Classifying Flood Severity Using Machine Learning

MSc Research Project
Data Analytics

Jayanta Behera
Student ID: x18188834

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Dr. Pramod Pathak

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jayanta Behera		
Student ID:	x18188834		
Programme:	Data Analytics	**Year:**	2020
Module:	MSc Research Project		
Supervisor:	Dr. Paul Stynes, Dr. Pramod Pathak		
Submission Due Date:	28th September 2020		
Project Title:	Classifying Flood Severity Using Machine Learning		
Word Count:	**7434** **Page Count 20**		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jayanta Behera

Date: 28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Flood Severity Classification Using Machine Learning

Jayanta Behera
x18188834

Abstract

Flood is one of the most devastating natural hazards that cause huge loss to human life and property. An early and accurate disaster prediction is helpful to prevent the damage. The complexity of factors contributing to flood prediction becomes a challenge in predicting its severity. This research illustrates a novel technique of combining the historical flood incidents with the meteorological and topographic features to predict flood severity by classifying its risk as high, low or moderate. To achieve this, random forest classifier is implemented along with support vector machine, k nearest neighbour, ensemble techniques and neural network. Each of the model is optimized and evaluated based on accuracy, pression, recall and F1-score where random forest classifier outperformed all other techniques with 83% accuracy. This novel technique of combination of historic data with climatic and topographic details showed potential improvement in predicting such catastrophic event which would help in planning proper evacuation and preventing loss of life and property.

Keywords: Flood severity, Random forest, Bagging, accuracy, precision, recall

1 Introduction

Flood is defined as a temporal water overflow into the dry land causing huge damage to property and lives. It is mainly caused due to heavy rainfall, tsunami, broken dam, snow melting or low pressure. Even with all the advancement in machine learning techniques, recent studies showed that of all the natural hazards, flood has caused most damage contributing to 24% of life loss in 127 major cases out of 315 registered in year 2017[1]. The world-wide flood incidents have also increased to 40% over the last two decades (Khosravi et al., 2019). This trend is expected to increase almost five times by 2050 and up to seventeen times by 2080 in Europe (Costache, Popa, et al., 2020). A report suggested that between 1995 to 2015, more than one hundred million people were affected by several flood incidents costing almost seventy-five billion USD per year (Karyotis et al., 2019). Even today, around twenty thousand people die because of flood. This shows that the current techniques are not capable enough of predicting the flood incident and its intensity accurately. Hence, further analysis is required.

In order to minimize the loss, different countries have taken numerous actions to monitor and examine the flood occurrences. Different machine learning techniques have been developed to identify the flood occurrence and classify the flood zones so that proper preparation could

[1] https://reliefweb.int/report/world/flooding-affected-more-people-2018-any-other-disaster-type-report-shows

be done to avoid life and property loss. Several researchers have used these machine learning techniques in flood risk assessments. However, most of these conventional approaches have relied on the hydrological measures to estimate the damages caused (Goel, 2020). Some other approaches have relied on image data to classify the flood zones. Still, the climatic conditions are not utilized as a major determining factor in the study of flood prediction. More so, the topographic factor like ground level elevation has not given much emphasis. Therefore, this research focuses on a novel approach of classifying the flood risk by combining the historical flood incidents that have happened since 1985 till date around the world with the ground elevation of the place and the weather data. The climatic data is collected for the day of flood incidence as well as previous four days to see its impact on flood severity. Figure 1 below shows various features that are used to classify the flood zone as high, low or moderate risk based on the data from department of flood history, meteorology and topography.

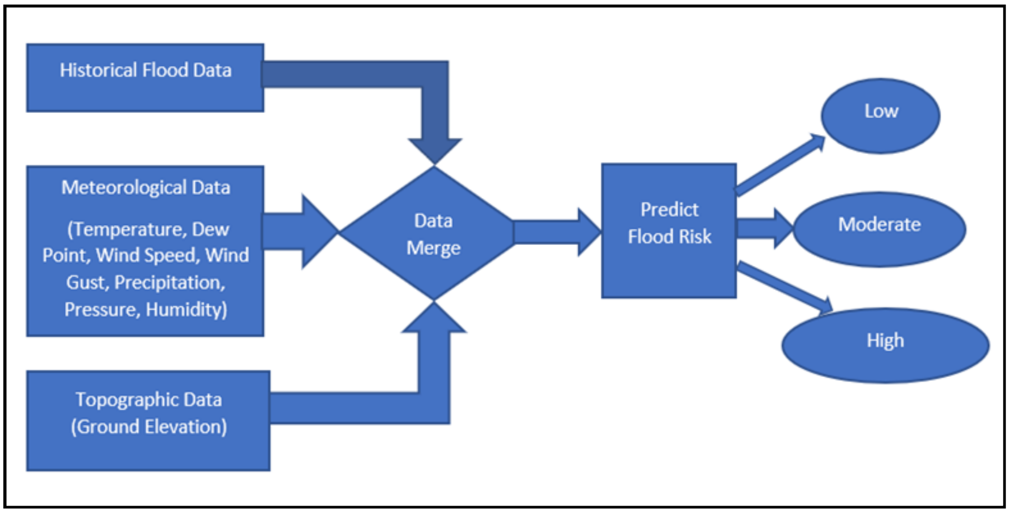

Historical Flood Data
Meteorological Data
(Temperature, Dew Point, Wind Speed, Wind Gust, Precipitation, Pressure, Humidity)
Topographic Data (Ground Elevation)
Data Merge
Predict Flood Risk
Low
Moderate
High

Figure 1: Flood Risk Classification using Historical, Meteorological and Topographic Data.

The aim of this research is "***To what extent, can machine learning techniques be used to predict the flood severity by classifying the risk as high, low or moderate based on the combination of historical flood data with climatic and topographic features?***"

In order to address the above research question, the following research objectives are defined-

- Objective 1- Extraction of Flood data and merging with the weather condition (that include temperature, precipitation, wind speed, dew point, humidity, pressure, wind gust, sky condition) and topographic data (ground level elevation from sea level) via web scrapping and Application Program Interface.
- Objective 2- Imputation of missing values with Multivariate Imputation via Chained Equations package.
- Objective 3- Selection of features contributing to flood risk prediction via Recursive Feature Elimination, Random Forest Classifier, Boruta package and Backward Feature Elimination.
- Objective 4- Data transformation by feature scaling, one hot encoding and standardisation.
- Objective 5- Handling class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) analysis.
- Objective 6- Data Clustering using K-Means and t-Distributed Stochastic Neighbour Embedding.

- Objective 7- Performing dimensionality reduction using Principle Component Analysis, t-distributed Stochastic Neighbour Embedding, Singular Value Decomposition and Independent component analysis.
- Objective 8- Implementation of machine learning algorithms on the engineered data using Random Forest, Decision Tree, k Nearest Neighbour, Support Vector Machine, Ensemble Methods and Neural Network.
- Objective 9- Cross validation and model optimization to perform prediction and evaluation using confusion matrix parameters such as precision, recall, F1-score and accuracy.

The result of this research would help the government and non-government officials to precisely identify the flood risk based on the past data such as land displaced, people died, flood duration and the climatic condition and predict the flood severity so that recursive measures can be taken to evacuate. The disaster recovery teams can thus accurately plan for the aftermath and migrate people to safer zones in the disastrous period.

This paper discusses the related work in section 2, methodology, design specification and implementation in section 3, 4 and 5 respectively. Section 6 comprises evaluation, followed by discussion and conclusion in the subsequent sections.

2 Related Work

This section critically investigates the implemented techniques on flood prediction. It is divided into several sub-sections, i.e. (1) Features influencing flood, (2) Significance of flood history and flood types, (3) Application of non-tree-based techniques like Support Vector Machine, k nearest neighbour for flood classification, (4) Image as input vectors in flood classification, (5) Neural networks and Deep learning in Flood Forecasting, (6) Traditional tree-based algorithms and ensemble techniques.

2.1 Feature Selection

Karyotis et al. (2019) performed research on urban flood using meteorological, hydrological, geographical features and developed a flood monitoring and forecasting platform. The research outcome signified the importance of climatic and geographic features in flood prediction study. Ogale and Srivastava (2019) developed a theoretical model to determine flash flood using artificial neural network and emphasized the importance of land elevation, surface run-off and location drainage. Along with the climatic and geographic features, Alipour et al. (2020) used storm data to predict the flood duration and classified the regions as damaged or not with the help of specificity index. All these studies showed range of complex features on flood occurrence and their significance in predicting flood severity which is confirmed in another study by Khosravi et al. (2019) where the authors have used lithological and vegetative data. However, Goel (2020) performed flood prediction with only rainfall data of three states in India and achieved good regression output with a little root mean squared error signifying the importance of rainfall in predicting flood as the major flood predicting parameter.

Contradicting to the above studies, Puttinaovarat and Horkaew (2020) discovered the limitation of rain gauge installation and included data from crowd source along with the geo-space to predict the flood map. However, the research was more leaned towards the technical concepts of Hadoop and hence given little importance. Another group of researchers used the depth of river water in measuring the impact on high intensity flood zones (Furquim et al., 2014). With the hydrological feature alone, the researchers improved the model's accuracy.
Thus, the significance of climatic and geographic features proved to be vital in the study of flood prediction.

2.2 Significance of Flood History

Several researchers used the historical flood incidents around the world to classify based on flood damages. Khalaf et al. (2018) attempted the flood mitigation issues with two thousand annotated flood events and classified them into normal, and high-risked. The researchers took the number of deaths, flood duration and displaced land. The historical flood incidents helped in creating flood risk zones without even considering environmental parameter. Some studies have implemented principle component analysis to reduce the dimensions of the input features, while others have used linear discriminant analysis. These steps confirm the importance of dimensionality reduction in flood prediction studies and the convenience of working with less dimensional data explaining most of the variances in dataset.
Alipour et al. (2020) took more than fourteen thousand flood incidents of south-east region of united states and classified the damaged regions and estimated flood duration, frequency and magnitude. More record counts in the dataset take care of under-sampling problem and encourage the usage of neural network. Mosavi, Ozturk and Chau (2018) confirmed in their studies that only the weather parameters can accurately predict short term floods like flash flood, rainfall run-off, while a decade long data can accurately predict long term flood incidents.
These studies showed the importance of flood archives to estimate the flood intensity and encourage to consider them as feature vectors.

2.3 Non-Tree-Based Machine Learning Techniques on Flood Prediction

Using only rainfall parameter, Goel (2020) performed flood regression and got good results with k nearest neighbour with a little root mean squared error. This shows the efficiency of k nearest neighbour algorithm in the flood study and its significance among non-tree-based algorithm. However, this was contradicted in a study performed by a group of researchers where they performed flood classification considering flood archive using k nearest neighbours, but it didn't show prominent accuracy when compared with the accuracy of support vector machine Khalaf et al. (2018). The researchers conveniently marked the flood zones with 76% accuracy using the later method. However, their study was limited to mere two years of data, hence the outcomes were not efficient enough. Also, the authors didn't provide any justification of the manually calculated columns used as input vectors such as magnitude of the flood, flood frequency. Nguyen and Chen (2020) used support vector machine algorithm for deterministic forecast and got that the model performed well with low root mean square error. However, probabilistic method performed much better with narrow bandwidth and provided a more practical prediction. The results of the models were quite poor due to which the researchers resampled the results with k nearest neighbour regressor to smoothen the probabilistic curve. Nguyen and Le (2019) used support vector regressor with the data from the downstream stations for tracking river water level. With Nash coefficient, the authors measured the accuracy economically which performed much better than neural networks and random forest. This indicates the importance of hydrological parameters in the flood studies.

Boukharouba et al. (2013) used clustering technique and applied machine learning models in each of the clusters using support vector regression and got much better result when compared to applying the model on global model. Their research distinguished the high-altitude ranges and the low ones as high altitudes have few flood incidents and low altitude have significantly more incidents. The significance of clusters formation using showed the legitimacy of classification problems achieved through exploratory data analysis. Another group of researchers used historical flood got more than 78% classification accuracy of each class with random forest compared to 77% in Levenberg-Marquardt learning algorithm (Khalaf et al., 2018). This leads to the scope of studying and implementing the tree-based algorithms as they seem to have more accuracy than non-tree-based counterpart in the classification problem. Thus, of all the non-tree-based algorithms, most of the researchers argued that support vector machine outperforms every otherwhile considering multiple dimensional datasets.

2.4 Neural Networks and Deep Learning in Flood Forecasting

Neural networks are used in most of the modern studies due to less training requirements. Their outcomes work well for complex problems such as multiclass classification, sequence determination etc. Ranit and Durge (2019) used the rainfall data with river runoff data and calculated the inflow and outflow of storage water by applying artificial neural network to control the reservoir storage. The result of this study showed the benefit of using neural networks in determining different flood types like flash flood, seasonal flood, coastal flood, urban flood etc. without training the dataset separately for each flood type with certain level of parameter tuning. Another study predicted the flood occurrence using artificial neural network in the interior region (Puttinaovarat and Horkaew, 2020) to track the precipitation details by combining support vector machine and random forest with the outcomes of artificial neural network using the meteorological, hydrological, geographic locations and got 97% accuracy and 0.10 root mean square error value. Artificial neural networks were also used with historical flood data as feature vectors to determine the flood severity (Khalaf et al., 2018). The outcomes were compared with traditional machine learning techniques and showed much better results without even tuning each of the parameter unlike the traditional technique implementation. Researchers have applied feedforward network in convolutional neural networks on the satellite images and produced flood susceptible map by feeding the model output to SVM. However, these studies were left open for further augmentation and usage of big datasets. Noymanee and Theeramunkong (2019) used Bayesian linear regression with neural network using flood parameters such as rainfall, water level, drainage etc. and developed a flood preparedness system which helped in error reduction much better than the former model. The result of this study indicated that probabilistic model outcomes are not quite efficient in the flood study compared to neural networks.

Ding et al. (2019) applied long short-term memory recurrent neural network to explore the relation of hydrological features with the final run-off and used time series for flow prediction. Even though the model showed better results compared to SVM, it required more optimization and justification of applying recurrent neural network as the work of forget gate was not justified. Researchers have used one dimensional convolutional neural network in flood monitoring and forecasting with the range of parameters discussed in the earlier section along with different optimizers to get minimal regression error (Karyotis et al., 2019). However, there was little work made on computer vision due to lack of data availability.

2.5 Tree Based Traditional and Ensemble Techniques

Even though neural networks are widely used in the study of flood prediction, it is often observed that the traditional techniques provide accurate results due to the inaccuracies in the dataset that are handled manually. Furquim et al. (2014) used a comparative application of multi-layer perceptron ANN and BF-Tree with 10-fold k-validation to predict flash flood in river water measurement. The results showed that BF-Tree classified the river behaviour quickly with ease, compared to neural network. The research also confirmed the significant of using k-fold cross validation techniques in flood studies to overcome the problem of overfitting. The study was limited to level value as the research gave no regression output. Chen et al. (2020) implemented decision tree and compared the results with random forest, which provided an accurate assessment of flood risk. Random forest outcome helped to alleviate the urban flood disasters by identifying the high flood-risk levels. The study confirmed the significance of using random forest classifier in the flood classification study. Felix and Sasipraba (2019) used ensemble technique by applying gradient boosting algorithm to classify the flood zones with water level measured via rain gauge. Even though the authors got good results, but the number of parameters considered were too less. More features as input vectors could have resulted in better reliability on outcome. Thus, the research encouraged the usage of ensemble technique in cross verifying the results against the conventional tree-based outcome. In contrast, k nearest neighbour ensemble technique used by Costache, Pham, et al. (2020) to determine flood potential index was more accurate than any other method. Using the statistical indices, the researchers confirmed the performance such as accuracy, sensitivity and specificity. This was also confirmed by Shahabi et al. (2020) where they used bagging cubic- k nearest neighbour ensemble technique to identify various flood zones and got good area under the curve value. The result of the study showed performance improvement by using bagging technique.

All these above studies showed that ensemble methods perform much better for such datatypes than any of the traditional techniques used so far.

2.6 Flood Severity Using Image as Feature Vector

Flood zones classification were also done using areal images as input parameters by (Akshya and Priyadarsini; 2019; Sachdeva, Bhatia and Verma; 2017; Opella and Hernandez; 2019;

Zaji, Bonakdari and Gharabaghi; 2019). Aerial images were used as input vectors for support vector machine hybridization and k-means clustering while others have combined with hydrological or topographic features to classify flood zones and to calculate the area affected by the river water discharge. Even though the images were considered in flood study, the current research focuses on using the flood archive, climatic and topographic feature and not the images.

From the above studies, it is evident that the flood severity prediction is quite complex as it is not limited to a few parameters. So far, the researchers in the field of machine learning have not incorporated various climatic and topographic features with flood archive to predict the flood severity. Therefore, there is a need to develop a model that would use these features to classify the flood severity as high, low or moderate. To incorporate this, the approach and the methodology are discussed elaborately in the next section.

3 Research Methodology

This section discusses various stages of flood severity classification and is based on Knowledge Discovery in Database (KDD) approach. An elaborate discussion of flood severity is explained in various stages such as (1) Business understanding, (2) Data acquisition, (3) Pre-processing, (4) Feature selection, (5) Clustering, (6) Dimensionality reduction as illustrated in Figure 2.

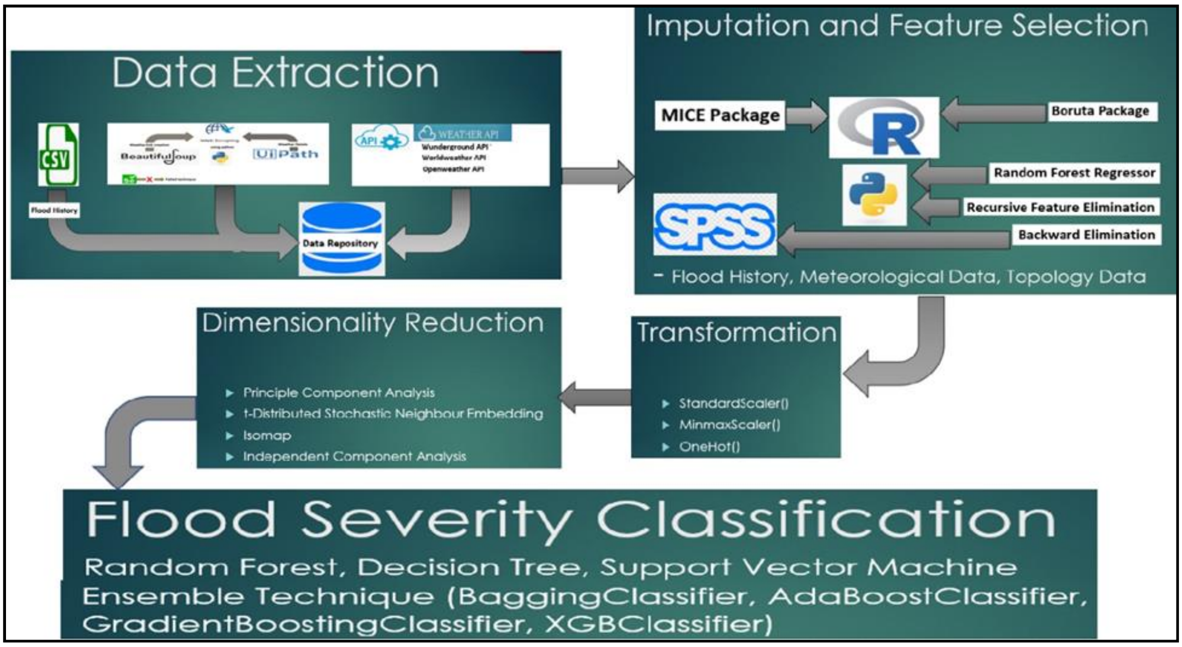

Data Extraction
CSV
Flood History
BeautifulSoup
UiPath
API
WEATHER API
Wunderground API
Worldweather API
Openweather API
Data Repository
Imputation and Feature Selection
MICE Package
R
Boruta Package
Random Forest Regressor
Recursive Feature Elimination
Backward Elimination
SPSS
- Flood History, Meteorological Data, Topology Data
Transformation
StandardScaler()
MinmaxScaler()
OneHot()
Dimensionality Reduction
Principle Component Analysis
t-Distributed Stochastic Neighbour Embedding
Isomap
Independent Component Analysis
Flood Severity Classification
Random Forest, Decision Tree, Support Vector Machine
Ensemble Technique (BaggingClassifier, AdaBoostClassifier, GradientBoostingClassifier, XGBClassifier)

Figure 2: Flood Severity Classification- Methodology

3.1 Business Understanding

Hydrological factors are common in flood prediction. However, this research focuses on a novel technique of combining flood archives with climatic and topographic aspects to predict the flood severity as high, moderate or low risk.

3.2 Data Acquisition

The worldwide flood incidents occurred from 1985-till date are taken from official website of Colorado[2]. The file is downloaded as comma separated file with 4926 flood incidents containing 14 flood attributes like Country, longitude, latitude, affected area, flood date, deaths, main cause, severity. Longitude, latitude and dates are used to extract the weather details of the flood incident. With these features, web-links were created using beautiful soup library in python to extract the location code and saved in the original dataset. The dataset is fed to UiPath (a robotics automation path tool) and a sequential workflow is designed to extract the weather attributes from the weatherunderground website via web scrapping. The extracted data is later saved to the dataset with column names as Day0_Temperature, Day0_Dew_Point, Day0_Humidity, Day0_Wind, Day0_Wind_Speed, Day0_Wind_Gust, Day0_Pressure, Day0_Precipitation, Day0_Condition.

The web scrapping process took prolonged time to run and the website didn't provide data for all the flood date. Hence, the website's application program interface was used to extract data which is much faster than the web scrapping technique. The weather data of previous 4 days of flood incident are extracted from different websites[345] via APIs to find the impact of climate on flood intensity. The ground elevation of the flood location above sea level is extracted from maps website via API. Initially, selenium library was used, but discarded later due to its limitations in python interface. The final dataset contained 56 columns as shown in Table 1. With this, the objective 1 is accomplished.

Table 1: Dataset Description

Dataset	Record Count	Attribute Count
Global Active Archive of Large Flood Events	4926	14
Historical Weather API	4926	9
Web Scrapping via UiPath for Flood Day	2200	9
Topological (Ground Elevation) API	4926	1

3.3 Data Pre-processing

The above features were merged, and missing values were handled as part of pre-processing.

3.3.1 Merging of Dataset

Based on flood date, latitude, longitude, weather and topographic data extracted via web scraping and API are merged with the flood archive as shown in Figure 3.

[2] http://floodobservatory.colorado.edu/Archives/

[3] https://www.wunderground.com/

[4] https://openweathermap.org/history

[5] https://www.worldweatheronline.com/

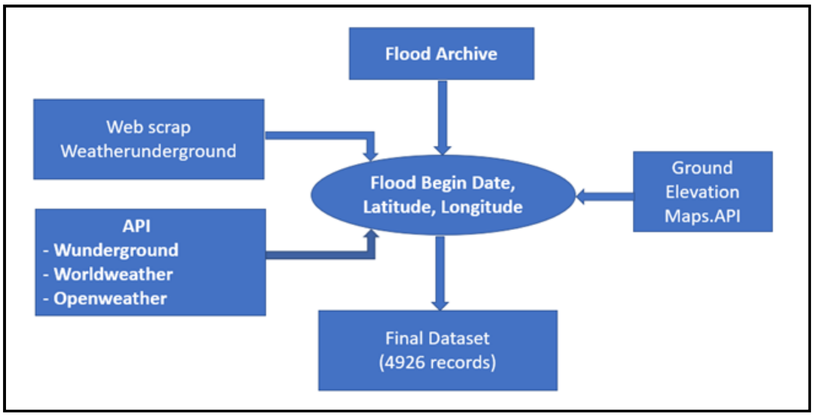


Figure 3: Merging of Dataset

3.3.2 Data Imputation

Using Amelia library's missmap function in RStudio, 23% of missing values in the dataset are identified. MICE (Multivariate Imputation via Chained Equations) package in RStudio is used to impute the missing. Figure 4, 5 shows distribution pattern of data imputation in Dew_Point, Wind_Gust happened in four stages with the blue ones are original values and red ones as imputed values. With this, objective 2 is accomplished.

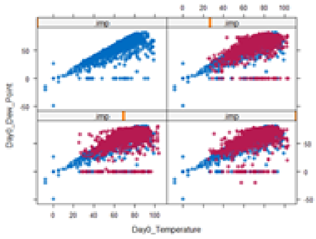


Figure 4: Imputing DewPoint Distribution

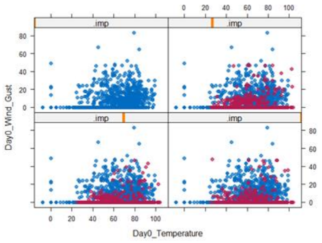


Figure 5: Imputing WindGust Distribution

3.4 Feature Engineering

Feature engineering deals with logarithmic transformation, one-hot encoding and feature scaling.

3.4.1.Logarithmic Transformation

In order to suffice normality distribution, logarithmic value of some of the variables are taken such as Area_Affected, Duration, Dead, Displaced. Figure 6(a), 6(b) shows the normal curve against the logarithmic curve for variables Area and Dead.

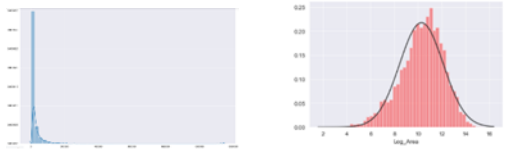


Figure 6(a): Area vs Log_Area

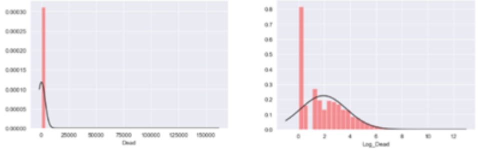


Figure 6(b): Dead vs Log_Dead

3.4.2.One-Hot Encoding

The categorical variables such as MainCause and Day_0_Weather_Condition, Day.1_Weather_Condition, Day.2_Weather_Condition, Day.3_Weather_Condition, Day.4_Weather_Condition are encoded to binary variables using one-hot encoding before feeding the data to the model for better prediction. For example, MainCause is converted to MainCause_0, MainCause_1 etc.

3.4.3.Feature Scaling

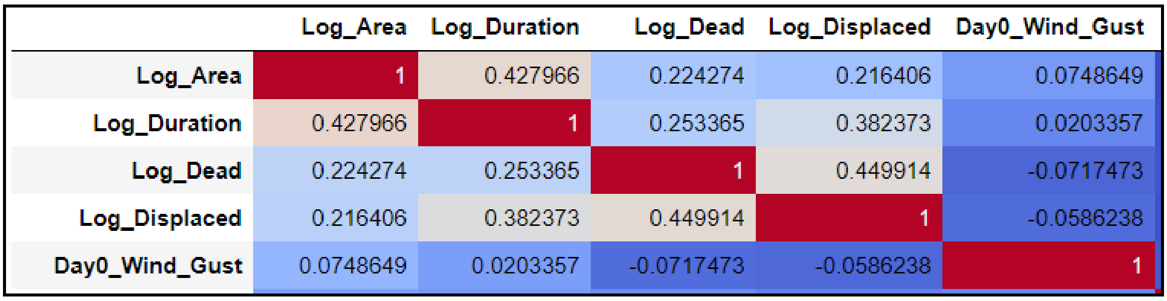
The original feature vectors are in different scales. Standardization is done to bring them into same scale so that vectors with higher scale range should not affect the lower ones. However, for tree-based models like Random Forest, XG Boost, these are not done as they build trees based on absolute values. With these experiments, objective 3 is accomplished.

3.5 Feature Selection

Feature selection is done using Backward feature elimination, Recursive feature elimination, random forest classifier, Boruta package and correlation matrix.

3.5.1.Correlation Matrix

Correlation matrix is plotted for all the numerical variables to check the correlation of the independent variables. It is observed that similar variables like temperature columns, dew point, wind speed etc. are correlated with each other. As there are lot of numerical features, a sample of the correlation matrix is shown in Figure 7.



	Log_Area	Log_Duration	Log_Dead	Log_Displaced	Day0_Wind_Gust
Log_Area	1	0.427966	0.224274	0.216406	0.0748649
Log_Duration	0.427966	1	0.253365	0.382373	0.0203357
Log_Dead	0.224274	0.253365	1	0.449914	-0.0717473
Log_Displaced	0.216406	0.382373	0.449914	1	-0.0586238
Day0_Wind_Gust	0.0748649	0.0203357	-0.0717473	-0.0586238	1

Figure 7: Correlation Matrix between Numerical Variables

3.5.2.Boruta using RStudio

Boruta package is used in RStudio to find out the significant features. It gave 45 significant features. Figure 8 shows some of the significant features highlighted in green and insignificant features in blue and red.

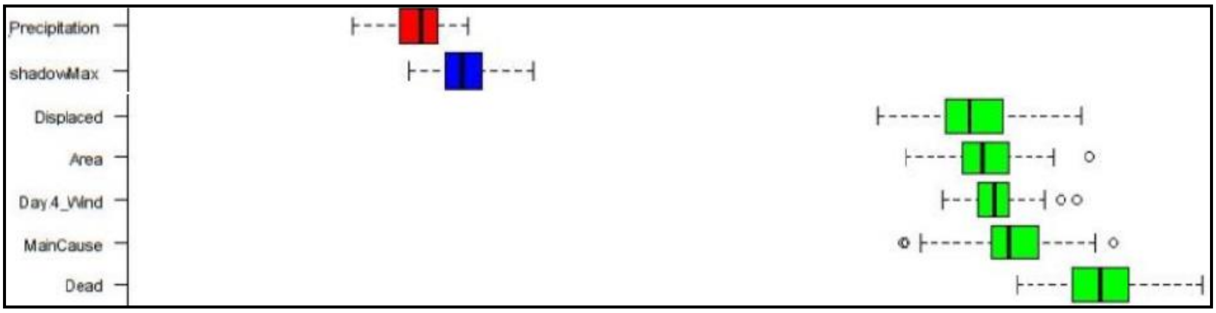


Figure 8: Boruta Feature Selection

Backward feature elimination was used in SPSS and Recursive elimination technique and random forest classifier were used in python. However, the features selected by Boruta gave the highest accuracy when implemented with a basic random forest model. Hence the output of Boruta is considered for implementation. With this, objective 4 is accomplished.

3.6 Class Imbalance

As the output variable is categorical with unequal number of output class shown in Figure 9(a), Synthetic Minority Over-sampling Technique is applied in python and the output distribution gets balanced as shown in Figure 9(b), thereby, accomplishing objective 5. The dataset now has 10,851 records.

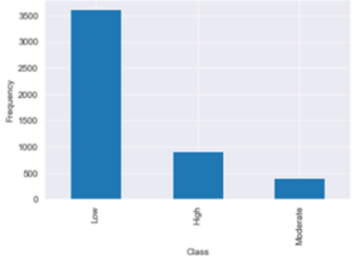


Figure 9(a): Class Imbalance before SMOTE

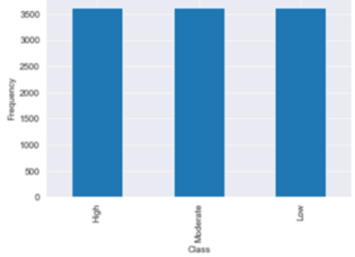


Figure 9(b): Class Imbalance after SMOTE

3.7 Clustering

K-Means clustering technique is used to check if dataset has any clusters. The elbow plot signified 3 clusters as it gets bend at cluster value 3 shown in Figure 10. When the dataset t-Distributed Stochastic Neighbour Embedding is applied, 3 distinct clusters are formed as shown in Figure 11. With this, objective 6 is accomplished.

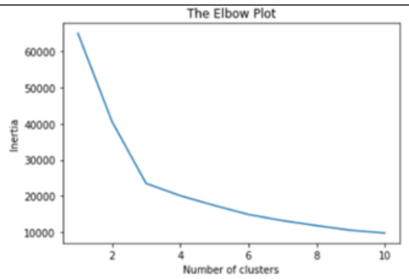


Figure 10: Elbow plot using K-Means Clustering

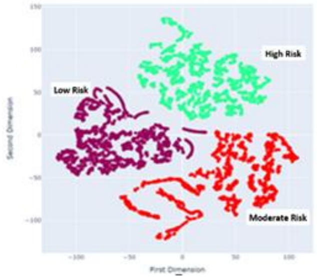


Figure 11: Clustering using t-SNE

3.8 Dimensionality Reduction

Dimensionality reduction technique is applied to reduce the input features to 2-3 components explaining around 95% of variance in the dataset. Below are the techniques used-

3.8.1.Principal Component Analysis

The 2 components formed by PCA do not explain the variance properly. Figure 12(a) shows that the scree-plot is not bent at any point, while Figure 12(b) shows the overlap of the components on each other. Hence, the components are not used in model implementation.

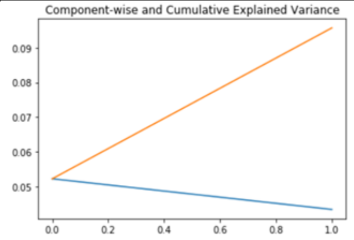


Figure 12(a): Scree plot – PCA

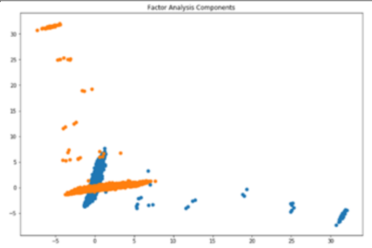


Figure 12(b): Scatter plot – PCA Components

3.8.2.Other Dimensionality Reduction Techniques

Some other techniques are used to reduce the input dimension, such as t-Distributed Stochastic Neighbour Embedding, Singular Value Decomposition, Independent Component Analysis, IsoMap in Figure 13(a), (b), (c), (d) respectively. However, in none of the techniques, the variances are explained properly as the components overlap on each other.

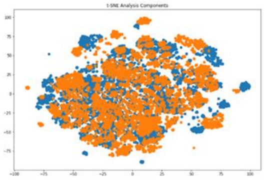


Figure 13(a): tSNE Components

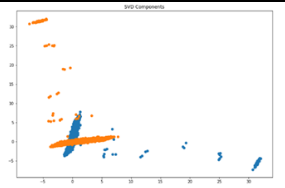


Figure 13(b): SVD Components

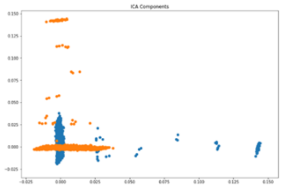


Figure 13(c): ICA Components

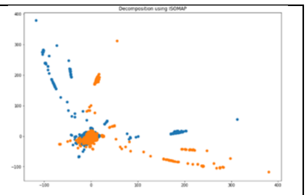


Figure 13(d): IsoMap Components

Hence, the components of dimensionality reduction are not considered. Instead, the original features are used while implementing the classification models. With this, objective 7 is accomplished.

4 Design Specification

Flood severity classification is designed in three phases, namely, Data preparation, modelling and visualization as shown in Figure 14 –

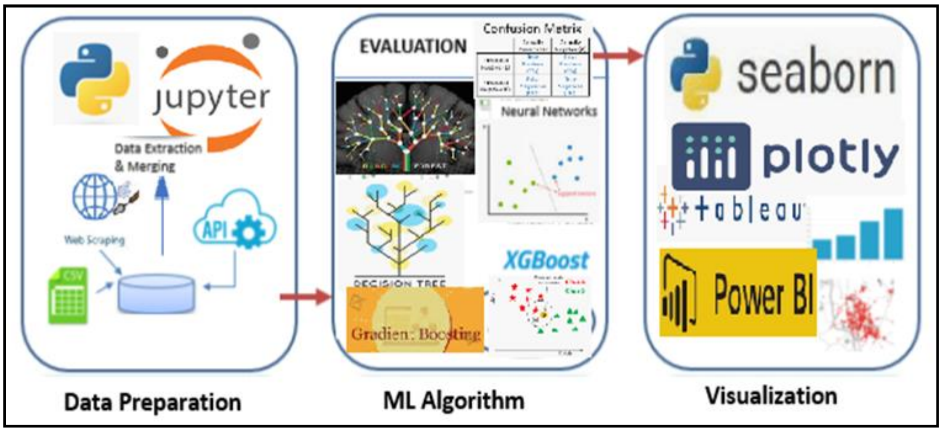


Figure 14: Flood Severity Classification – Design Flow

- Data Preparation phase is explained elaborately in section 3.
- Modelling phase deals with implementation of various classification algorithms such as random forest, decision tree, bagging, boosting etc.
- Visualization phase deals with presentation of the data using graphs and plots.

5 Implementation

This section precisely discusses the implementation of the current research in multiple phases. Traditional machine learning models are applied with 2 non-tree-based and 2 tree-based algorithms. The accuracies are also compared with ensemble techniques and neural networks. Figure 15 shows list of algorithms used in implementation of flood severity prediction.

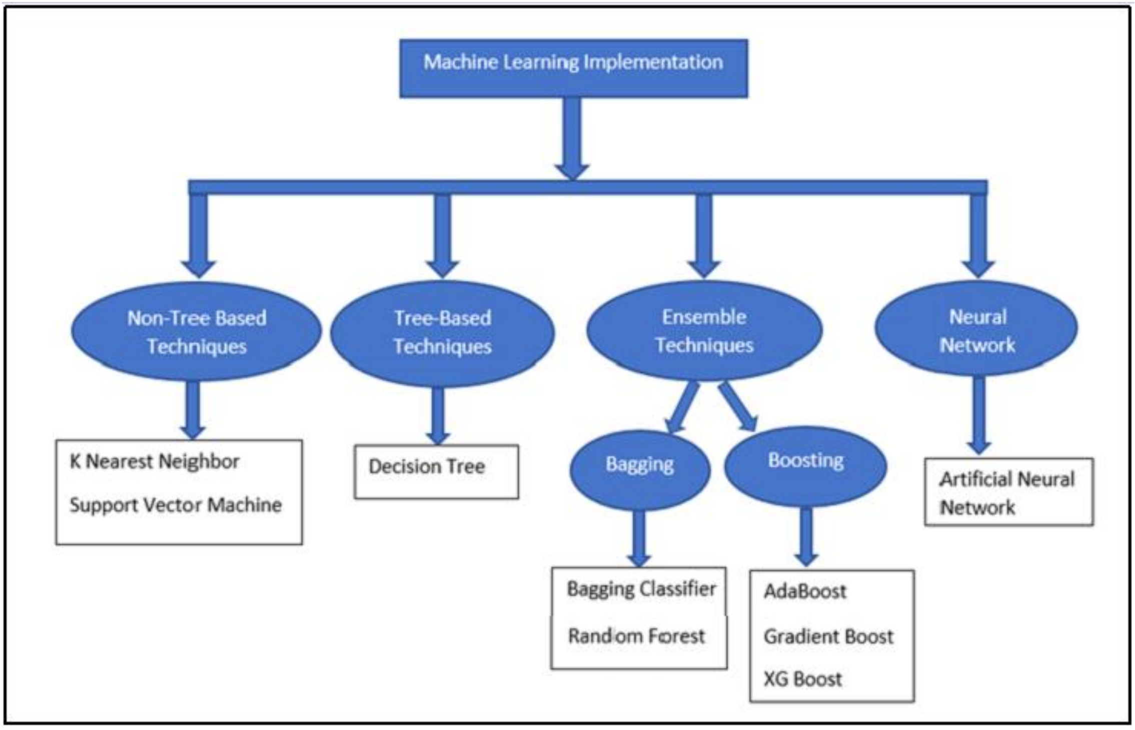


Figure 15: Implementation of Flood Severity Prediction

5.1 Data Preparation

Post feature engineering and standardizing, the final dataset has 10,851 records. Before applying the models, the dataset is split into 80% training and 20% testing by train_test_split function available under sklearn's model_selection library in python. All the models are applied with default parameters. As part of hyperparameter tuning, GridSearchCV function is used and the best parameters obtained are implemented.

5.2 Implementing Machine Learning Models for Flood Prediction

Below are different models used for flood severity predictions available under sklearn library in python-

- **k-Nearest Neighbour**

The kNN is considered as the basic model to predict flood severity as it classifies the output classes based on nearest neighbour. A range of k values are taken in the lazy learning from 1 to 95, as it is the closest value to the square root of the number of records (Lantz, 2020). The model is developed by KNeighborsClassifier function. The best k value obtained is 6 where the model is implemented with rest all default parameters.

- **Support Vector Machine**

For high dimensional spaces, Support Vector machine algorithm is more effective. Hence, SVM is used as there are many input dimensions in the dataset where a range of different parameters used in gridsearch to identify the most efficient combination. SVM model is developed by SVC function and run for regularized parameter(C) value 10, gamma value 0.1, kernel value rbf to get the best result.

- **Decision Tree Classifier**

Decision tree considers all possible outcomes and track each of the branch to conclusion and improves accuracy. Hence, decision tree is considered to in flood classification developed by

DecisionTreeClassifier function with the best set of parameters obtained from gridsearch like criterion as entropy, max_depth value as 29, min_Samples_split as 2.

- **Random Forest Classifier**

Random forest works well with large input parameters and estimates the important variables quite well. Hence, this model is selected as a comparative approach for the above tree-based algorithm developed by RandomForestClassifier function with the best parameter set obtained from gridsearch as n_estimator value 1000 and criterion as gini.

- **Bagging Classifier**

The variance in decision tree classifier is addressed by bagging classifier as it breaks training data into several random subsets where each is used to train the decision tree by averaging all the subsets as aggregate value for prediction. The model is selected to improve the accuracy and developed by BaggingClassifier function. The algorithm is run with default parameters with random_state as 1.

- **AdaBoost and Gradient Boost Classifier**

Adaptive boosting is the simplest algorithm based on multiple sequential models. Gradient Boosting classifier works on forming a strong learner by combining weak learners where the base learner is the regression tree. Each new model in boosting is correction of previous models' error. Models are developed by AdaBoostClassifier and GradientBoostCLassifer functions and run with the default parameters.

- **XGBoost Classifier**

Extreme Gradient Boosting is much faster and advanced gradient boosting technique that improves model performance by including a variety regularization to reduce overfitting. It is used to improve the boosting accuracy developed by XGBClassifier function.

- **Artificial Neural Network**

Artificial Neural Network is used as part of deep learning to extract a better prediction. In python, ANN is implemented using keras library. The model is run with different set of batch sizes and epochs. However, due to high dimension, there was no significant learning in the training loss versus the validation loss as well as in training accuracy versus validation accuracy. This curse of dimensionality is overcome by reducing the number of input parameters, even though with a little less accuracy rate. With this, all the research objectives are accomplished.

6 Evaluation

This section discussed each of the evaluation techniques used in feature selection, dimensionality reduction, model evaluation. Evaluation of the models are done by comparing different parameters of confusion matrix such as Precision, Recall, F1-score and Accuracy that estimates the correctness of the predictions.

6.1 Experiment 1 – Evaluation of Feature Selection

Table 2 shows the number of parameters selected by each feature selection method along with their accuracy when implemented by a sample random forest classifier model. Boruta is considered for further processing as it showed highest accuracy. Even with almost same accuracy, recursive feature elimination technique is not selected because it almost identified all the features as significant.

Table 2: Feature Selection Evaluation

Feature Elimination Methods	Number of Features	Accuracy using Random Forest
Backward Elimination in SPSS	49	67.25%
Recursive Feature Elimination	121	69.30%
Random Forest Classifier	38	68.25%
LMG Classifier	35	68.45%
Boruta in R	**45**	**69.50%**

6.2 Experiment 2 – Evaluation of Dimensionality Reduction

Table 3 shows the accuracy percentage of each dimensionality reduction techniques implemented with kNN, Decision Tree and Random Forest algorithms. Except t-SNE, the components of all the techniques performed dismally. However, the accuracy obtained by the components of tSNE is quite less when compared with the accuracy obtained by the original inputs. Therefore, none of these dimensionality reduction techniques are used in model building.

Table 3: Dimensionality Reduction Evaluation

Dimensionality Reduction Methods	Accuracy using kNN	Accuracy using Decision Tree	Accuracy using Random Forest
PCA Components	48.27%	42.88%	48.91%
t-SNE Analysis Components	**68.44%**	**69.73%**	**63.33%**
SVD Components	45.60%	43.11%	42.60%
ICA Components	45.46%	42.46%	50.11%
ISOMAP Components	52.46%	49.60%	53.01%

6.3 Experiment 3 – Evaluation of Machine Learning Models

Table 4 shows the classification report of the 3 models with highest accuracy percentage. Precision is the total positive predicted values, while recall measures the completeness of the results (Lantz, 2020). The combination of both precision and recall as a measure of model performance is called as F-measure or F1-score (Lantz, 2020).

Table 4: Machine Learning Model Evaluation

Models	Class	Precision	Recall	F1-score
Random Forest Classifier	**High**	**0.82**	**0.73**	0.77
	Low	**0.78**	**0.92**	0.85
	Moderate	**0.88**	**0.83**	0.85
Bagging Classifier	High	0.73	0.77	0.75
	Low	0.78	0.80	0.79
	Moderate	0.85	0.79	0.82

Support Vector Machine	High	0.74	0.66	0.70
	Low	0.74	0.85	0.79
	Moderate	0.80	0.77	0.78

6.3.1. Evaluation of Random Forest Classifier

From table 4, random forest classifier results showed 0.82 precision value for class High, 0.78 for Low and 0.88 for Moderate. This signifies that out of all predictions of class High, 82% of them are actually High. While out of all the predictions of class low and moderate, 78% and 88% of them are actually class low and moderate respectively.

The recall value of 0.73 for class High signifies that out of all the classes that are actually High, 73% are classified as class High. Recall value of 0.92 for class low suggests that out of all the classes that are actually low, 92% are classified as class low. Similarly, 0.83 for class moderate signifies that out of all the classes that are actually moderate, 83% are classified as class moderate.

Results of this model is much better than all of the earlier research's predictions. The F1-score is the harmonic mean of the precision and recall values.

6.3.2. Evaluation of Bagging Classifier

Bagging classifier results showed 0.73 precision value for class High, 0.78 for Low and 0.85 for Moderate. This signifies that out of all predictions of class High, 73% of them are actually High. While out of all the predictions of class low and moderate, 78% and 85% of them are actually class low and moderate respectively.

The recall value of 0.77 for class High signifies that out of all the classes that are actually High, 77% are classified as class High. Recall value of 0.80 for class low suggests that out of all the classes that are actually low, 80% are classified as class low. Similarly, 0.79 for class moderate signifies that out of all the classes that are actually moderate, 79% are classified as class moderate.

6.3.3. Evaluation of Support Vector Classifier

SVC results showed 0.74 precision value for class High, 0.74 for Low and 0.80 for Moderate. This signifies that out of all predictions of class High, 74% of them are actually High. While out of all the predictions of class low and moderate, 74% and 80% of them are actually class low and moderate respectively.

The recall value of 0.66 for class High signifies that out of all the classes that are actually High, 66% are classified as class High. Recall value of 0.85 for class low suggests that out of all the classes that are actually low, 85% are classified as class low. Similarly, 0.77 for class moderate signifies that out of all the classes that are actually moderate, 77% are classified as class moderate.

6.4 Discussion

This study is done to with a novel idea of combining the impact of weather and topography with flood archive to predict the flood intensity unlike earlier flood studies implemented by Khalaf et al. (2018) and Alipour et al. (2020), where only the flood archive data is considered for the prediction. The current research successfully extracted the weather conditions of four consecutive days before flood incident using web scraping and application program interface and merged with the flood archive along with topographic features based on flood date and geographical coordinates. Out of many extracted fields, some of the correlated features are removed as part of pre-processing. Certain important parameters like hydrological and vegetative details are not considered unlike Puttinaovarat and Horkaew (2020) due to unavailability of global application programming interface with these details. Therefore, the study relied on the climatic, topographic and archival data to predict the flood severity. Before implementing machine learning models, data is scaled, and dimensionality reduction is done. The components formed by dimensionality reduction techniques do not perform well. Hence, the original parameter values are considered in model building after applying over-sampling technique to address the class imbalance of output variable.

Out of several classification algorithms, two non-tree-based and two tree-based traditional techniques are implemented. The results are compared with ensemble techniques and neural network. Hyperparameter tuning is done using gridsearch function unlike any earlier studies. Random forest classifier outperformed other models with 83% accuracy, followed by bagging classifier with 79% and support vector machine with 76%. Figure 16 shows accuracy percentage of all the implemented models.

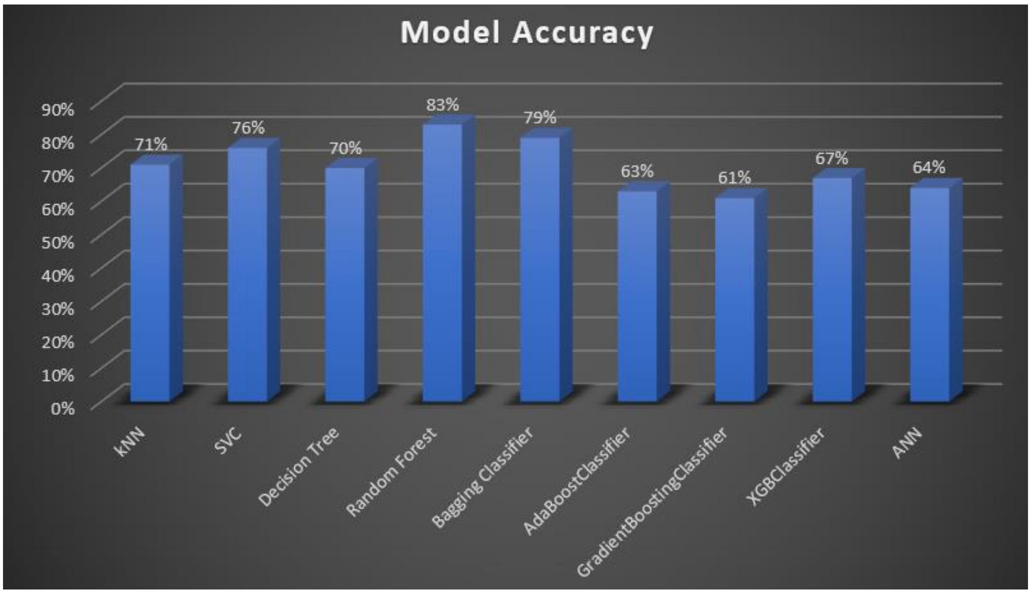


Figure 16: Performance Comparison of Machine Learning Techniques

Ideally, boosting techniques are expected to perform well than random forest and bagging. Yet, the maximum accuracy obtained by XGBoost is 67% which is quite low compared to the bagging accuracy for the used dataset. This is due to overfitting the difficulty of single model. More so, the neural networks also underperformed compared to kNN, SVM and decision tree.

Based on the results of earlier studies, Khalaf et al. (2018) obtained 78% of overall accuracy and Alipour et al. (2020) got 81%, the current research got remarkable result with 83% overall accuracy considering much complex input features. Therefore, unlike the earlier studies, this research outcome based on combining weather, topography and flood history has immense potential in classifying the flood severity which could help the disaster management teams with early severity predictions. One permissible solution to improve the results further is to include other input features like data from department of hydrology, vegetation, crowd sourcing etc.

7 Conclusion and Future Work

In this research, a novel idea of combining the weather and topographic data with the flood archive was implemented to predict the flood severity by classifying the risk as high, low or moderate. The weather data included features such as temperature, dew point, precipitation, wind speed, humidity and topographic data had ground elevation of flood location from sea level. These were merged with flood affected area, duration, number of dead, land displaced. With all these complex data as input features, an efficient machine learning model is built to help in accurate and early flood severity prediction. Random forest proved to be the best suited classifier model in predicting the flood severity giving 83% accuracy, closely followed by bagging classifier with 79% accuracy. These accuracies are confirmed by the high values of precision and recall obtained in respective models. Even though, neural networks showed less accuracies, it can be improved by increasing the hidden layers and epochs. The outcome of this research will help the disaster management teams in accurate prediction of flood severity and take necessary actions to prevent the damage. The study is left open to application of recurrent neural networks in estimating the regression output of flood affected area based on these input features. Also, further analysis is required to estimate the impact of hydrological and vegetative data on flood severity prediction.

Acknowledgement

I would like to thank my supervisors Dr. Paul Stynes and Dr. Pramod Pathak for regularly motivating and supporting me throughout the research. I would like to show my sincere gratitude for answering all my queries and guiding me properly. I would also like to thank my family for support me and my friends for inspiring me.

References

Akshya, J. and Priyadarsini, P. L. K. (2019) ‘A hybrid machine learning approach for classifying aerial images of flood-hit areas’, *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*. IEEE, pp. 1–5. doi: 10.1109/ICCIDS.2019.8862138.

Alipour, A. *et al.* (2020) ‘Leveraging machine learning for predicting flash flood damage in the Southeast US’, *Environmental Research Letters*. IOP Publishing, 15(2). doi: 10.1088/1748-9326/ab6edd.

Boukharouba, K. *et al.* (2013) ‘Flash flood forecasting using Support Vector Regression: An event clustering based approach’, *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2(1). doi: 10.1109/MLSP.2013.6661958.

Chen, J. *et al.* (2020) ‘A machine learning ensemble approach based on random forest and radial basis function neural network for risk evaluation of regional flood disaster: A case study of the yangtze river delta, China’, *International Journal of Environmental Research and Public Health,* 17(1), pp. 1–21. doi: 10.3390/ijerph17010049.

Costache, R., Pham, Q. B., *et al.* (2020) ‘Flash-flood susceptibility assessment using multi-criteria decision making and machine learning supported by remote sensing and GIS techniques’, *Remote Sensing,* 12(1). doi: 10.3390/RS12010106.

Costache, R., Popa, M. C., *et al.* (2020) ‘Spatial predicting of flood potential areas using novel hybridizations of fuzzy decision-making, bivariate statistics, and machine learning’, *Journal of Hydrology*. Elsevier, 585(February), p. 124808. doi: 10.1016/j.jhydrol.2020.124808.

Ding, Y. *et al.* (2019) ‘Spatio-Temporal attention lstm model for flood forecasting’, *Proceedings - 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and So*. IEEE, pp. 458–465. doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00095.

Felix, A. Y. and Sasipraba, T. (2019) ‘Flood Detection Using Gradient Boost Machine Learning Approach’, *Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019*. IEEE, pp. 779–783. doi: 10.1109/ICCIKE47802.2019.9004419.

Furquim, G. *et al.* (2014) ‘Combining wireless sensor networks and machine learning for flash flood nowcasting’, *Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014*, (Section V), pp. 67–72. doi: 10.1109/WAINA.2014.21.

Goel, R. (2020) ‘Flood Damage Analysis Using Machine Learning Techniques’, *Procedia Computer Science*. Elsevier B.V., 173(C), pp. 78–85. doi: 10.1016/j.procs.2020.06.011.

Karyotis, C. *et al.* (2019) ‘Deep learning for flood forecasting and monitoring in urban environments’, *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, pp. 1392–1397. doi: 10.1109/ICMLA.2019.00227.

Khalaf, M. *et al.* (2018) ‘A Data Science Methodology Based on Machine Learning Algorithms for Flood Severity Prediction’, *2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings*. IEEE, pp. 1–8. doi: 10.1109/CEC.2018.8477904.

Khosravi, K. *et al.* (2019) ‘A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods’, *Journal of Hydrology*. Elsevier, 573(March), pp. 311–323. doi: 10.1016/j.jhydrol.2019.03.073.

Lantz, B. (2020) *Machine Learning With R*. doi: 10.4018/978-1-7998-2718-4.ch015.

Mosavi, A., Ozturk, P. and Chau, K. W. (2018) ‘Flood prediction using machine learning models: Literature review’, *Water (Switzerland),* 10(11). doi: 10.3390/w10111536.

Nguyen, D. T. and Chen, S.-T. (2020) ‘Real-Time Probabilistic Flood Forecasting Using Multiple Machine Learning Methods’, *Water*, 12(3), p. 787. doi: 10.3390/w12030787.

Nguyen, T. T. and Le, H. T. T. (2019) ‘Water Level Prediction at TICH-BUI river in Vietnam Using Support Vector Regression’, *Proceedings - International Conference on Machine Learning and Cybernetics*. IEEE, 2019-July, pp. 1–6. doi: 10.1109/ICMLC48188.2019.8949273.

Noymanee, J. and Theeramunkong, T. (2019) ‘Flood Forecasting with Machine Learning Technique on Hydrological Modeling’, *Procedia Computer Science*. Elsevier B.V., 156, pp. 377–386. doi: 10.1016/j.procs.2019.08.214.

Ogale, S. and Srivastava, S. (2019) ‘Modelling and short term forecasting of flash floods in an urban environment’, *25th National Conference on Communications, NCC 2019*. IEEE, pp. 1–6. doi: 10.1109/NCC.2019.8732193.

Opella, J. M. A. and Hernandez, A. A. (2019) ‘Developing a Flood Risk Assessment Using Support Vector Machine and Convolutional Neural Network: A Conceptual Framework’, *Proceedings - 2019 IEEE 15th International Colloquium on Signal Processing and its Applications, CSPA 2019*. IEEE, (March), pp. 260–265. doi: 10.1109/CSPA.2019.8695980.
Puttinaovarat, S. and Horkaew, P. (2020) ‘Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques’, *IEEE Access*. IEEE, 8, pp. 5885–5905. doi: 10.1109/ACCESS.2019.2963819.

Ranit, A. B. and Durge, P. V. (2019) ‘Flood Forecasting by Using Machine Learning’, *Proceedings of the 4th International Conference on Communication and Electronics Systems, ICCES 2019*, (Icces), pp. 166–169. doi: 10.1109/ICCES45898.2019.9002579.

Sachdeva, S., Bhatia, T. and Verma, A. K. (2017) ‘Flood susceptibility mapping using GIS-based support vector machine and particle swarm optimization: A case study in Uttarakhand (India)’, *8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. doi: 10.1109/ICCCNT.2017.8204182.

Shahabi, H. *et al.* (2020) ‘Flood detection and susceptibility mapping using Sentinel-1 remote sensing data and a machine learning approach: Hybrid intelligence of bagging ensemble based on K-Nearest Neighbor classifier’, *Remote Sensing*, 12(2). doi: 10.3390/rs12020266.

Zaji, A. H., Bonakdari, H. and Gharabaghi, B. (2019) ‘Applying upstream satellite signals and a 2-d error minimization algorithm to advance early warning and management of flood water levels and river discharge’, *IEEE Transactions on Geoscience and Remote Sensing*. IEEE, 57(2), pp. 902–910. doi: 10.1109/TGRS.2018.2862640.