For this research, the authors are trying to utilize this facility for flood disaster management specifically for the state of Kelantan, Malaysia. This research is considering the ordinal type data obtained from the state authorities and proposing on data manipulation through statistical inferences and big data analytics. As a result, the research is expecting for an early warning system can be developed from this study. Nonetheless, the added value approach to the analytics carried out was also considering the method of semantic network and flood ontology to design its algorithms.

2.7 NNARX FLOOD PREDICTION MODEL USING "TRAINGD" AND "TRAINOSS" TRAINING FUNCTION

Research on flood disasters are popular area of researches and more and more advanced tools are uses to obtain reliable predicted results. This due to flood causes harm to people and their property specifically. Thus, many hydrological models are propose and developed by researchers around the world to forecast flood disaster ahead of time for evacuation purposes. However, it is very difficult to develop flood model because all physical parameters that represent the flood behaviour must be included in the modelling. Parameters such as the depth of river basin, river flow rate and sediment factors data are very difficult to get. Hence, this paper proposes flood prediction model using black-box model. It is call black-box model because the model is develop using only input and output data without the needs to know the physical parameters that contribute to flood disaster. Neural Network Autoregressive with Exogenous Inputs (NNARX) is one type of black-box model that is widely applied to solve the nonlinear problems.

2.8 FLOOD PREDICTION SYSTEM FOR MIDDLE REGION OF MYANMAR

The natural disaster are extremely hit on humans' society and people are unwanted to face it. However, people are not able to control the causing of disaster. Floods are one of the most affected disasters which heavily hit to people. People cannot elude from hit of flood but they can protect from it by predicting of floods. The flood prediction system is developed for Mone

based on hydrological and hydrodynamic models and research based on data driven models.

SWMM , Mike, and Storm are widely used hydrological and hydrodynamic models in flood prediction, among which the SWMM model developed by the U.S. Environmental Protection Agency is one of the most widely used models to simulate urban runoff and drainage [30], [31]. Zhou *et al.* [32] comprehensively considered the land use type, surface impermeability and drainage system to establish a SWMM model and estimated the flood volume and risk under urbanization and climate change. Zhu *et al.* [33] simulated the influence of different pavement structures (drainage surface, permeable pavement and permeable road) on reducing surface runoff by constructing a SWMM model. Kim and Cho [34] employed SWMM and a 2D surface model to simulate the inundation area and range of the city under 320 different rainfall situations.

2.5 SATELLITE FLOOD INUNDATION ASSESSMENT AND FORECAST

The SMAP FW data effectively captured surface water dynamics during the severe tropical cyclone Idai event, indicating potential utility for regional flood monitoring to inform disaster assessments. The regional inundation and soil moisture information acquired from SMAP was further combined with Landsat observations and GFS precipitation forecasts to establish a GEE-based machine-learning approach for effective regional flood forecasts. The resulting 1-day (24-h) FW forecast predictions were highly correlated ($R = 0.87$) with contemporaneous Landsat observations and showed relatively low errors (RMSE = 0.68%; nRMSE = 25.6%). A model feature importance analysis showed that timely satellite measurements of surface wetness over the study area are crucial for determining the 1-day forecast inundation extent from a rainfall-driven flood event, while the cumulative precipitation over a longer period and surface wetness information for the surrounding region become more important for longer (3-day) forecasts. The 1-day forecasts for the Idai event captured the flood inundation temporal dynamics and 30-m spatial pattern consistent with independent satellite SAR observations. The approach provides new capacity for global flood monitoring and forecasts from synergistic satellite observations, including data sparse regions of Africa.

2.3 IOT-ENABLED FLOOD PREDICTION MACHINE LEARNING MODULE

According to a recent research study by Gartner [1], 6.4 billion connected objects/things were identified in 2016, representing an increase of over 3% compared to 2015, and expected to reach 20.8 billion by 2020. Some of these 'things' include a variety of sensors that might be useful for improving the quality of data collected for the purpose of making better decisions. IoT is an increasingly growing topic and widely available for such purposes [2], [3]. It permits things to be controlled or sensed remotely across several network environments, providing an interface for direct control over the physical world [4]. To extract useful and effective data, ML offers an appealing method for predicting water levels, for example. The vast majority of environmental monitoring centers have adopted IoT to assist in environmental protection [5], [6].

Sun and Scanlon [7] indicate that the use of machine learning has significantly improved the detection of early flood warning using powerful deep learning algorithms. Panchal *et al.* [8] show that gait characteristics can be utilised to capture flood levels and used machine learning algorithms including support vector machine and random forest for the analysis of the data. While Furquim *et al.* [9] propose a flood detection system based on IoT, machine learning and Wireless Sensor Networks (WSNs) in which fault-tolerance was embedded in their system to anticipate any risk of communication breakdown. Belal *et al.* [10] indicate that the lack of information about the quality of drinking water and the difficulty of early prediction of the flood has inspired various researchers to monitor and detect flood. In this research, we use multi-sensor data that originate from monitoring flood centres located in different countries around the world to determine rivers' water levels. To this end, a variety of advanced predictive models and learning algorithms were developed (i.e., Artificial Neural Networks (ANN), Random Forest (RF), K-Nearest Neighbour classifier (KNN), Long-Short Term Memory (LSTM) and Support Vector Machine (SVM)). The aim is to utilize machine learning algorithms to analyse flood sensors log datasets, characterized by nonlinearities and dynamic characteristics.

2.4 REAL-TIME PREDICTION OF THE WATER ACCUMULATION

Urban flood prediction is an effective means to help urban flood management personnel reduce the losses caused by urban flood. Many scholars have done considerable research on the theory and technology of urban flood prediction in recent years . To the authors' best knowledge, there are two main types of urban flood forecasting research: research

2.2 MULTIAGENT SYSTEM FOR FLOOD MONITORING, PREDICTION, AND RESCUE

Flood is a natural disaster event that affects millions of people each year in Punjab province, Pakistan. It results in huge volumes of water, much more than the natural or artificial conveyance system (i.e. channels, canals, streams, rivers, dams, creeks, basins, culverts, and estuaries) can manage. In Punjab, climate change is the major cause of floods resulting in the rapid melting of snow in northern Himalayan mountains, and excessive monsoon rains. The flood damages can be reduced by modern-day technological and scientific advancements. Modern-day early warning systems can inform stakeholders in time, rescue system can inform stakeholders about rescue services (i.e. shelter, clean drinking water, cooked food, medical facilities, medicines, etc). In Punjab province, poverty results in poorly planned illegal human settlements in flood risk areas (i.e. floodplains adjacent to riverbanks, areas not protected by flood banks). These settlements result in high casualties and losses during floods.

In semi-arid and arid regions, the floods are also beneficial events, their benefits are:

1. Flood water is the only source of groundwater recharge. In cities and towns near the river the groundwater recharges, its quality improves, and its level rises. The aquifers are re-filled and their water quality improves.
2. Flood water is also the most important source of enrichment of agriculture lands with important minerals. This fertile soil yields better crops for the next few years, which is beneficial for the agriculture economy of Punjab.
3. In some regions of Punjab, there are run-off floodwater harvesting (i.e. harvesting crops in agriculture lands flooded by floodwater).
4. The floodwater also re-fills the fresh water ponds, lakes, dams, and water reservoirs.
5. South Punjab is among the hottest regions of the world. In these regions after a flood, water evaporation increases, which in turn causes rain. The overall weather becomes cooler and these rains are extremely beneficial for these hot, low-rain arid regions.

CHAPTER 2

LITERATURE SURVEY

2.1 MODEL INTEGRATED WITH ALGORITHMS FOR HOURLY WATER LEVEL PREDICTION

The establishment of reliable water level prediction models is vital for urban flood control and planning. In this paper, we develop hybrid models (GA-XG Boost and DE-XG Boost) that couple two evolutionary models, a genetic algorithm (GA) and a differential evolution (DE) algorithm, with the extreme gradient boosting (XG Boost) model for hourly water level prediction. The Jungrang urban basin located on the Han River, South Korea, was selected as a case study for the proposed models. Hourly rainfall and water level data were collected between 2003 and 2020 to construct and evaluate the performance of the selected models. To compare the prediction efficiency, two other tree-based models were chosen: classification and registration tree (CART) and random forest (RF) models. A comparison of the results showed that two hybrid models, GA-XGBoost and DE-XGBoost, outperformed RF and CART in the multistep-ahead prediction of water level, and the relative errors of the hybrid model ranged from [2.18%-9.21%], compared to [3.76%-10.41%] and [2.99%-11.88%] for the RF and CART, respectively. Reliable performance was also supported by other measures. In general, the GA-XGBoost and DE-XGBoost models displayed relatively similar performance despite their small differences. The CART model was not preferable for multistep-ahead water level predictions, even though it yielded the lowest Akaike information criterion (AIC) value. This study verifies that despite having some drawbacks when considering long step-ahead prediction and model complexity, hybrid XGBoost models might be superior to many existing models for hourly water level prediction.

PREDICTION RESULT GENRATION

CLASS 1: "FLASH FLOOD MAY OCCUR": If the given input is higher than 2400mm, then the flash flood class is set to 1.

CLASS 2: "FLASH FLOOD MAY NOT OCCUR": If the given input is lesser than 2400mm, then the flash flood class is set to 0.

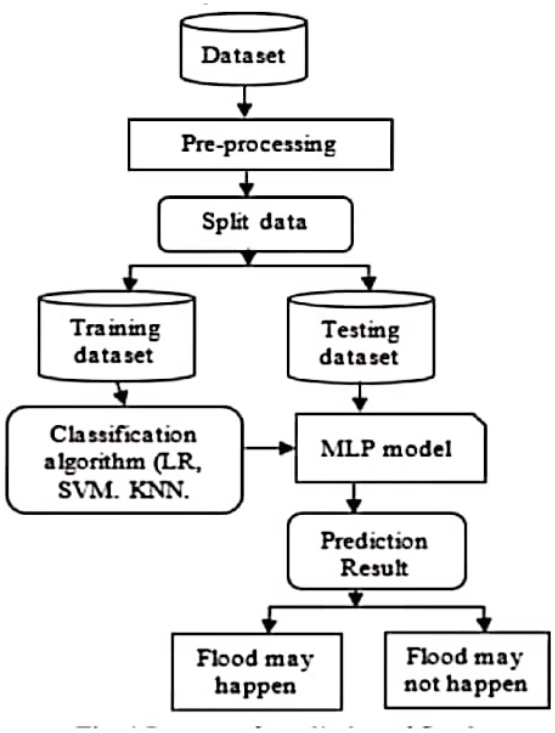


FIG 4.4 PREDICTION PROCESS OF FLOOD

4.2 MATERIALS AND METHODS

The dataset used for the analysis is an Kerala rainfall data, between the periods of 1901 to 2018. The dataset is created in CSV file format in month-wise and also sub-division of the state along with district. The unit used to measure the rainfall is in millimeter (mm). The dataset was collected from the metrological department of 36 locations on a monthly basis. ML method is useful to estimate the future by analyzing historical data like occurrences of flood in earlier times. The ML Performance are measured using algorithms like LR, SVM, KNN, and MLP.

DECISION TREE CLASSIFICATION ALGORITHM

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.

The logic behind the decision tree can be easily understood because it shows a tree-like structure. In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

Fitting K-NN classifier to the Training data:

Now we will fit the K-NN classifier to the training data. To do this we will import the

K-NeighborsClassifier class of **Sklearn Neighbors** library. After importing the class, we will create the Classifier object of the class. The Parameter of this class will be

N-neighbors: To define the required neighbors of the algorithm. Usually, it takes 5.

Metric='Minkowski': This is the default parameter and it decides the distance between the points.

P=2: It is equivalent to the standard Euclidean metric.

And then we will fit the classifier to the training data.

THE K-NN WORKING CAN BE EXPLAINED ON THE BASIS OF ALGORITHM:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

STEPS TO IMPLEMENT THE K-NN ALGORITHM:

1. Data Pre-processing step
2. Fitting the K-NN algorithm to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

K-NEAREST NEIGHBOR

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. K-Nearest Neighbor(KNN) Algorithm for Machine Learning Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. Hence, to label a new point, it looks at the closest labeled points to that new point and has those neighbors vote, so whichever label most of the neighbors have is the label for the new point. This algorithm makes predictions about the validation set using the whole training set. Only by rummaging through the whole training set to seek out the closest instances, the new instance is predicted. Closeness is a value that is determined using a proximity measurement across all the features involved.

LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Steps in Logistic Regression

To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

CHAPTER 6

SUMMARY AND CONCLUSION

This will lead to an insight into the information and preparedness requirements of local communities and the development of solutions adapted to the social realities.

Secondly, it will lead to a closer cooperation and coordination for flood predection and warning services of public institutions based on user needs.

Technically the project will focus on developing a collaborative platform that will link citizen, public authorities and other stakeholders and on enabling the public to be warned so that actions can be taken to reduce the adverse effects of the flood.

Damages that occur due to flash flood to living and non-living are very large. In this paper, flash flood prediction model is built. Indian rainfall data collected between the periods of 1901 to 2018 is used for analysis. The pre-processed rainfall data was split into 80% training data and 20% testing data. The dataset is trained with Support Vector Machine, Logistic regression, K-nearest neighbor, Decision Tree Algorithm and Random Forest . The performance factors like precision, recall, F1 score, sensitivity, specificity was calculated for each technique. Confusion matrix with TP, TN, FP and FN were calculated. The machine learning system ignited by data cleaning and processing, replacing or removing the null values, model building and evaluation. The classification accuracy achieved by and SVM is(Above 90%). Among the five techniques SVM performed with highest accuracy. The SVM flash flood prediction model predicts whether "flood may happen or not" based on the rainfall range for particular locations. This prediction model can be used by disaster management department to forecast flash flood. In future, we aim to use other artificial intelligence techniques to improve the prediction accuracy. The process can be automated by displaying the result of prediction in webpage or desktop application.

5.3 PERFORMANCE ANALYSIS OF ALGORITHMS

The following steps stated that the proposed model provides a very easy efficient method for predicting flood:

STEP1: The collected dataset of rainfall is pre-processed.

STEP2: The dataset of rainfall is randomly partitioned into testing and training.

STEP3: The dataset trained with LR, SVM, KNN, RF, and DT algorithm.

STEP4: The model is constructed using the SVM algorithm with the highest accuracy and validated with the parameters such as precision, f1-score, accuracy, and confusion matrix

STEP5: Input test data to the prediction model and validate the results.

Accuracy of the algorithm calculated from the F1 score and Confusion Matrix.

From (1) and (2), the precision measured.

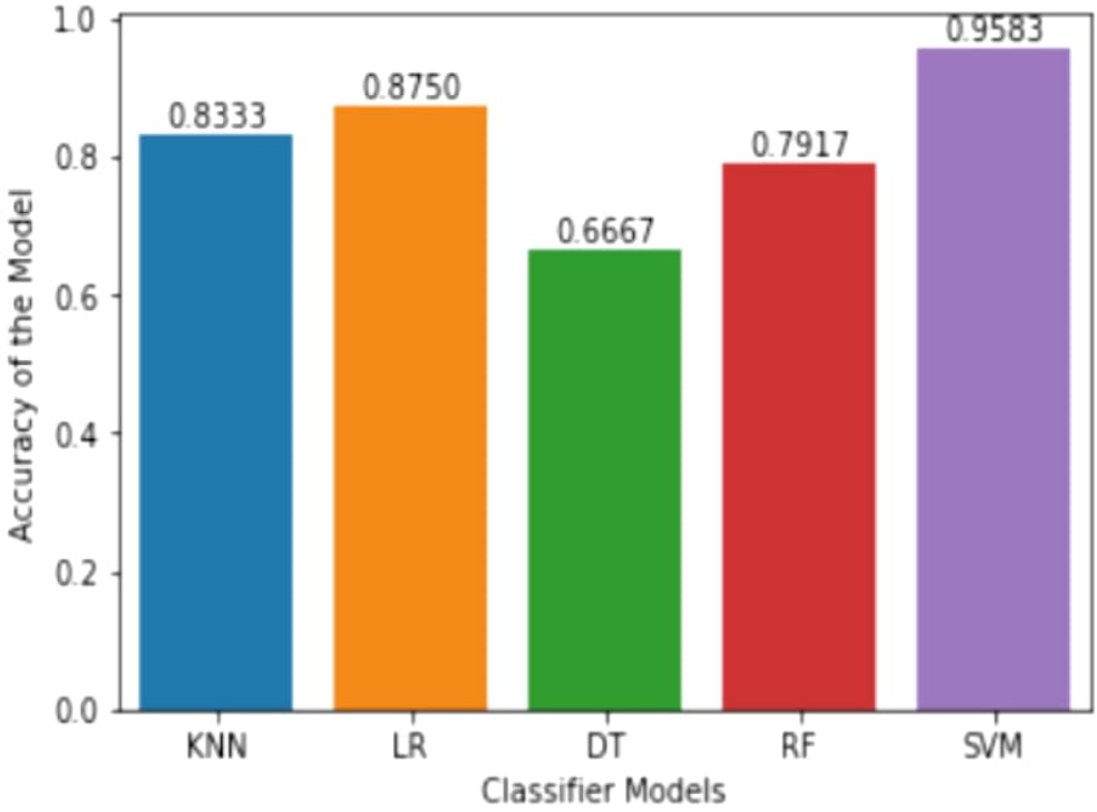


FIG 5.1 PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHM

susceptibility in the area which necessitates the identification of the factors prevalent in the region and their significance.

Also, it is worthy to note the importance of remote sensing which facilitated the derivation of most of the influencing factors, verification of the past flood events and generally plays a huge role in assessing potential areas susceptible to natural hazards. Subsequently, in achieving accurate results and high prediction based on the novelty of region efficient ML algorithms namely DT, SVM and RF integrated with IOE, and each model as a stand-alone model was implemented in deriving flood susceptibility maps for the study area. The proposed work is a way to evaluate the rainfall dataset to predict the flash flood using machine learning techniques with higher accuracy.

THE FOLLOWING STEPS STATED THAT THE PROPOSED MODEL PROVIDES A VERY EASY EFFICIENT METHOD FOR PREDICTING FLOOD:

STEP1: The collected dataset of rainfall is pre-processed.

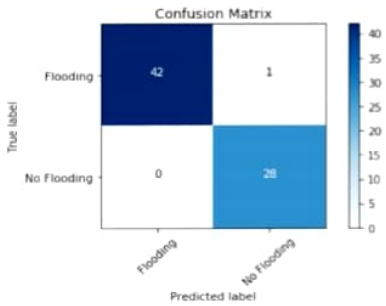
STEP2: The dataset of rainfall is randomly partitioned into testing and training.

STEP3: The dataset trained with LR, DT, KNN, RF and MLP algorithm.

STEP4: The model is constructed using the SVM algorithm with the highest accuracy and validated with the parameters such as precision, recall, f1-score, sensitivity, specificity, and accuracy.

STEP5: Input test data to the prediction model and validate the results. Accuracy of the algorithm calculated .

STEP6: Also pop up of flood warning system



	Name	Score
0	KNN	0.833333
1	LR	0.875000
2	DT	0.666667
3	RF	0.791667
4	SVM	0.958333

FIG5.0 PREDICTION LABLE AND ACCURACY

CHAPTER 5

RESULT DISCUSSION AND PERFORMANCE ANALYSIS

5.1 RESULT

The proposed work is a way to evaluate the rainfall dataset to predict the flash flood using machine learning techniques with higher accuracy. The performance assessment and prediction capability of the flood models area unit evaluated using each the coaching and therefore the testing datasets through the mythical monster curve and applied mathematics metrices. The major focus of this analysis is to develop and utilize machine learning-based flood susceptibility models in account flood inclined maps considering a various vary of factors relative to the study space and to spot the impact of the factors on flood incidence in the study space. This was performed taking account of the human-induced factors and other natural-caused factors acknowledged in previous studies and verified to possess a precise influence on flood incidence within the study space. Moreover, feature engineering was performed to research the prognostic ability, significance, and relation among the influencing factors before the most modelling and every one the factors had a precise influence and were thus used for the modelling. what is more, feature engineering was performed as there aren't any existing works associated with flood suscpectableness within the space that necessitates the identification of the factors prevailing within the region and their significance.

5.2 DISCUSSION

The major focus of this research is to develop and utilize machine learning-based flood susceptibility models in deriving flood susceptible maps considering a diverse range of factors relative to the study area and to identify the impact of the factors on flood occurrence in the study area. This was performed taking account of the human-induced factors and other natural-caused factors acknowledged in previous studies and proven to have a certain influence on flood occurrence in the study area. Moreover, feature engineering was performed to investigate the predictive ability, significance, and interrelationship among the influencing factors before the main modelling and all the factors had a certain influence and were therefore utilized for the modelling. Furthermore, feature engineering was performed as there are no existing works related to flood

SYSTEM REQUIRMENT SOFTWARE AND HARDWARE

SOFTWARE REQUIRMENTS:

- JUPTER NOTE BOOK
  (anaconda3)
- PYTHON 3
- MACHINE LEARNING LIBRARYS (pandas, nuppy, Matplotlib ….etc)
- Gmail-SMTP (smtplib, ssl)

HARDWARE REQUIRMENTS:

- **PROCESSOR:** Intel(R) Core (TM) i3-8145U CPU @ 2.10GHz 2.30 GHz
- **SYSTEM TYPE:** 64-bit operating system, x64-based processor
- **RAM:** > 4.00 GB
- **OS:** WINDOWS 11

INSTALLATION SET UP

- Go to Jupiter Notebook
- Go to File
- Select your File destination
- Select the source code
- Run the code
- PRINT EMAIL SENT SUCCESSFUL
- We Will Receive E-main

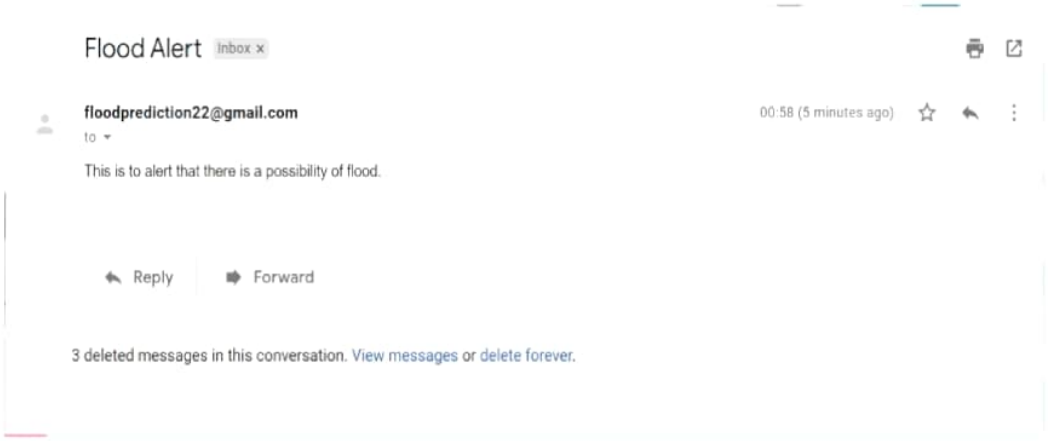
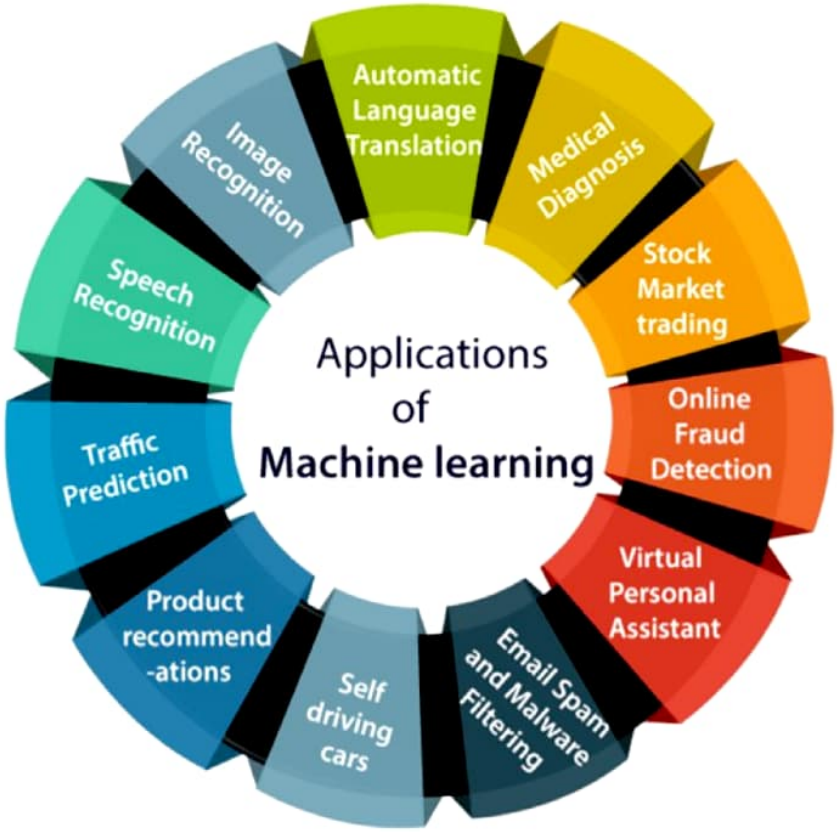
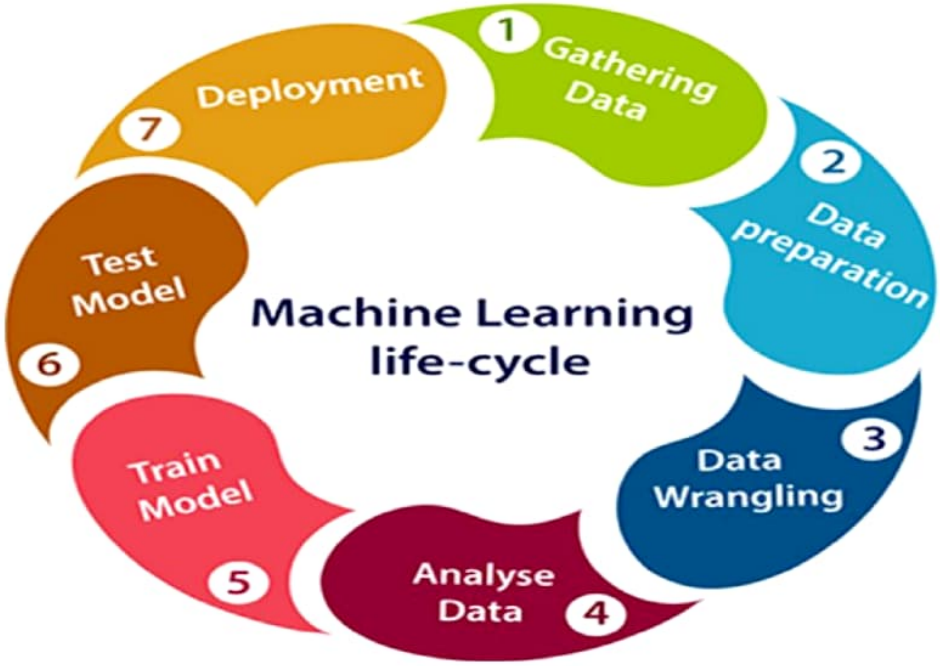


FIG 4.7 POP OF FLOOD ALERT

APPLICATION OF ML



LIFE CYCLE OF ML



MACHINE LEARNING

FIG 4.6 ML LOGO

In iOS applications, we use Core ML to incorporate machine learning in IOS applications. It is Apple's framework to use pre-trained models in IOS applications. In this tutorial, we will discuss what machine learning is, different types of it, including some real-life examples of machine learning. Machine learning is the field of study that allows computers to learn without being explicitly programmed. Using machine learning, we don't need to provide explicit instructions to Computers for reacting to some special situations. We need to provide training to the computers to find real-time solutions for the specific problems. The chess game is a famous example where machine learning is being used to play chess. The code lets the machine learn and optimizes itself over repeated games.

Machine Learning is broadly classified into two main categories.

- ***Supervised Machine Learning***

  Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

- ***Unsupervised Machine Learning***

  Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

FUNCTION'S OF PYTHON

Functions are the most important aspect of an application. A function can be defined as the organized block of reusable code, which can be called whenever required.

Python allows us to divide a large program into the basic building blocks known as a function. The function contains the set of programming statements enclosed by {}. A function can be called multiple times to provide reusability and modularity to the Python program.

The Function helps to programmer to break the program into the smaller part. It organizes the code very effectively and avoids the repetition of the code. As the program grows, function makes the program more organized.

Python provide us various inbuilt functions like **range()** or **print()**. Although, the user can create its functions, which can be called user-defined functions

There are mainly two types of functions.

- **User-define functions** - The user-defined functions are those define by the **user** to perform the specific task.
- **Built-in functions** - The built-in functions are those functions that are **pre-defined** in Python.

FIG 4.5 LOGO AND FOUNDER OF PYTHON

4.3 EXPREMENTAL ANALYSIS

BRIEF VIWE OVER THE TECHNOLOGY USED

PYTHON:

PYTHON is a widely used general-purpose, high-level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code.

In the late 1980s, history was about to be written. It was that time when working on Python started. Soon after that, Guido Van Rossum began doing its application based work in December of 1989 by at Centrum Wiskunde & Informatica (CWI) which is situated in Netherlands. It was started firstly as a hobby project because he was looking for an interesting project to keep him occupied during Christmas. The programming language which Python is said to have succeeded is ABC Programming Language, which had the interfacing with the Amoeba Operating System and had the feature of exception handling. He had already helped to create ABC earlier in his career and he had seen some issues with ABC but liked most of the features. After that what he did as really very clever. He had taken the syntax of ABC, and some of its good features. It came with a lot of complaints too, so he fixed those issues completely and had created a good scripting language which had removed all the flaws. The inspiration for the name came from BBC's TV Show – 'Monty Python's Flying Circus', as he was a big fan of the TV show and also he wanted a short, unique and slightly mysterious name for his invention and hence he named it Python! He was the "Benevolent dictator for life" (BDFL) until he stepped down from the position as the leader on 12th July 2018. For quite some time he used to work for Google, but currently, he is working at Dropbox.

The language was finally released in 1991. When it was released, it used a lot fewer codes to express the concepts, when we compare it with Java, C++ & C. Its design philosophy was quite good too. Its main objective is to provide code readability and advanced developer productivity. When it was released it had more than enough capability to provide classes with inheritance, several core data types exception handling and functions.

SUPPORT VECTOR MAHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Support Vector Machines (SVM), a fast and dependable classification algorithm that performs very well with a limited amount of data to analyze. Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Classification and Regression analysis use Support vector machine (SVM) algorithm for evaluating data. They have two categories in which data sets are trained. However, when new data comes, it can be easily categorized under a particular group. Thus, it is called a non-binary linear classifier.

SVM -STEPS INVOLVED:

1. Data Pre-processing
2. Fitting the SVM
3. Classifier to training set
4. Predicting the test set result
5. Creating Confusion matrix
6. Visualizing the train set result
7. Visualizing the test set result

RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier.There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

RANDOM FOREST ALGORITHM WORK

1. Select random K data points from the training set.
2. Build the decision trees associated with the selected data points (Subsets).
3. Choose the number N for decision trees that you want to build.
4. Repeat Step 1 & 2.
5. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

RANDOM FOREST ALGORITHM STEPS:

1. Data Pre-processing step
2. Fitting the Random forest algorithm to the Training set
3. Predicting the test result
4. Test accuracy of the result (Creation of Confusion matrix)
5. Visualizing the test set result.

DECISION TREE TERMINOLOGIES:

ROOT NODE: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

LEAF NODE: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

SPLITTING: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

BRANCH/SUB TREE: A tree formed by splitting the tree.

PRUNING: Pruning is the process of removing the unwanted branches from the tree.

PARENT/CHILD NODE: The root node of the tree is called the parent node, and other nodes are called the child nodes.

PROCESS OF DECISION TREE ALGORITHM:

1. Begin the tree with the root node, says S, which contains the complete dataset.
2. Find the best attribute in the dataset using Attribute Selection Measure (ASM).
3. Divide the S into subsets that contains possible values for the best attributes.
4. Generate the decision tree node, which contains the best attribute.
5. Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

DECISION TREE ALGORITHM STEPS:

1. Data Pre-processing step
2. Fitting a Decision-Tree algorithm to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

REFERENCE

[1] Y. Wang, Z. Fang, H. Hong, and L. Peng, "Flood susceptibility mapping using convolutional neural network frameworks," J. Hydrol., vol. 582, no. March, p. 124482, 2020, doi: 10.1016/j.jhydrol.2019.124482.

[2] R. Mind'je et al., "Flood susceptibility modeling and hazard perception in Rwanda," Int. J. Disaster Risk Reduct., vol. 38, no. April 2018, p. 101211, 2019, doi: 10.1016/j.ijdrr.2019.101211.

[3] W. Chen et al., "Modeling flood susceptibility using data-driven approaches of naïve Bayes tree, alternating decision tree, and random forest methods," Sci. Total Environ., vol. 701, 2020, doi: 10.1016/j.scitotenv.2019.134979.

[4] R. Costache et al., "Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning," J. Hydrol., vol. 585, no. December 2019, p. 124808, 2020, doi: 10.1016/j.jhydrol.2020.124808.

[5] I. E. Olorunfemi, A. A. Komolafe, J. T. Fasinmirin, A. A. Olufayo, and S. O. Akande, "A GIS-based assessment of the potential soil erosion and flood hazard zones in Ekiti State, Southwestern Nigeria using integrated RUSLE and HAND models," Catena, vol. 194, no. January, p. 104725, 2020, doi: 10.1016/j.catena.2020.104725.

[6] P. T. Padi, G. Di Baldassarre, and A. Castellarin, "Floodplain management in Africa: Large scale analysis of flood data," Phys. Chem. Earth, vol. 36, no. 7–8, pp. 292– 298, 2011, doi: 10.1016/j.pce.2011.02.002.

[7] I. Ajibade, G. McBean, and R. Bezner-Kerr, "Urban flooding in Lagos, Nigeria: Patterns of vulnerability and resilience among women," Glob. Environ. Chang., vol. 23, no. 6, pp. 1714–1725, 2013, doi: 10.1016/j.gloenvcha.2013.08.009

[8] J. Ntajal, B. L. Lamptey, I. B. Mahamadou, and B. K. Nyarko, "Flood disaster risk mapping in the Lower Mono River Basin in Togo, West Africa," Int. J. Disaster Risk Reduct., vol. 23, no. October 2016, pp. 93–103, 2017, doi: 10.1016/j.ijdrr.2017.03.015.

[9] I. Douglas, "Flooding in African cities, scales of causes, teleconnections, risks, vulnerability and impacts," Int. J. Disaster Risk Reduct., vol. 26, no. September, pp. 34–42, 2017, doi: 10.1016/j.ijdrr.2017.09.024

[10] C. C. Olanrewaju, M. Chitakira, O. A. Olanrewaju, and E. Louw, "Impacts of flood disasters in Nigeria: A critical evaluation of health implications and management," 52 Jamba J. Disaster Risk Stud., vol. 11, no. 1, pp. 1–9, 2019, doi: 10.4102/jamba.v11i1.557.

Classification Accuracy:

It is one of the important parameters to determine the accuracy of the classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers.

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+FN+TN}$$

NEED OF CONFUSION MATRIX:

- It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
- It not only tells the error made by the classifiers but also the type of errors such as it is either type-I or type-II error.
- With the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, etc.

3.RECALL SCORE

RECALL – The recall is the ratio **TP/ (TP+ FN)** where **TP** is the number of true positives and **FN** the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The best value is **1** and the worst value is **0**.

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved. For example, for a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned.In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough..

FORMULA OF F1 – SCORE

Recall score takes the (1) and (2) for the calculation. It gives the average weight of them.

RECALL= TP/ (TP+ FN)

EVALUATING A CLASSIFICATION MODEL:

1. PRECISION:

Precision is the proportion of positive prediction that observed to be correct. Precision refers to the amount of information that is conveyed by a number in terms of its digits; it shows the closeness of two or more measurements to each other. It is independent of accuracy. Precision refers to a value in decimal numbers after the whole number, and it does not relate with accuracy. The concepts of accuracy and precision are almost related, and it is easy to get confused. Precision is the amount of information that is conveyed by a value. Whereas Accuracy is the measure of correctness of the value in correlation with the information.

PRECISION FORMULA (PRECISION = TP / (TP + FP) TP)

Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

In the same way, we can write the formula to find the accuracy and recall.

Therefore,

Accuracy = (True positives + True Negatives)/ (True positives + True negatives + False positives + False negatives)

= (TP + TN)/ (TP + TN + FP + FN)

Recall = True positives/ (True positives + False negatives)

= TP/ (TP + FN)

2.CONFUSION MATRIX:

The confusion matrix provides us a matrix/table as output and describes the performance of the model.

It is also known as the error matrix.

The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions.

CLASSIFICATION

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1.** Classes can be called as targets/labels or categories. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output. The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data. Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes. Hence in this the SVM ,Random forest , KNN, LR and Decision tree classification algorithm is used for classifying the Rainfall data. 4 Types of Classification Tasks in Machine Learning

Machine learning is a field of study and is concerned with algorithms that learn from examples. Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as "spam" or "not spam." There are many different types of classification tasks that you may encounter in machine learning and specialized approaches to modeling that may be used for each. In this tutorial, you will discover different types of classification predictive modeling in machine learning.

Classification predictive modeling involves assigning a class label to input examples.

Binary classification refers to predicting one of two classes and multi-class classification involves predicting one of more than two classes. Multi-label classification involves predicting one or more classes for each example and imbalanced classification refers to classification tasks where the distribution of examples across the classes is not equal.

There are many different types of classification algorithms for modeling classification predictive modeling problems. There is no good theory on how to map algorithms onto problem types; instead, it is generally recommended that a practitioner use controlled experiments and discover which algorithm and algorithm configuration results in the best performance for a given classification task.

SPLITTING DATASET INTO TRAIN AND TEST DATA

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose, if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance

TRAINING SET: A subset of dataset to train the machine learning model, and we already know the output.

TEST SET: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

STEPS INVOLVED:

1. splitting arrays of the dataset into random train and test subsets.
2. In the second line, we have used four variables for our output that are

- **X train:** features for the training data
- **X test:** features for testing data
- **Y train**: Dependent variables for training data
- **Y test:** Independent variable for testing data

3. In train test split () function, we have passed four parameters in which first two are for arrays of data, and test size is for specifying the size of the test set. The test_size maybe .5, .3, or .2, which tells the dividing ratio of training and testing sets.

4. The last parameter random state is used to set a seed for a random generator so that you always get the same result.

MODEL SELECTION AND PREDICTION

Model selection is a method that may be used to pick models of various sorts. To anticipate floods, many machine learning methods are applied. SVM (support vector machine), Random forest, KNN, and Decision tree and LR Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Apart from this, it also depends on associated attributes, the volume of the available dataset, the number of features, complexity, etc. However, in practice, it is recommended that we always start with the simplest model that can be applied to the particular problem and then gradually enhance the complexity & test the accuracy with the help of parameter tuning and cross-validation.

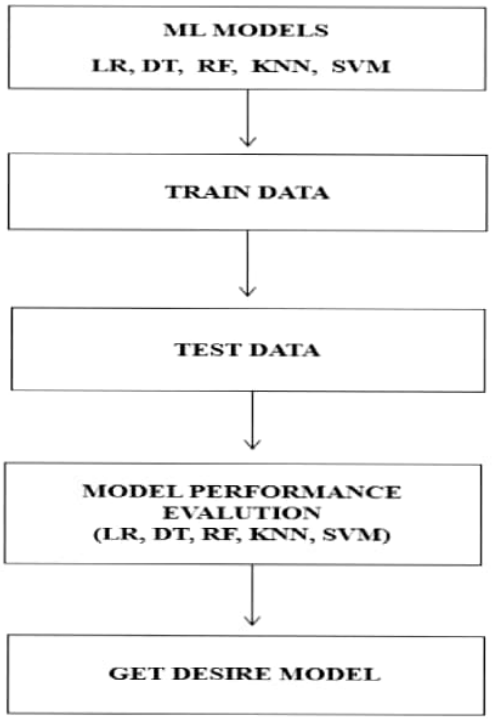


FIG 4.3 ML MODEL SELECTION PROCESS

Matplotlib:

The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code. It will be imported as below:

1. **import matplotlib.pyplot as mpt**

Here we have used mpt as a short name for this library.

Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library. It will be imported as below:

Here, we have used pd as a short name for this library. Consider the below image:

```
1 # importing libraries
2 import numpy as nm
3 import matplotlib.pyplot as mtp
4 import pandas as pd
5
```

3) IMPORTING THE DATASETS

Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory. To set a working directory in Spyder IDE, we need to follow the below steps:

1. Save your Python file in the directory which contains dataset.
2. Go to File explorer option in Spyder IDE, and select the required directory.
3. Click on F5 button or run option to execute the file.

Here, in the below image, we can see the Python file along with required dataset. Now, the current folder is set as a working directory.

GET THE DATASET

To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset.

Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient. So each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV file. However, sometimes, we may also need to use an HTML or xlsx file.

WHAT IS A CSV FILE?

CSV stands for "Comma-Separated Values" files; it is a file format which allows us to save the tabular data, such as spreadsheets. It is useful for huge datasets and can use these datasets in programs.

We can also create our dataset by gathering data using various API with Python and put that data into a .csv file.

2) IMPORTING LIBRARIES

In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:

1. **import numpy as nm**

Here we have used nm, which is a short name for Numpy, and it will be used in the whole program.

DATA PRE-PROCESSING

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task. Dataset undergoes pre-processing, and therefore, it has no missing values, no null values, and no duplicate values. The labels are converted into numeric forms to create an efficient machine-readable format. A new column flood is created, based meteorological data. The process of deleting undesirable data from a dataset is known as data pre-processing. STEPS: Import Numpy and Pandas libraries, which will be used to process the dataset. One of the most crucial steps is to import the dataset. The approach of removing data with missing values is inefficient. We are using null values instead of missing data. The prepared data is saved as a csv (Comma Separated Values) file, which is then loaded into pandas

STEPS INVOLVED IN DATA PRE-PROCESSING

1. Getting the dataset
2. Importing libraries
3. Importing datasets
4. Finding Missing Data
5. Encoding Categorical Data
6. Splitting dataset into training and test set
7. Feature scaling

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	1
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	1
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	1
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	1
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	0

FIG4.1 DATA PRE-PROCESSED

DATA PREPARATION

It is a research project that aims to prepare and assemble rainfall data in various parts of India, mostly in Kerala. The dataset is created and labelled appropriately once the Flood Prone region has been studied. Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence (BI), analytics and data visualization applications. The components of data preparation include data pre-processing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources. Data preparation is often referred to informally as *data prep*. It's also known as *data wrangling*, although some practitioners use that term in a narrower sense to refer to cleansing, structuring and transforming data; that usage distinguishes data wrangling from the data Pre-processing stage. Data preparation work is done by information technology (IT), BI and data management teams as they integrate data sets to load into a data warehouse, NoSQL database or data lake repository, and then when new analytics applications are developed with those data sets. In addition, data scientists, data engineers, other data analysts and business users increasingly use self-service data preparation tools to collect and prepare data themselves. Data preparation is often referred to informally as *data prep*. It's also known as *data wrangling*, although some practitioners use that term in a narrower sense to refer to cleansing, structuring and transforming data; that usage distinguishes data wrangling from the data pre-processing stage.

STEPS IN THE DATA PREPARATION PROCESS

1. Data collection. Relevant data is gathered from operational systems, data warehouses, data lakes and other data sources
2. Data discovery and profiling
3. Data cleansing
4. Data structuring
5. Data transformation and enrichment
6. Data validation and publishing

RAINFALL DATA SET

A dataset with the amount of rainfall and if a flood had occurred in a particular area/state/city, in the previous years, will be used. The dataset will have the rainfall data for a duration of 3 months approx. We are going to predict floods just for the state of Kerala depending on the rainfall dataset we are going to use. However, this method can be used for prediction for any state of India, with the given data. https://drive.google.com/file/d/13A5iG_F_8iim4px9Mj8fwFNL1Z7mSH_6/view

Using this dataset, for every 10 days, we take average rainfall and map it on a graph to display it.

We take this annual rainfall data as feedback to our model of machine learning and whether or not it induces a flood as output labels. We train and save our model (depending on certain daily rainfall threshold value in the dataset).

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	\
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	
..	...	...	...	...	...	...	...	...	...	...	
113	KERALA	2014	4.6	10.3	17.9	95.7	251.0	454.4	677.8	733.9	
114	KERALA	2015	3.1	5.8	50.1	214.1	201.8	563.6	406.0	252.2	
115	KERALA	2016	2.4	3.8	35.9	143.0	186.4	522.2	412.3	325.5	
116	KERALA	2017	1.9	6.8	8.9	43.6	173.5	498.5	319.6	531.8	
117	KERALA	2018	29.1	52.1	48.6	116.4	183.8	625.4	1048.5	1398.9	

	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	197.7	266.9	350.8	48.4	3248.6	YES
1	491.6	358.4	158.3	121.5	3326.6	YES
2	341.8	354.1	157.0	59.0	3271.2	YES
3	222.7	328.1	33.9	3.3	3129.7	YES
4	217.2	383.5	74.4	0.2	2741.6	NO
..	...	...	...	...	...	...
113	298.8	355.5	99.5	47.2	3046.4	YES
114	292.9	308.1	223.6	79.4	2600.6	NO
115	173.2	225.9	125.4	23.6	2176.6	NO
116	209.5	192.4	92.5	38.1	2117.1	NO
117	423.6	356.1	125.4	65.1	4473.0	YES

[118 rows x 16 columns]

FIG 4.1RAINFALL DATASET

Thus, the methodology of this literature review aims to include the most effective flood resource variables in the search queries.

A combination of these flood resource variables and ML methods was used to implement the complete list of search queries. Note that the ML methods for flood prediction may vary significantly according to the application, dataset, and prediction type. For instance, ML methods used for short-term water level prediction are significantly different from those used for long-term streamflow prediction.

4.1 WORKING OF MODULE

The prediction accuracy of the different model is evaluated using data validation and result are compared to get accuracy. The accuracy of training dataset, accuracy of the testing dataset, false-positive rate specification precision, and recall are calculated by comparing algorithm

The steps involved:

- Define a problem
- Preparing data
- Evaluating algorithms
- Prediction result

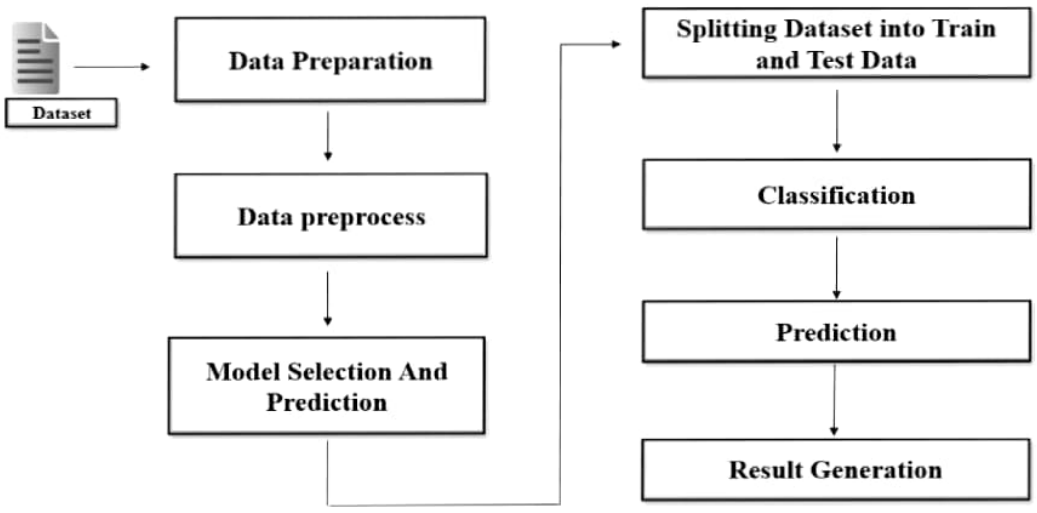


FIG 4.0 SYSTEM ARCHITECTURE

CHAPTER 4

EXPREMENTAL OR MATERIAL METHODS ALGORITHM

The rainfall data for Kerala was utilized for the analysis, and it spans the years 1901 to 2018. The data is organized in a CSV file format by month, with sub-divisions within Kerala. The unit of measurement for rainfall is millimeter (mm). The data was gathered from the Indian Metrological Department. The machine learning approach is used to evaluate the performance of various algorithms such as KNN, LR, SVM, DT, and RF. his survey identifies the state of the art of ML methods for flood prediction where peer-reviewed articles in top-level subject fields are reviewed. Among the articles identified, through search queries using the search strategy, those including the performance evaluation and comparison of ML methods were given priority to be included in the review to identify the ML methods that perform better in particular applications. Furthermore, to choose an article, four types of quality measure for each article were considered, i.e., source normalized impact per paper (SNIP), Cite Score, SC Imago journal rank (SJR), and h-index. The papers were reviewed in terms of flood resource variables, ML methods, prediction type, and the obtained results.

The applications in flood prediction can be classified according to flood resource variables, i.e., water level, river flood, soil moisture, rainfall–discharge, precipitation, river inflow, peak flow, river flow, rainfall–runoff, flash flood, rainfall, streamflow, seasonal stream flow, flood peak discharge, urban flood, plain flood, groundwater level, rainfall stage, flood frequency analysis, flood quantiles, surge level, extreme flow, storm surge, typhoon rainfall, and daily flows .Among these key influencing flood resource variables, rainfall and the spatial examination of the hydrologic cycle had the most remarkable role in runoff and flood modeling . This is the reason why quantitative rainfall prediction, including avalanches, slush flow, and melting snow, is traditionally used for flood prediction, especially in the prediction of flash floods or short-term flood prediction. .However, rainfall prediction was shown to be inadequate for accurate flood prediction. For instance, the prediction of streamflow in a long-term flood prediction scenario depends on soil moisture estimates in a catchment, in addition to rainfall. Although, high-resolution precipitation forecasting is essential, other flood resource variables were considered in the

Therefore main aim of this is to get all the rainfall data of India and from a dataset containing yearly rainfall data. By providing real time input to different models of machine learning, those are Logistic Regression, Support Vector Machine, K-Nearest Neighbors Decision Tree Classifier and Random Forest algorithm. The input provided to models are pre-processed and patterns are extracted by getting maximum accuracy. The data provided is split into a Training set and Test set. It is split in the ratio of 7:3. The all five models are used to predict and by comparing all the results of model and considering the confusion matrix of all the models the accuracy is determined. The best model is chosen by comparing the accuracy of each model.

3.1 SCOPE OF FLOOD PREDICTION USING MACHINE LEARNING

The Scope of Flood Prediction using Machine Learning is to design a model to predict the flood using the rainfall data.

The prediction of different models is taken and compared within each other to find the best model that has high accuracy.

The flood can be predicted in different states of India in different months.

The confusion matrix of different models in Machine learning is considered to evaluate the accuracy and precision of the system.

To empower local communities to directly participate in the design of emergency services dealing with mitigation actions for floods.

To harness the power of new technologies, such as social media and mobile technologies, to increase the efficiency of public administrations in raising public awareness and education regarding floods risks, effects and impact.

To make use of the best available data in order to identify the location and potential impacts that natural hazards as floods can have on people, property and natural environment.

To improve the systems of warning and emergency communications.

The goal of Flood Prediction is to issue advance warning about water level or discharge large enough that threatens safety of structures and flood plain activities.

As observed in previous module, an advance warning of this nature help authorities adopt a series of measures to contain adverse impacts of flood.

Unlike several other disasters, approaching flood can be predicted ahead of its occurrence with advance collection of hydro-meteorological data, and its transformation into flood water level or flood hydrograph.

CHAPTER 3

AIM AND SCOPE

The aim of this project is to develop a flood prediction model which is real time. This could be helpful in the areas where the flash flood occurs. This system takes the input as the rainfall data all over the India process it using different machine learning models and the best model is determined with the help of accuracy of different algorithms, which would help people prior, save lives and also save lots of meteorological efforts.

THE MAIN AIM OF THE PROJECT

- Flood warning service is to detect and Predict threatening flood events so that the public can be alerted in advance and can undertake appropriate responses to minimise the impact of the event.

- As such, the primary objective of a flood warning system is to reduce exposure to coastal flooding.

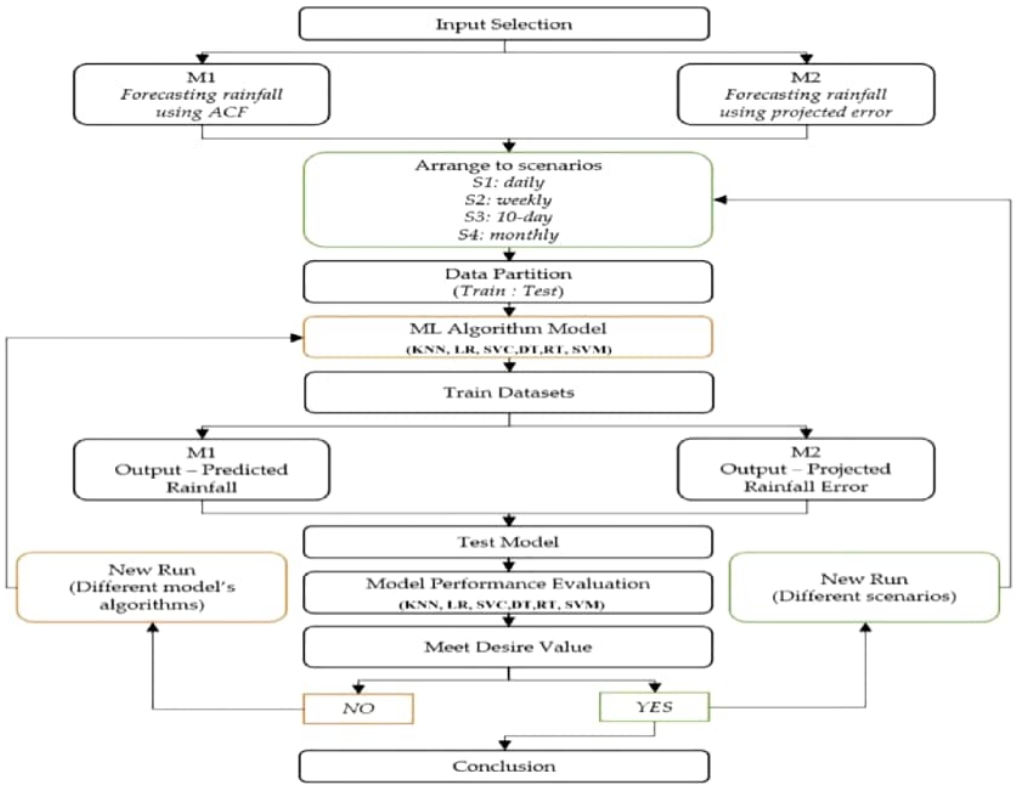


FIG 3.0 WORK SYSTEM OF FLOOD PREDICTION