Rainfall Prediction in Various Parts of Bhutan With Machine Learning

CSA301 Data Science
Bachelor of Science in Computer Science
(Year III, Semester I)

RESEARCHER(S)

Sherab Tharchen Dorji (12200020)
Ugyen Tenzin (12200033)
Tshewang Dema (12200030)
Sonam Tobden (12200024)

GUIDED BY

Yonten Jamtsho

Gyalpozhing College of Information Technology
Gyalpozhing : Mongar

1 Abstract

Rainfall is one of the most complex and difficult elements of the hydrology cycle to understand and to model due to the complexity of the atmospheric processes that generate rainfall and the tremendous range of variation over a wide range of scales both in space and time. Rainfall is a crucial phenomenon within a climate system, whose chaotic nature has a direct influence on water resource planning, agriculture and biological systems. Machine learning techniques can effectively predict rainfall by extracting the hidden patterns among available features of past weather data. In this project we make use of past data to train machine learning models and then use the best machine learning model to make rainfall prediction. Rainfall prediction would help agricultural industries to keep crops safe and ensure the production of seasonal fruits and vegetables. It is significant for the flood management authorities as more precise and accurate prediction for heavy monsoon rains will keep the authorities alert and focused for an upcoming event that of which the destruction could be minimized by taking precautionary measures. It will also help the people to manage and plan their social activities accordingly. We use four algorithms to train our model i.e., Decision Tree, Random Forest, KNN and SVM using pipeline. It was found that KNN makes more accurate prediction with less MSE and RMSE compared to other algorithms. It can be concluded that KNN algorithm is best suitable for the dataset we used. To make more accurate prediction in the future we recommend using more features and more dataset to train the machine learning model.

Keyboards: Machine learning; Supervised learning; Rainfall Prediction; Accuracy; K-Nearest Neighbors Algorithm (KNN); Mean Square Error (MSE); Root Mean Square Error (RMSE)

2 Introduction

Rainfall prediction remains a serious concern and has attracted the attention of governments, industries, risk management entities, as well as the scientific community. Rainfall is a climatic factor that affects many human activities like agricultural production, construction, power generation, forestry and tourism, among others. To this extent, rainfall prediction is essential since this variable is the one with the highest correlation with adverse natural events such as landslides, flooding, mass movements and avalanches. These incidents have affected society for years. Therefore, having an appropriate approach for rainfall prediction makes it possible to take preventive and mitigation measures for these natural phenomena [1].

Rainfall is a crucial phenomenon within a climate system, whose chaotic nature has a direct influence on water resource planning, agriculture and biological systems. Within finance, the level of rainfall over a period of time is vital for estimating the value of a financial security. Over recent years, scientists' abilities in understanding and predicting rainfall have increased, due to numerous models developed for increasing the accuracy of rainfall prediction [2].

Rainfall is one of the most complex and difficult elements of the hydrology cycle to understand and to model due to the complexity of the atmospheric processes that generate rainfall and the tremendous range of variation over a wide range of scales both in space

and time.

Accuracy of rainfall forecasting has great importance for countries like Bhutan whose economy is largely dependent on hydro-power project and agriculture. Due to dynamic nature of atmosphere, Statistical techniques fail to provide good accuracy for rainfall forecasting. Thus, accurate rainfall prediction is one of the greatest challenges in operational hydrology. On a worldwide scale, large numbers of attempts have been made by different researchers to predict rainfall accurately using various techniques. But due to the nonlinear nature of rainfall, prediction accuracy obtained by these techniques is still below the satisfactory level.

This study is focusing on predicting monthly rainfall using machine learning over some identified places in Bhutan. The rainfall prediction will not just assist in analyzing the changing patterns of rainfall but will also help in organizing the precautionary measures in case of disaster and its management. The rainfall prediction would also assist in planning the policies and strategies to deal with the increasing global issue of ozone depletion. The rainfall prediction could also contribute to the well-being and comfort of the people by keeping them informed by tracking the rainfall patterns and predicting the rainfall using machine learning. The rainfall predictions help people to deal with hot and humid weather. Technological development in the modern world has expanded the space for innovation and revolution. Although the issues concerned are probably associated with these technological advancements, one needs to consider the range of possibilities and opportunities that this technological evolution has opened to human beings.

It is significant for the flood management authorities as more precise and accurate prediction for heavy monsoon rains will keep the authorities alert and focused for an upcoming event that of which the destruction could be minimized by taking precautionary measures. Over the past few decades, major landslide events and water shortage issues were reported over the places in Bhutan. In order to overcome those issues and to reach out for information beforehand, we are going to develop a model using machine learning algorithms where rainfall can be predicted precisely and accurately by studying the datasets of the past 5 years.

The dataset for the study was obtained from the National Center For Hydrology and Meteorology. The dataset provides the intensity of rainfall received by specified places during the period of 2010 - 2021 and the dataset has records of rainfall received daily, monthly and yearly. These parameters had either zero or very few missing values that will be handled during data preprocessing.

The main objective of this project is to predict rainfall more accurately using the best suitable machine learning algorithm.

3 Related Work

3.1 Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan

Rainfall prediction is an important task due to the dependence of many people on it. In this study, we carry out monthly rainfall predictions over Simtokha, a region in the capital of Bhutan, Thimphu. and the Meteorology Department (NCHM) of Bhutan. Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional Long Short Term Memory (BLSTM) this paper proposes a BLSTM-GRU based model which outperforms the existing machine and deep From the six different existing models under study, LSTM recorded the best Mean The proposed BLSTM-GRU model outperformed LSTM by 41.1% model can achieve lower MSE in rainfall prediction systems [3].

The study of deep learning methods for rainfall prediction is presented in this paper, and a BLSTM-GRU-based model is proposed for rainfall prediction over the Simtokha region in Thimphu, 2 value of 0.50), which is widely used for rainfall prediction, did not perform well in comparison to the recent deep learning models on weather station data. 1024 neurons performed better than the others, with an MSE score of 0.013, and a correlation value of 0.90, furthermore, the proposed model presented an improved correlation value of 0.93 and R. Predicting actual rainfall values has become more challenging due to the changing weather patterns.

3.2 Monthly Rainfall Prediction Using Various ML Algorithms for Early Warning of Landslide Occurrence

Natural calamities like landslides cause major human casualties and severe damage to infrastructure and natural resources. Preventing landslides is beyond human capabilities but their impact can be minimized if they can be predicted prior to their occurrence, producing an optimal forecast. Machine learning algorithms for predicting monthly rainfall that monthly forecast are more accurate than weekly or daily forecasts when compared with the actual rainfall data. Hence, the daily rainfall data were converted into monthly months' rainfall data with the past three years' average forecast monthly rainfall well in advance. Based on the results obtained, it can be inferred that BPNN is the best machine learning algorithm for forecasting rainfall followed by LSTM. Coupling the previous three months' rainfall data with the past three years' average rainfall of the targeted month aided the network in between the forecasted rainfall intensity and the optimum predicting the occurrence of rainfall-induced landslides. The study aims at developing a feasible machine learning model for forecasting rainfall which can be used for predicting rainfall-induced landslides. models are capable of predicting low as well as medium-intensity rainfalls effectively, however under performed in mapping high-intensity rainfalls accurately. The predictive accuracy of the models can be increased by introducing other input variables such as humidity, temperature, wind speed etc. The study is conducted specifically for the Narendra Nagar region of Uttarakhand but can be generalized to any area vulnerable to rainfall-induced landslides [4].

4 Methodology

4.1 System Overview

The flowchart below show the steps involved in model training and deployment.

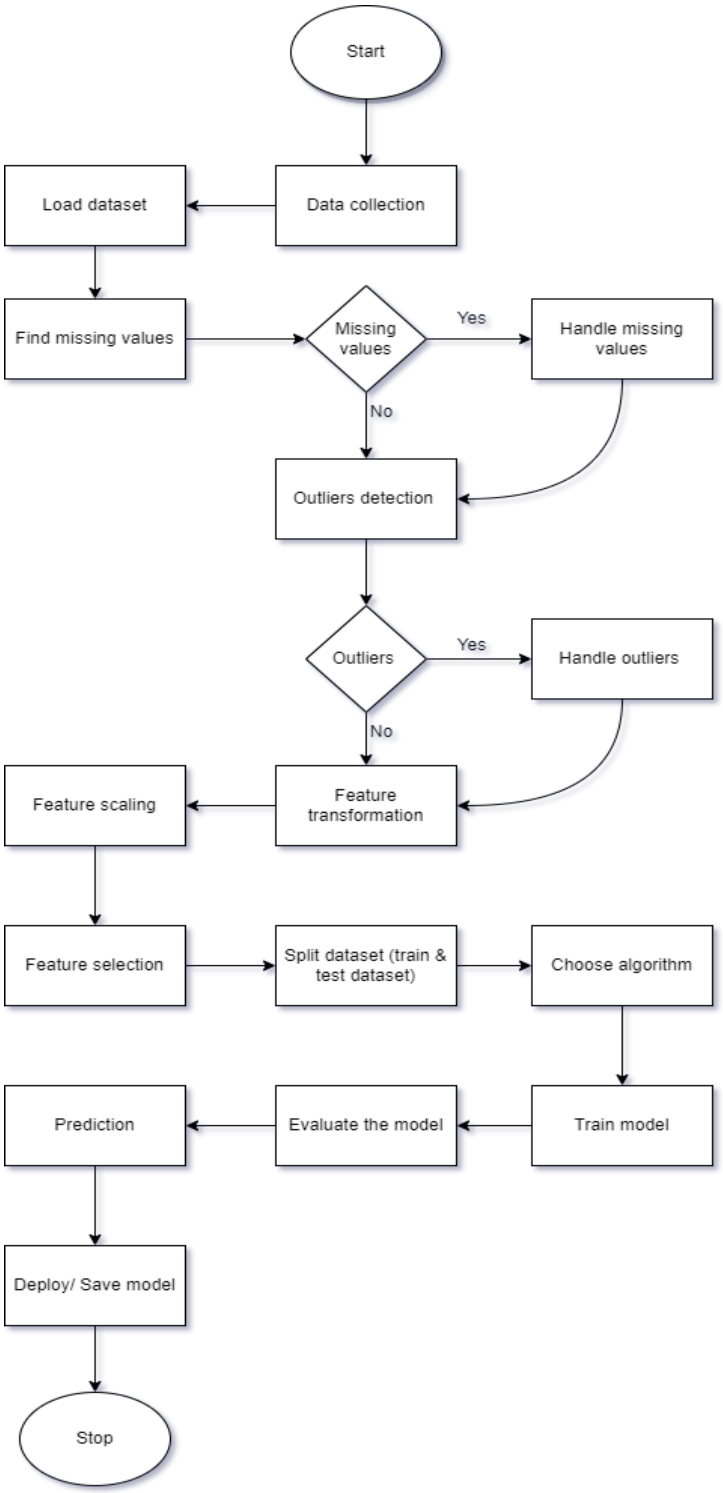

Figure 1: System Overview

4.2 Algorithm

Random Forest(RF)

Random forest (RF) models are machine learning models that make output predictions by combining outcomes from a sequence of regression decision trees. Each tree is constructed independently and depends on a random vector sampled from the input data, with all the trees in the forest having the same distribution.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

RF works by building several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

The RF algorithm involves the following steps:

- In Random forest n number of random data points are taken
- Individual decision trees are constructed for each sample
- Each decision tree will generate an output
- Final output is considered based on Majority Voting or Averaging for Classification and regression respectively

According to the RF algorithm, it is efficient for large datasets and a good experimental result is obtained using large datasets having a large proportion of the data missing.

Advantages of using Random Forest:

- Runs efficiently on a large dataset
- Better accuracy than other classification algorithms.
- It can automatically handle missing values
- It can automatically handle outliers
- Works well with non-linear data
- Lower risk of overfitting
- The algorithm is very stable. If a new data point is introduced in the dataset, the overall algorithm is not impacted much since the new data may affect only one tree, but it is improbable to affect all trees.
- Random forest is comparatively less affected by noise

Decision Tree Algorithm

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

Advantages of using Decision Tree Algorithm:

- Easy to understand and interpret, perfect for visual representation
- Can work with numerical and categorical features
- Requires little data preprocessing: no need for one-hot encoding, dummy variables, and so on
- Non-parametric model
- Fast for inference
- Feature selection happens automatically

KNN (K-Nearest Neighbors) Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems.

It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement.

Advantages of using KNN Algorithm:

- Fast to develop & simple to implementation
- It can be used for both classification and regression problems
- Easy to understand and implement and does not require a training period

Support Vector Machine Algorithm

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outlier detection.

The advantages of support vector machines are Effective in high-dimensional spaces. Still effective in cases where the number of dimensions is greater than the number of samples.

SVMs are sensitive to feature scaling as it takes input data to find the margins around hyperplanes and gets biased for the variance in high values. Although SVMs often work effectively with balanced datasets, they could produce suboptimal results with imbalanced datasets. More specifically, an SVM classifier trained on an imbalanced dataset often produces models which are biased towards the majority class and have low performance on the minority class.

4.3 Dataset

The dataset for this project is collected from the National Center of Meteorology and Hydrology. The dataset contain features such as location, year, month, date, maximum temperature, minimum temperature, relative humidity, wind speed, and rainfall as shown in the table below.

The data was recorded from 2020 to 2021 from 10 different locations. The meteorology station records the values of the environmental variable every day for each year directly from the devices in the station.
The dataset is in CSV format.

Table 1: Features

Parameters	**Measurement unit**
Location	-
Year	-
Month	-
Max Temperature	°C
Min Temperature	°C
Relative Humidity	%
Wind Speed	m/s
Rainfall	mm

4.4 Evaluation Metrics

The trained model can be evaluated using mean square error, root mean square error and R2 score for regression problem.

Mean Squared Error (MSE)
The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.
There is no correct value for MSE. Simply put, the lower the value the better and 0 means the model is perfect. Since there is no correct answer, the MSE's basic value is in selecting one prediction model over another.

Root Mean Square Error (RMSE)
Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.
The closer the value of RMSE is to zero , the better is the Regression Model.

R-Squared or R2 Score
R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

The standards for a good R-Squared reading can be much higher, such as 0.9 or above. In finance, an R-Squared above 0.7 would generally be seen as showing a high level of correlation, whereas a measure below 0.4 would show a low correlation.

4.5 Experimental Setup

Python
Python is a high-level, general-purpose programming language designed by Guido Van Rossum. Its design philosophy emphasizes code readability with the use of significant indentation. Its language constructs an object-oriented approach to help programmers write clear, logical code for small and large-scale projects. This project also uses Python programming language in collaboration with the Django web framework.

Jupyter Notebook
The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience. Jupyter Notebook (formerly known as IPython Notebook) is an interactive way of running Python code in the terminal using the REPL model (Read-Eval-Print-Loop).

Pandas
Panda is a Python package which provides a data structure designed to make working with relational or labelled data both easy and intuitive. In this project, pandas would be used for analyzing and manipulating the tabular data set of rainfall in identified places of Bhutan. Additionally, the Panda can be used in handling missing values, inserting and deleting data, intelligent slicing, indexing, and reshaping

Scikit-learn
Scikit-learn is a software machine learning library for python programming language. It features various classification, regression and clustering algorithms including a support vector machine, random forest, etc.

Matplotlib
It is a cross-platform used for data visualization and graphical plotting library for Python and it is imported as NumPy. It will be used to visualize the data on rainfall for better understanding. This would make things easy and helps to create quality plots while doing the project

Seaborn
Seaborn is a library for making statistical graphics in python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps explore and understand the data. Its plotting operates on data frames and arrays containing whole datasets and internally performs the necessary semantic mapping and statistical aggregation to produce informative plots.

Pickle
Pickle is a useful Python tool that allows users to save their ML models, to minimise lengthy re-training and allow users to share, commit, and re-load pre-trained machine learning models.

5 Results and Discussions

Results

The project model prediction is based on real-time rainfall dataset of the 10 Dzongkhags of Bhutan, granted by NHMC. The dataset used in this project spans over 21 years (2000 to 2021) and consists of 2640 instances and 8 features. First, 7 features are the independent features, which are given as input to the proposed framework in order to predict the 8th feature, which is the output class (dependent feature). The output class indicates how much rainfall will occur in a given month. The dataset is divided into two parts: 80% of the data is reserved for training(2065), and 20% of the data is reserved for testing (517). The activities of the pre-processing stage, including cleaning and normalization, are performed on the rainfall dataset before building a model. To predict, Four machine learning techniques are used: KNN, Decision Tree, Random Forest and SVM. The best model is selected as the best technique to predict the rainfall.

5.1 Train & Test Set Accuracy of Algorithms

The accuracy of each algorithms' train and test set are as shown in figure below. We used pipeline for each algorithms. Random forest algorithm has the highest train set accuracy, followed by SVM, KNN and Decision tree respectively. KNN has the highest test set accuracy, followed by Random forest, Decision tree and SVM respectively. Even though

Table 2: Train & Test Set Accuracy

Algorithm	Decision Tree	Random Forest	KNN	SVM
Train Set Accuracy	72.35	94.13	72.90	90.99
Test Set Accuracy	64.97	69.41	71.34	41.20

the training accuracy of random forest and SVM are higher we cannot consider them as a good algorithm as their test accuracy is relatively low which will cause overfitting in the later part of the prediction. Based on the train and test accuracy results, it can be seen that KNN is the best algorithm for a provided dataset.

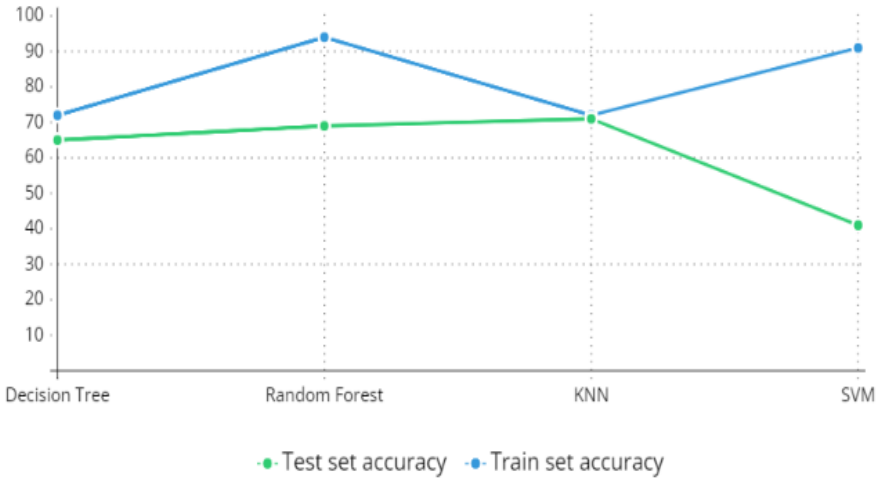

Figure 2: Train & Test Set Accuracy

5.2 Algorithms comparison based on MSE & RMSE

Mean squared error(MSE) aims to determine the average of error squares i.e. the average squared difference between the predicted values and true value.

Root mean squared error (RMSE) indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.

Both MSE and RMSE with lower the value indicates the better model and 0 means the model is perfect.

From the table and figure given below we can conclude that KNN and Random forest have the least MSE (0.2691 and 0.2874 respectively) and RMSE (0.5189 and 0.5361 respectively). It indicates that these two algorithms are make less error while predicting compared to other algorithms.

Table 3: MSE & RMSE

Algorithm	MSE	RMSE
Decision Tree	0.3291	0.5736
Random Forest	0.2874	0.5361
KNN	0.2691	0.5189
SVM	0.5452	0.7383

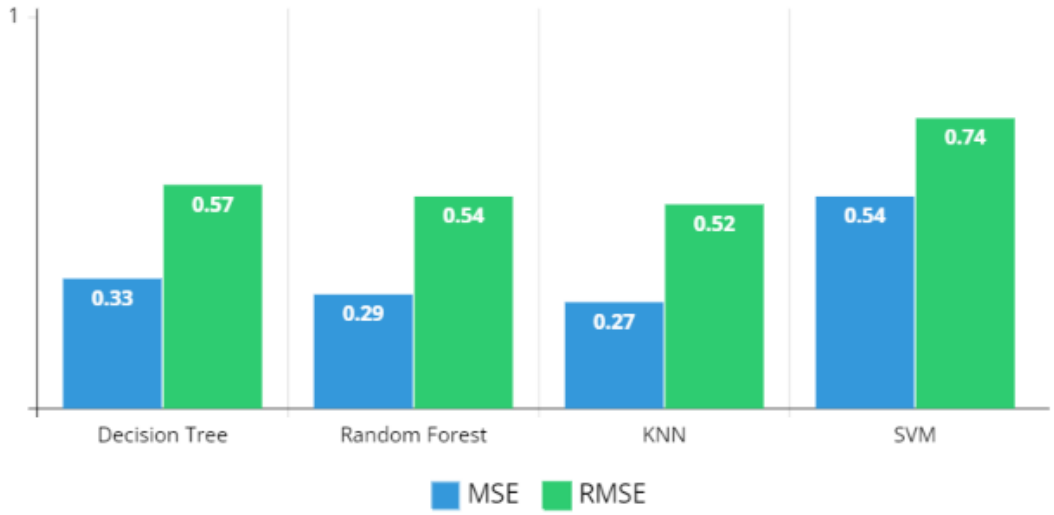

Figure 3: MSE & RMSE

Discussions

After going through all the result analyses, we concluded that KNN is the best machine learning algorithm for rainfall prediction as it has equal train and test accuracy which will avoid overfitting during prediction and it has less MSE and RMSE as compared to other algorithms which helps us to get more accurate results with less error.

6 Conclusion

Comparing all the machines learning models it was found that KNN machine learning model predict more accurate with least MSE and RMSE compared to other machine learning algorithms. Even though KNN algorithm is more accurate and has less MSE and RMSE compared to other algorithm i.e., Decision tree, Random forest and SVM, it has its own limitations. A major drawback of KNN is that it becomes significantly slows as the size of the dataset grows. KNN model has MSE of 0.2691 and RMSE of 0.5289 which is less compared to other algorithm. KNN model have train set accuracy of 72.90% and test set accuracy of 71.34%. Even though some algorithm have high train set accuracy, we can not consider it as a good model as it may lead to overfitting. The KNN model is best suitable machine learning model for the dataset we have used.

Future recommendation, more accurate prediction can be made by adding more features such as moisture-bearing winds, ocean currents, distance inland from the coast and many more and having more training and test data.

7 References

[1] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert Systems with Applications*, vol. 85, pp. 169–181, 2017. [Online]. Available: https://doi.org/10.1016/j.eswa.2017.05.029

[2] N. Oswal, "Predicting Rainfall using Machine Learning Techniques,"Oct. 2019. [Online]. Available: https://arxiv.org/pdf/1910.13827.pdf

[3] M. Chhetri, S. Kumar, P. Pratim Roy, and B.-G. Kim, "Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan," *Remote Sensing*, vol. 12, no. 19, p. 3174, Sep. 2020, doi: 10.3390/rs12193174.

[4] S. Srivastava, N. Anand, S. Sharma, S. Dhar and L. K. Sinha, "Monthly Rainfall Prediction Using Various Machine Learning Algorithms for Early Warning of Landslide Occurrence," *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154184.

[5] V. Kumar, V. K. Yadav, and Er. S. Dubey, "Rainfall Prediction using Machine Learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 5, pp. 2494–2497, May 2022, doi: 10.22214/ijraset.2022.42876.