Development of LR-Multi Predicting Cross-Validation Model for an Imbalanced Dataset in a Flood Susceptible Area

B.I. Ayinla
University of Ibadan, Ibadan, Nigeria

Akande Oremei C.
University of Ibadan, Ibadan, Nigeria

ABSTRACT

Climate change has a profound impact on human well-being and health. It threatens the fundamental aspects of a good quality of life if not effectively managed. Changes in the frequency and intensity of heavy rainfall events can lead to shifts in the scale and occurrence of river floods, altering how floods happen. However, situations like floods, droughts, and famines raise global concerns. These complex alterations entail calamities and necessitate comprehensive analysis for effective prediction and counteraction. Machine learning algorithms and cross-validation techniques have been employed in the past for flood forecasting to identify patterns from various indicators. While traditional K-FOLD is an effective and commonly used cross-validation technique, the structure of each fold during randomization in terms of convergence and divergence of the dataset is unclear. This research introduces a logistic regression multi-predicting cross-validation (LRMPCV) to address overfitting in imbalanced datasets. The 20,543 tuples of the flooding dataset for Bangladesh from the Kaggle site were used for the experiment. This was divided into two sets, training and test, at a ratio of 80:20%. A Logistic Regression(LR) algorithm checks the distribution of data points for each fold in the three validation techniques during the 10-fold validation processes. Random Forest (RF) and LR models were eventually built from the best folds in each round for prediction. The area under the precision-recall curve (AUPRC) was the critical metric due to data imbalance. The new hybridized model demonstrates a marked improvement when the result is compared with the models built from traditional validation methods. The Random Forest had 99% AUPRC, against the previous result of 84.96% from the traditional KNN and other models. This underscores the power of meticulous model validation in enhancing model selection.

General Terms

Climate change

Keywords

Climate Change, Machine Learning, Prediction Model, Multi-Cross-Validation, Skewed datasets, Random Forest

1. INTRODUCTION

Globally, climatic change is a significant area of concern known for its highly unpredictable nature and, if not adequately monitored, may influence different climatic hazards such as drought, flooding, famine, etc. Climate change is an intricate, multifaceted scientific subject[7]. Climate change describes long-term variations in precipitation and temperature that take place over centuries or millennia[11]. The effects of these changes are flooding, droughts, earthquakes, etc. The impact of climatic changes and the hazards accompanying such changes are more felt in areas susceptible to these hazards, which may lead to loss of lives, as in the case of flooding. People live close to water basins for various reasons, such as wealthy plants, easy access to water for irrigation purposes, drinking, and resort centers. These advantages should not be cut short by flooding. A flood is a rising and overflowing body of water, especially unto normally dry land[12]. Floods are dangerous due to the immense damage to lives and properties. These occurrences account for 84% of all natural disaster deaths worldwide[13]. While climatic changes occur dramatically and change wet and dry seasons, accurately studying climatic changes to understand and predict occurrences like flooding will help humans slowly adapt to these changes and carry out safety measures should the climatic changes give rise to hazards [16]. Several methods, including machine learning, have been employed to predict flooding accurately.

Machine learning algorithms act on data to discover patterns to accurately and efficiently carry out classifications, improvements on existing systems, detections, or future predictions. The K-FOLD Cross-Validation technique is one common and efficient cross-validation technique used to measure an algorithm's performance on a given dataset. It is a technique in which the dataset is randomly split into k-subsets, where each k subset is used as a test set, and other k-1 subsets are used for training purposes. While K-FOLD is widely accepted, it is essential to note that when used as a validation technique on a skewed dataset, the resulting outcome, such as the classification accuracy, although looking good, may be flawed and dangerously misleading[3]. Different variations of the K-Fold have been introduced to solve this overfitting problem that can be introduced by the K-Fold randomization technique on a skewed dataset, like the Multi Predicting Cross-Validation technique and the Stratified K-Fold.

To improve model selection, this study uses logistic Regression and random forest algorithms for model building and compares them against different cross-validation techniques. This study uses a precompiled dataset used by Gauhur, 2022 on GitHub. The flood-predicting model was trained using Random Forest and Logistic Regression and cross-validated using techniques like traditional K-fold, Stratified K-fold, and Repeated K-fold.

To investigate the performance of the model, the auprc (area under the precision-recall curve) score of the model was compared with Gauhar, 2022 and the auprc score was 93.5% as against that of Gauhar, which is 84.96%.

This research further shows that if the validation of a model is given importance, it will increase the chances of model selection.

2. LITERATURE REVIEW

Current studies in the field of flood forecasting and prediction use various machine learning algorithms to harness their ability to identify patterns from historical data[9]. These algorithms include DecisionTree (DT), Random Forest (RF), Linear

Regression (Linreg), Logistic Regression (LR), ExtremeGradient Boosting (XGBoost), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN) have been implemented in flood prediction, yielding reliable results[14].

Prusty et al., conducted a study in 2022 to make predictions for cervical cancer. The stratified k-FOLD was considered since the issue of increased variance is yet to be solved for an imbalanced dataset unless by increasing the value of k (subset division, which increases the computation time). Prusty et al., in 2022, made use of an improvement on this weakness of k-fold: Stratified k-fold cross-validation for predicting cervical cancer. Stratified K-Fold ensures that each subset contains all classes or labels in the dataset[14]. Different machine learning methods (Support vector classification, Random Forest, K-nearest neighbors, and Extreme gradient boosting) were used to build the machine learning model, and stratified k-fold was used for cross-validation. The Random Forest model performed the best[14]. It was found that the model RF6 scored 98.10 percent for Hinselmann, 95.80 percent for Schiller, RF8 scored 97.49 percent for Cytology, and RF9 scored 97.95 percent for Biopsy. This provides various accuracies in various folds for the various types of cervical cancer. This study shows the importance of model validation.

According to Ulker 2022 in "Forecasting Precipitation by Machine Learning Algorithms to Adapt Climate Change," the 30-year precipitation data from Two different climatic regions and cities were utilized to perform prediction. The five most popular regression model algorithms in Python were used to build different models, and the best model was checked to obtain the best prediction of previous years' rainfall. With the proposed model, the precipitation in the coming years could be foreseen, measures could be taken, and the cities could adapt to the coming climate change impacts. The regression methods used were Linear Regression, Decision tree regression, Polynomial Regression, Random Forest regression, and Support Vector Regression. The study showed Random Forest to be more accurate in predicting the precipitation in the two regions for the next five years. In Diyarbakır, it was observed that the precipitation rate will be below average for the next four years.

Similarly, Ladi et al.,2022, in the article "Applications of machine learning and deep learning methods for climate change mitigation and adaptation," explored the most widely used machine learning and deep learning techniques for climate change mitigation and mitigation. The report also identified the most widespread mitigation and adaptation initiatives researched using machine learning and deep learning techniques, emphasizing metropolitan regions. To achieve this, this study used topic modeling and word frequency analysis, specifically the Latent Dirichlet allocation (LDA), as a machine learning technique. According to the findings, artificial neural networks are the most widely used machine learning technology for both reducing the effects of climate change and adapting to them. Geoengineering and land surface temperature are the two climate change adaptation and mitigation research fields that have incorporated machine learning and deep learning algorithms the most.

In the paper titled "Enhancing Data Classification with K-Nearest Neighbors, K-Fold Cross-Validation, and Analytic Hierarchy Process" by Tembusai et al. in 2021, the authors conducted a thorough analysis of the performance of the k-Nearest Neighbors (KNN) method. They employed the K-Fold Cross-Validation algorithm as an evaluation tool and integrated the Analytic Hierarchy Process (AHP) for feature selection in the data classification process. The primary aim was to identify the optimal level of accuracy and machine learning model for their task. The most promising test results were observed in fold-3, where the model achieved an impressive accuracy rate of 95%. However, it's important to note that a potential limitation of this algorithm lies in its suitability for reduced datasets.

Bajpai & He, 2020 in the study titled "Evaluating KNN Performance on WESAD Dataset," explored the effectiveness of K-Nearest Neighbors (KNN) models on the WESAD dataset. The researchers investigated how varying parameters, such as K-fold cross-validation and the number of nearest neighbors, impacted the model's performance using the Python Sklearn library. They aimed to determine the optimal number of nearest neighbors for accurate classification. The paper highlighted a significant finding: altering the number of nearest neighbors led to noticeable changes in the performance of KNN models, regardless of the dataset used. This observation emphasized the importance of balancing achieving optimal performance and managing computational costs. By limiting the number of neighbors, the researchers could control the model's complexity, making deploying resource-constrained devices like Raspberry Pi, multicore microcontrollers, and low-power IoT devices more feasible. These models would be particularly suitable for classifying sensor data in portable embedded systems. However, the study acknowledged a limitation: the WESAD dataset exhibited varying class variances. To address this, the researchers suggested implementing and testing Quadratic Discriminant Analysis (QDA) models on portable embedded devices. Moreover, the proposed KNN model should undergo testing in clinical trials using real-time patient data. In summary, the research shed light on the trade-off between model performance and computational efficiency in KNN-based machine learning applications, paving the way for their potential utilization in practical scenarios involving wearable devices and healthcare applications.

In the article titled "Multiple Predicting K-fold Cross-Validation for Model Selection," authored by Jung in 2018, a novel approach to cross-validation (CV) within the K-fold CV framework was introduced. This innovative method divides the dataset into K subsets or "folds," where one fold is used for constructing the model, and the remaining folds are used for model validation. This process generates predicted values for each observation, which are then averaged to derive a final predicted value. The critical contribution of this approach lies in model selection, which is based on the averaged predicted values. This technique helps mitigate the variation in the assessment process due to averaging. The paper also establishes the variable-selection consistency of this method. Its effectiveness compared to traditional K-fold CV was investigated across various scenarios, including linear, non-linear, and high-dimensional models. The conclusion drawn from this study is that the proposed Multiple Predicting K-fold Cross-Validation (MPCV) method can enhance the traditional K-fold CV. It achieves this by reducing the variability in the validation error, leading to more stable model construction[8].

In the paper "Prediction of Flood in Bangladesh Using k-Nearest Neighbors Algorithm"[4], the research centered on predicting floods in Bangladesh, a country prone to floods, by leveraging the k-nearest neighbors (k-NN) algorithm. The primary goal was to establish a robust flood management system. The study explored various correlation coefficients for selecting essential features to achieve this. Through this approach, the study achieved remarkable results. The KNN machine learning model showcased its effectiveness with a testing accuracy of 94.91%, an average precision of 92.00%, and an average recall of 91.00%. These outcomes signified the model's potential for accurate flood prediction and highlighted its value in contributing to a more effective flood management strategy.

3. METHODOLOGY

This section details the proposed Logistic Regression Multi Predicting Cross-Validation (LRMPCV) to address overfitting in an imbalanced dataset. Multiple models were built from the flood dataset, such as Random Forest, KNN, and Logistic regression models using both traditional and LRMPCV validations. In this study, importance was given to data convergence and divergence in each fold during validation while building the models. The cross-validation techniques: Stratified K-fold, K-fold, and Repeated K-fold were structured to partition the dataset into ten (10-fold) in each round of randomization. The Linear Regression Algorithm (LRA) was employed to check the data point distribution and request for further randomization where necessary. Each model's performance was checked across the three validation methods, and the overall best-performing model was picked for the prediction. The schematic diagram representing the method used in this study is shown below.

The first phase involves data collection and preparation. In the second phase, a prediction model is built with machine learning algorithms (Random Forest and Logistic Regression) and traditional k-fold. The distribution of data points is evaluated for each fold using Logistic Regression. Data points can be switched between folds; for example, if class A is more in fold two than in fold three, and fold three performs better than fold two, some data points can be moved to fold three. The performance metrics are then measured again to see the new data point's impact on that fold's overall performance. The Models were built during validations, and validation test sets were used to examine the performance. The model from the best folds was selected as the final model for prediction using the isolated test set of 20% of the total set. A pictorial representation of the methodology is shown below.

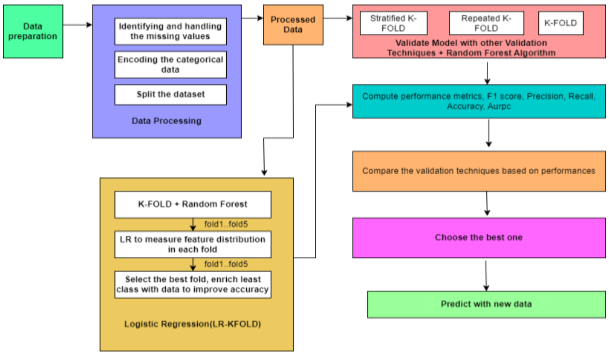

Data preparation
Identifying and handling the missing values
Encoding the categorical data
Split the dataset
Data Processing
Processed Data
Stratified K-FOLD
Repeated K-FOLD
K-FOLD
Validate Model with other Validation Techniques + Random Forest Algorithm
Compute performance metrics, F1 score, Precision, Recall, Accuracy, Aurpc
Compare the validation techniques based on performances
Choose the best one
Predict with new data
K-FOLD + Random Forest
fold1..fold5
LR to measure feature distribution in each fold
fold1..fold5
Select the best fold, enrich least class with data to improve accuracy
Logistic Regression(LR-KFOLD)

Fig 1: Pictorial representation of the proposed methodology

3.1 Data collection and preprocessing

The first phase involves data collection and preparation. The dataset for this study was obtained from GitHub (Gauhar et al., 2021). It is a flooding dataset for Bangladesh. It is 1.97 MB in size and consists of 20,543 rows and 18 columns. The features include Station_Names, Year, Month, Max_Temp, Min_Temp, Rainfall, Relative_Humidity, Wind_Speed, Cloud_Coverage, Bright_Sunshine, Station_Number, X_COR, Y_COR, LATITUDE, LONGITUDE, ALT, Period, and Flood?. The dataset was preprocessed by checking and filling in the missing values and checking duplicate values, and it was also encoded to make it coherent and easy to read by the computer. The preprocessing was carried out using Jupyter-lab on PyCharm IDE. After preprocessing, the distribution of the dataset was seen as shown in Figure 2, showing the imbalanced nature of the dataset, and Figure three shows how correlated the variables are to each other; the closer to 1 and the lighter the shade, the more correlated the features are. As illustrated in Figure 3, the variable Cloud_coverage was plotted against Min_temp, and the map showed a value of 0.82, which means it is more correlated.

3.2 Cross-validation

This project involves building a Flood Prediction model using the Random Forest and Logistic Regression algorithm, but aside from the model construction, it gives importance to model validation. Just like Prusty et al.,2022 in "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," cross-validation was carried out on the prediction model, but instead of using Stratified K-fold alone, K-fold, Stratified K-fold, and Repeated K-fold were used to validate the model. The best fold was chosen based on the evaluation by the Logistic Regression model. The final model was constructed using a Random Forest algorithm.

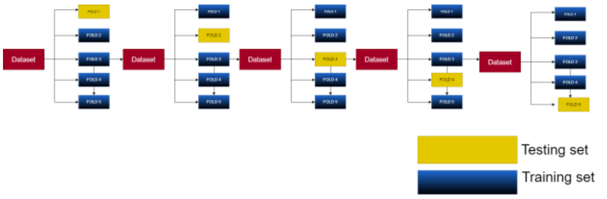


Fig 2: A pictorial description of how k-fold randomly splits a dataset for training and testing

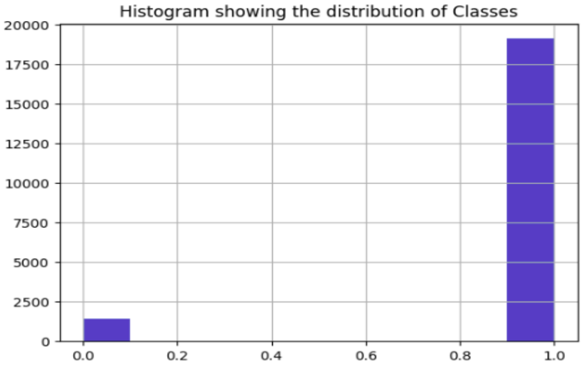


Fig 3: Histogram showing the distribution of classes for flood (1) and no flood (0)

Figure 3 above reveals how skewed the dataset looks, which can be deceptive if models are built using the set without adequate data validation.

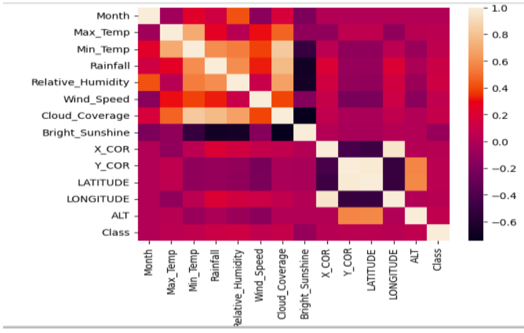


Fig 3: Correlation plot of the variables

Figure 4 shows the level of relationship in the dataset. It simply helps to eliminate any variable that is not

3.3 Evaluation of each fold

The performance metrics of each fold are taken. In this project, we are applying the following metrics:

1. F1 score: The harmonic mean of precision and recall, balancing both metrics.

   F1-Score = 2[(Precision*Recall)/(Precision + Recall)] ……………….…….. (3.1)

   OR

   F1-Score = 2TP/[(2TP)+FP+FN] …………(3.2)

   Where:

   TP = True Positive, FP = False Positive, FN = False Negative

2. Precision score: The number of true positive predictions divided by the total predicted positive instances (true positive and false positive). It measures the model's ability to identify positive instances correctly.

   Precision = TP/TP + FP ………………..(3.3)

3. Recall score(Sensitivity): The number of true positive predictions divided by the total actual positive instances (true positive and false negative). It measures the model's ability to capture all positive instances.

   Recall = TP/TP + FP …………………….(3.4)

4. Area Under the Recall Precision Curve (AURPC): The AURPC measures the model's ability to distinguish between positive and negative instances across various threshold values. It provides an aggregate performance metric.

3.4 Model building and logistic regression cross-validation techniques

After splitting of the data in each fold, the random forest classifier is built. The Python Scikit library was used to build the Random Forest classifier. The model was trained on the training data for that round and validated on the test data for that fold.

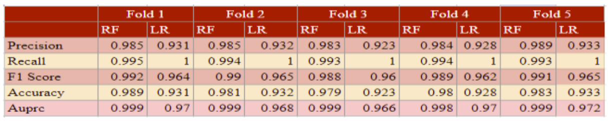
A flood prediction model is built with a Random Forest algorithm in the second phase. The models were then validated using different cross-validation techniques: K-fold, Stratified K-fold, and Repeated K-fold, and the Logistic Regression (LR) was then applied for data point distribution. The performance of each fold was evaluated using the test set of that round. Due to the imbalanced nature of the dataset, more is needed to measure the accuracy of the model alone [1]; the recall, precision, and auprc scores were also measured. The performance of each fold was compared, and the best fold was selected.

This algorithm in Figure 4 shows the step-by-step analysis of the experiment. It simply highlights the steps involved and how processes flow into each other.

4. RESULTS

This section gives insight into the results obtained in this study. When the prediction models (Random Forest and Logistic Regression) were built and validated using K-fold with a split of 5, the performance evaluation is shown in Table 1. Table 2 shows the performance evaluation for the models and validation with Stratified K-fold, and Table 3 shows the performance of the models when validated with Repeated K-fold. In selecting the best-performing fold for each cross-validation technique, priority was given to the auprc score and Logistic regression model as Random Forest generally performed very well across all folds.

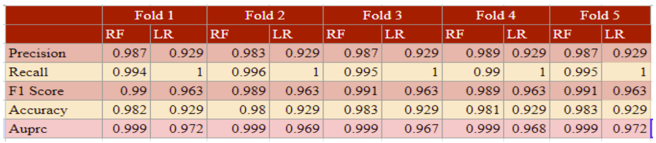
Table 1. Random forest and logistic regression performance on the dataset using k-fold cross-validation



	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR
Precision	0.985	0.931	0.985	0.932	0.983	0.923	0.984	0.928	0.989	0.933
Recall	0.995	1	0.994	1	0.993	1	0.994	1	0.993	1
F1 Score	0.992	0.964	0.99	0.965	0.988	0.96	0.989	0.962	0.991	0.965
Accuracy	0.989	0.931	0.981	0.932	0.979	0.923	0.98	0.928	0.983	0.933
Auprc	0.999	0.97	0.999	0.968	0.999	0.966	0.998	0.97	0.999	0.972

Fold 5 performed best with an auprc score of 0.972.

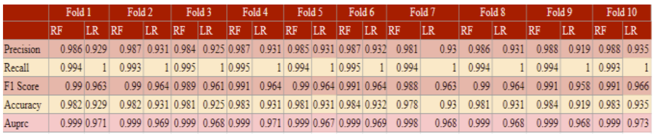
Table 2. Random forest and logistic regression performance on the dataset using stratified k-fold cross-validation



	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR
Precision	0.987	0.929	0.983	0.929	0.987	0.929	0.989	0.929	0.987	0.929
Recall	0.994	1	0.996	1	0.995	1	0.99	1	0.995	1
F1 Score	0.99	0.963	0.989	0.963	0.991	0.963	0.989	0.963	0.991	0.963
Accuracy	0.982	0.929	0.98	0.929	0.983	0.929	0.981	0.929	0.983	0.929
Auprc	0.999	0.972	0.999	0.969	0.999	0.967	0.999	0.968	0.999	0.972

Fold 1 performed best with an auprc score of 0.972.

Table 3. Random forest and logistic regression performance on the dataset using repeated k-fold cross-validation



	Fold 1		Fold 2		Fold 3		Fold 4		Fold 5		Fold 6		Fold 7		Fold 8		Fold 9		Fold 10	
	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR
Precision	0.986	0.929	0.987	0.931	0.984	0.925	0.987	0.931	0.985	0.931	0.987	0.932	0.981	0.93	0.986	0.931	0.988	0.919	0.988	0.935
Recall	0.994	1	0.993	1	0.995	1	0.995	1	0.994	1	0.995	1	0.994	1	0.994	1	0.994	1	0.993	1
F1 Score	0.99	0.963	0.99	0.964	0.989	0.961	0.991	0.964	0.99	0.964	0.991	0.964	0.988	0.963	0.99	0.964	0.991	0.958	0.991	0.966
Accuracy	0.982	0.929	0.982	0.931	0.981	0.925	0.983	0.931	0.981	0.931	0.984	0.932	0.978	0.93	0.981	0.931	0.984	0.919	0.983	0.935
Auprc	0.999	0.971	0.999	0.969	0.999	0.968	0.999	0.971	0.999	0.967	0.999	0.969	0.998	0.968	0.999	0.968	0.999	0.968	0.999	0.973

Fold 10 had the highest performance with an auprc score of 0.973.

Table 4. Shows a summary of the performance evaluation of the three cross-validation techniques using logistic Regression in selecting the best fold and the auprc as the primary metrics for selection.

Metric	K-fold	Stratified k-fold	Repeated K-fold
Precision	0.933	0.929	0.935
Recall	1	1	1
F1 Score	0.965	0.963	0.966
Accuracy	0.933	0.929	0.935
Auprc	0.972	0.972	0.973

Further evaluation for the best-selected fold was done. Figure 5 is a graphical representation of the most important features contributing to the performance.

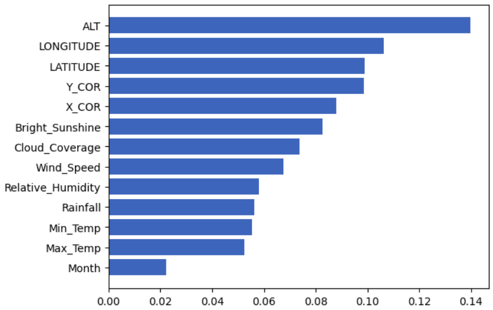


Fig 5: Feature Importance of the best fold in descending order

The outcome of the precision-recall curve is shown in Figure 6, while Figure 7 shows the confusion matrix from the performances of the two metrics.

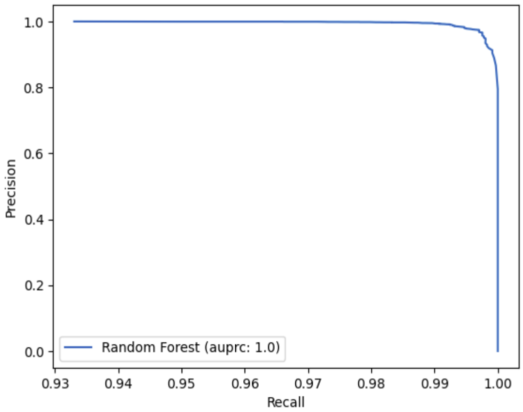


Fig 6: Precision-Recall curve of the best model.

Fig 7: The confusion matrix of the best-performing model.

Figure 8 shows the metrics obtained from Gahur in 2021 with the AUPRC score of 84.96%. as against the AUPRC score of the best-performing algorithm of this research, which was 99.99% and 93.5% for the lowest performing algorithm across all folds for the different cross-validation-techniques.

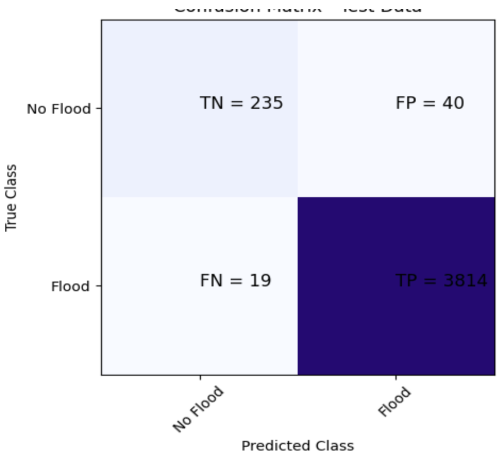
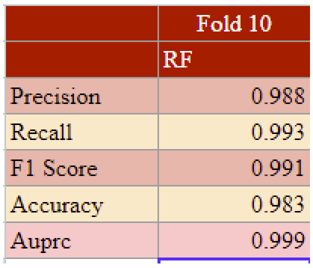


Fig 8: Shows the performance result for Gauha, 2021



	Fold 10
	RF
Precision	0.988
Recall	0.993
F1 Score	0.991
Accuracy	0.983
Auprc	0.999

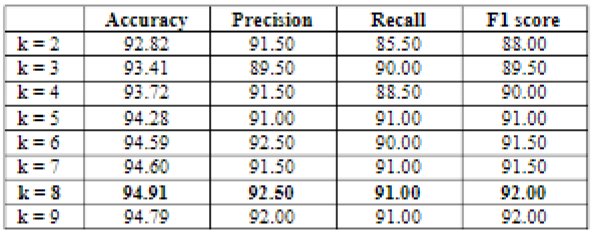
Fig 9: Shows the summary of the performance metrics of the best-performing fold

4.1 Result Discussion

Jung, 2018, in the paper "Multiple predicting K-fold cross-validation for model selection," mentioned that studies have shown that when model validation is done well and given importance as well as model construction is, it could lead to increase model selection. In Gauhar's 2021 paper "Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm," Knn was used as the algorithm for building the flood prediction model, and the best performance was seen when the value of the k-nearest neighbor was 8. This study used the same dataset as Gauhar, 2021 and on exploration of the dataset, the dataset is highly imbalanced and skewed. Brownlee, 2022, in the write-up "Failure of Classification Accuracy for Imbalanced Class Distributions," mentioned that when standard methods are used on an imbalanced dataset, poor results will be obtained, although they may look good. From the study done by Gauhar 2021, the results obtained are shown below:

It is shown that when k=8, there was an accuracy of 94.91%, precision of 92.50%, recall of 91%, and f1 score of 92%. Gauhar 2021 proposed using more advanced machine learning tools to improve the flood prediction model's performance and increase the chances of the model selection for future work. This study used cross-validation alongside the Random Forest algorithm and logistic regression in selecting the best fold. It can be seen that a higher performance was extracted on the dataset, as shown in Figure 4.8, as opposed to what Gauhar got.

These further buttresses the point of Brownlee 2022; to solve this overfitting problem that can be introduced by the K-Fold randomization technique on a skewed dataset, different variations of K-Fold should be introduced like the Multi Predicting Cross-Validation technique and the Stratified K-Fold. The flood-predicting model was trained using Random Forest and Logistic Regression and cross-validated using techniques like traditional K-fold, Stratified K-fold, and Repeated K-fold were employed. The performance of the model of the auprc (area under the precision-recall curve) score of the model was compared with Gauhar, 2022 and the auprc score was 93.5% as against that of Gauhar, which is 84.96%. This further agrees with the point submitted by Brownlee that classification accuracy alone is typically not enough information to make a decision on the best model.



	Accuracy	Precision	Recall	F1 score
k = 2	92.82	91.50	85.50	88.00
k = 3	93.41	89.50	90.00	89.50
k = 4	93.72	91.50	88.50	90.00
k = 5	94.28	91.00	91.00	91.00
k = 6	94.59	92.50	90.00	91.50
k = 7	94.60	91.50	91.00	91.50
k = 8	**94.91**	**92.50**	**91.00**	**92.00**
k = 9	94.79	92.00	91.00	92.00

In "Forecasting Precipitation by Machine Learning Algorithms to Adapt Climate Change"[16], two different climatic regions' precipitation data were used to perform regression model algorithms on 30 years of precipitation data from two cities. The best model obtained was Random Forest, which further validated the outcome of this experiment. The study shows auprc values for Random Forest and Logistic regression models are 99% and 93.5%, respectively.

The use of PCA to determine feature importance during model building, as referred to by Gupta et al.,2022 as a good method to improve the Random Forest model, was fully supported by the results of this study. The result obtained from this study, as depicted in Fig. 4 and Figure 4.8, vividly illustrated the potency of the two techniques.

5. CONCLUSION

In this study, a model was developed using Random Forest and Logistic Regression by carrying out cross-validation on the model to select the best fold in the different cross-validation techniques: Stratified K-fold, K-fold, and Repeated K-fold. The evaluation criteria, including accuracy, precision, recall, and F1-Score, auprc shed light on how well the model performed when compared to Gauhar, 2022 who used KNN for flood prediction on the same dataset as this study. According to the results, the best fold was found in Repeated K-fold in fold ten, thereby improving on the existing model built by Gauhar, 2022(Figure 7) using KNN and identified the model with k nearest neighbor as eight as the best with an auprc score of 84.96% by using cross-validation and Random Forest and Logistic Regression; the auprc score increased to 93.5%. From further look into the feature distribution using descriptive analysis, The experiment could not get any pattern in relation to feature distribution and the performance on each model as the dataset in each fold were closely related. Although from the research on this dataset, the "Alt" feature was the most important feature that contributed to the best fold.

In conclusion, according to Jung, 2017, when model validation is done well and given as much importance as model construction, it could lead to increased model selection.

6. LIMITATION

It is significant to emphasize that the lack of a bigger and more varied dataset is a drawback of this study, which may restrict the generalizability of the results. The lack of an African dataset is also a limitation of this study, particularly in countries like Malawi and Nigeria.

7. RECOMMENDATION

It is recommended that for future work on this study, a bigger and more varied dataset should be explored. African researchers should open their data to use, and the agencies involved should provide access or reliable sources to relevant datasets. Also, other approaches should be explored as changes in climate are not constant, and more improvement is required for prediction.

8. ACKNOWLEDGEMENT

Our thanks go to all who made this paper successful.

```
Step 1: Start
Step 3: Upload the dataset
        df = pd.csv_read("FloodPrediction.csv", index_col=0).
Step 3: Data preprocessing. Handle missing values, label encoding, feature selection, etc.
        cleaned_data = clean(data)
Step 4: Split data into X (features) and y (target)
        y = df["Class"]
        X = df.drop(columns=["Class"])
Step 5: Normalize features
        X_norm = X / X.max()
Step 6: Initialize variables
        kf = KFold(n_splits=5, shuffle=True)
                skf = StratifiedKFold(n_splits=5, shuffle=True)
                rkf = RepeatedKFold(n_repeats=2, n_splits=5)
Step 7: Loop through each fold in K-Fold
        for each fold in kf:
            train_index, test_index = kf.split(X_norm, y)
            X_train, X_val = X_norm[train_index], X_norm[test_index]
            y_train, y_val = y[train_index], y[test_index]
Step 8: Loop through each fold in Stratified K-Fold (similar steps as K-Fold)
Step 9: Loop through each fold in Repeated K-Fold (similar steps as K-Fold)
Step 10: Visualize feature importance for Random Forest model
Step 11: Plot the precision-recall curve for Random Forest, Logistic Regression
Step 12: Visualize feature importance for Logistic Regression model
Step 13: Plot line charts to display most important feature values for each class
Step 14: Choose the best-performing model based on evaluation metrics: auprc, precision, recall
Step 15: Optionally, perform hyperparameter tuning for the selected model
Step 16: Make predictions on unseen data using the selected model
```

Fig 10: Algorithm showing step-by-step implementation of the experiment

9. REFERENCES

[1] Baldini, G., & Geneiatakis, D. (2019). A Performance Evaluation on Distance Measures in KNN for Mobile Malware Detection. In Proceedings of the 2019 International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 193-198). doi: 10.1109/CoDIT.2019.8820510.

[2] Bajpai, D., & He, L. (2020). Evaluating KNN Performance on WESAD Dataset. In 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 60-62). Bhimtal, India. doi: 10.1109/CICN49253.2020.9242568.

[3] Brownlee, J. (2020, January 1). Failure of Classification Accuracy for Imbalanced Class Distributions -

MachineLearningMastery.com. Machine Learning Mastery. Retrieved August 23, 2023, from https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

[4] Gauhar, N., Das, S., & Moury, K. S. (2021). Prediction of Flood in Bangladesh using k-Nearest Neighbors Algorithm. In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 357-361). DHAKA, Bangladesh. doi: 10.1109/ICREST51555.2021.9331199.

[5] Gupta, I., Sharma, V., Kaur, S., & Singh, A. (2022). PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification.

[6] He, J., & Fan, X. (2019). Evaluating the Performance of the K-fold Cross-Validation Approach for Model Selection in Growth Mixture Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 26(1), 66-79. doi: 10.1080/10705511.2018.1500140.

[7] Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

[8] Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. Journal of Nonparametric Statistics. https://doi.org/10.1080/10485252.2017.1404598.

[9] Kim, W. S., & Hong, J. (2022). An Application of Machine Learning Algorithms and a Stacking Ensemble Method for Mass Appraisal of Apartments. Han-Guk Gyeong-Yeong Gonghak Hoeji. https://doi.org/10.35373/kmes.27.2.6

[10] Ladi, T., Jabalameli, S., & Sharifi, A. (2022). Applications of machine learning and deep learning methods for climate change mitigation and adaptation. Environment and Planning B: Urban Analytics and City Science, 49, 239980832210852. https://doi.org/10.1177/23998083221085281.

[11] Lieber, M., Chin-Hong, P., Kelly, K., Dandu, M., & Weiser, S. D. (2022). A systematic review and meta-analysis assessing the impact of droughts, flooding, and climate variability on malnutrition. Global Public Health, 17(1), 68-82. https://doi.org/10.1080/17441692.2020.1860247

[12] Merriam-Webster. (n.d.). Flood. In Merriam-Webster.com dictionary. Retrieved August 24, 2023, from https://www.merriam-webster.com/dictionary/flood

[13] Panahi, M., Jaafari, A., Shirzadi, A., Shahabi, H., Rahmati, O., Omidvar, E., Bui, D., & Lee, S. (2020). Deep learning neural networks for spatially explicit prediction of flash flood probability. Geoscience Frontiers. https://doi.org/10.1016/j.gsf.2020.09.007.

[14] Prusty S, Patnaik S and Dash SK (2022), SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Front. Nanotechnol. 4:972421.doi: 10.3389/fnano.2022.972421

[15] Tembusai, Z. R., Mawengkang, H., & Zarlis, M. (. (2021). K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification. Retrieved from https://www.neliti.com/publications/396954/k-nearest-neighbor-with-k-fold-cross-validation-and-analytic-hierarchy-process-o

[16] Ulker, E. (2022). Forecasting of Precipitation by Machine Learning Algorithms to Adapt Climate Change. Journal of Environmental and Natural Studies, 4(2), 109-118. DOI: https://doi.org/10.53472/jenas.1150975.