

# **EARLY WARNING FOR PRE AND POST FLOOD RISK MANAGEMENT**

2021-124

Final (Draft) Report

Vinobaji. S

(IT17181648)

BSc (Hons) in Information Technology Specializing in  
Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

October 2021

## DECLARATION

I declare that this is our work, and this final report does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or institute of higher learning and to the best of our knowledge and belief it does not have any material previously published or written by another person except where the acknowledgment is made in the text.

Name	Registration Number	Signature
Vinobaji. S	IT17181648	S. vinobaji

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor: ..... Date.....

## **ABSTRACT**

Predicting rainfall is one of the most difficult aspects of weather forecasting. Accurate and timely rainfall forecasting can be extremely useful in preparing for ongoing construction, fishing, agricultural, air operations, and flood situations, and other aspects. Due to global warming, deforestation, and other human-influenced factors in nature, heavy rainfall is becoming more unpredictable. In this study, machine learning models were used in three different locations in Sri Lanka that were selected based on city and coastal, countryside, and hill areas. The Sri Lankan meteorological department provides the necessary historical weather data for the prediction model. This system predicts rainfall and rainfall volume ranges (mm) by using machine learning models such as Logistic Regression for binary classification and Support Vector Machine for binary and multi-classification. The primary goal is to compare the accuracy of each location model while using the same attributes such as temperature maximum, relative humidity minimum, and sea level pressure based on the day as the time frame and month used for extra attributes. The rainfall prediction data will be used in the main flood prediction model.

**Keywords – Machine Learning Model, Logistic Regression, Support Vector Machine, Rainfall Prediction, Flood Prediction Model.**

## **ACKNOWLEDGEMENT**

There are many people and things for which I am thankful as I reflect on a year of hard work and dedication to my Research Project studies. First and foremost, we would want to offer our deepest appreciation to my supervisor, Mr. Samantha Rajapaksha, for providing me with the incredible chance to put myself to the ultimate test. You've helped me grow from a fresh graduate student to a seasoned researcher by energizing and assisting me. I've learned a lot from you, both intellectually and emotionally, and it's aided us in getting to where we are today in our studies. Thank you very much for all your help and suggestions with my study. I'd also want to thank the Sri Lankan meteorological service for supplying all the weather data used in this research.

# CONTENTS

<b>DECLARATION.....</b>	<b>2</b>
<b>ABSTRACT.....</b>	<b>3</b>
<b>LIST OF FIGURES.....</b>	<b>7</b>
<b>LIST OF TABLES.....</b>	<b>8</b>
<b>LIST OF ABBREVIATIONS.....</b>	<b>9</b>
<b>1 INTRODUCTION.....</b>	<b>11</b>
<b>1.1 Background &amp; Literature Survey.....</b>	<b>11</b>
<b>1.2 Research Gap.....</b>	<b>19</b>
<b>1.3 Research Problem.....</b>	<b>22</b>
<b>1.4 Research Objectives.....</b>	<b>23</b>
<b>1.4.1 Main Objective.....</b>	<b>23</b>
<b>1.4.2 Specific Objectives.....</b>	<b>23</b>
<b>2 METHODOLOGY.....</b>	<b>24</b>
<b>2.1 Methodology.....</b>	<b>24</b>
<b>2.1.1 System Overview.....</b>	<b>24</b>
<b>2.1.2 Feasibility Study.....</b>	<b>25</b>
<b>2.1.3 Data Collection &amp; Processing.....</b>	<b>27</b>
<b>2.2 Commercialization Aspect of the Product.....</b>	<b>30</b>
<b>2.3 Testing and Implementation.....</b>	<b>31</b>
<b>2.3.1 Testing.....</b>	<b>31</b>
<b>2.3.1.1.....</b>	<b>32</b>
<b>2.3.1.3.....</b>	<b>32</b>
<b>2.3.1.4.....</b>	<b>33</b>
<b>2.3.1.5.....</b>	<b>33</b>
<b>2.3.1.7.....</b>	<b>34</b>
<b>2.3.2 Implementation.....</b>	<b>37</b>
<b>3 RESULTS &amp; DISCUSSION.....</b>	<b>44</b>
<b>3.1 Results.....</b>	<b>44</b>
<b>3.1.1 Models Accuracy Comparison with Data Set Changes - 1.....</b>	<b>45</b>
<b>3.1.2 Model Accuracy Comparison with Data Set Changes - 2.....</b>	<b>46</b>
<b>3.1.3 All Models Accuracy Comparison.....</b>	<b>47</b>

<b>3.2 Research Findings.....</b>	<b>48</b>
<b>3.3 Discussion.....</b>	<b>49</b>
<b>4 BUDGET &amp; JUSTIFICATION .....</b>	<b>50</b>
<b>5 SUMMARY OF STUDENT CONTRIBUTION .....</b>	<b>51</b>
<b>6 CONCLUSIONS .....</b>	<b>52</b>
<b>REFERENCE LIST.....</b>	<b>53</b>
<b>APPENDICES .....</b>	<b>55</b>

## LIST OF FIGURES

Figure 1: System Diagram .....	24
Figure 2: Process Diagram .....	25
Figure 3: Colombo Data .....	28
Figure 4: Katugastota Data .....	29
Figure 5: Vavuniya Data .....	29
Figure 6: Level of Testing .....	31
Figure 7: Firebase Result .....	36
Figure 8: Model Training .....	38
Figure 9: Logistic Function .....	39
Figure 10: SVM Binary & Multi classification .....	40
Figure 11: Prediction Flow Chart .....	41
Figure 12: Weather data input form .....	42
Figure 13: Prediction Page(Sunny Day) .....	43
Figure 14: Prediction Page(Rainy Day) .....	43
Figure 15: SVM (Binary-Colombo) .....	44
Figure 16: LR Katugstota model .....	44
Figure 17: LR Model Comparison with Data Set Changes - 1 .....	45
Figure 18: LR Model Comparison with Data Set Changes - 2 .....	46
Figure 19: All Models Accuracy Comparison .....	47
Figure 20: Data Collection1 .....	55
Figure 21: Data Collection 2 .....	55
Figure 22: Data Collection 3 .....	56

## LIST OF TABLES

Table 1:Research Gap .....	21
Table 2:Attributes .....	28
Table 3:Test Case1 .....	34
Table 4:Test Case2.....	34
Table 5:Budget.....	50
Table 6:Student Contribution.....	51



## LIST OF ABBREVIATIONS

Abbreviations	Description
ML	Machine Learning
LR	Logistic Regression
RP	Rainfall Prediction
FPM	Flood Prediction Model
SVM	Support Vector Machine
RF	Random Forest
WS	Wind Speed
DP	Dew Point
ELM	Extreme Learning Machine
ANN	Artificial Neural Network
MLP	Multilayer Perceptron
DCT	Decision Tree
ET	Extra Tree
MSG	Meta set Second Generation
MLM	Machine Learning Model
MSE	Mean Square Error
RMSE	Root Mean Square Error
KNN	K-Nearest Neighbor
NB	Naive Bayes
CNN	Convolutional Neural Network

LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
ARIMA	Auto-Regressive Integrated Moving Average
GIS	Geographic Information System
AUC	Area Under the Curve
SVR	Support Vector Regression
LSVM	Lagrangian Support Vector Machine
IoT	Internet of Things
API	Application Programming Interface
IDE	Integrated Development Environment

# 1 INTRODUCTION

## 1.1 Background & Literature Survey

Rainfall is one of the natural events that happen in the water cycle. Rainfall is now recognized as one of the primary causes of most major events around the world. Agriculture is mainly dependent on rainfall in Sri Lanka, and it is regarded as one of the most significant variables in determining the country's economy. Apart from that, knowing the quantity of rainfall in a certain region is critical in coastal areas all over the world. To install a rainwater harvester in some of the locations where there is water scarcity, rainfall forecasting should be done ahead of time.

The two monsoon seasons are becoming increasingly well-known in Sri Lanka. The southwestern monsoon brings rain to Sri Lanka's southwest from May to September, while the dry season lasts from December to March. The weather in the country's north and eastern coastal areas is influenced by the north-eastern monsoon, which brings wind and rain between October and January, and dry weather between May and September. There is also an inter-monsoonal period in October and November when rain and thunderstorms are likely across the island.

Rainfall tends to diminish as deforestation progresses, but this is not always the case. The extent, pattern, and placement of land clearance vary. Deforestation in a continuous block reduces the amount of oxygen in the atmosphere. Rainfall is more than the average quantity of woodland clearance in a variety of locations. Although partial deforestation has fewer negative consequences, it still influences temperature and rainfall.

Rainfall prediction is one of the most common predictions in metrology. Rainfall is caused by many factors. However, the contribution of each factor changes with location and time. However, four main factors are most often used to predict rainfall, such as

1. Temperature
2. Relative humidity
3. Sea Level Pressure
4. Wind speed

In some cases, small factors like global warming and deforestation are also able to become considerable factors when it breaks down in long term. It can create abnormal patterns in rainfall. Even though it's not possible to consider too many factors while predicting. So, better to find replacement factors or environments to make the prediction. When looking for replacement factors, a better solution is seven SI base units related factors. Such as

- Length - meter (m)
- Time - second (s)
- Amount of substance - mole (mole)
- Electric current - ampere (A)
- Temperature - Kelvin (K)
- Luminous intensity - candela (cd)
- Mass - kilogram (kg)

Length and Mass base factor location and time base factor Day and month are selected as replacement factors to create a prediction environment Location use as constant so each location creating a different machine learning model and day replacement factor going to use as a time frame for collect other attributes and month attributes going to use as direct replacement factor.

There are many methods used to predict rainfall. Using machine learning (ML) is also one of the methods to predict rainfall. There are a lot of algorithms that can be used to predict rainfall in machine learning. Even though there are a lot of methods to predict rainfall, but not enough accurate methods to predict rainfall volume range. There is a lot of exciting research there to be done to increase the accuracy of rainfall prediction by changing machine learning models and algorithms and some research using a lot of factors to predict rainfall. In machine learning better to use minimum factors and get the best prediction. Here needed to avoid many factors because it will be complicated and costly when using many factors in the machine learning model. So better to use the main three rainfall factors such as temperature, relative humidity, sea level pressure as usual, and month as replacement factors. The project compares machine learning techniques based on location and compares the effect of other factors on rainfall based on the accuracy of each location model, then depicts the most efficient way for rainfall prediction.

Jitendra Shree Mali and colleagues' research article "Rainfall Prediction for Udaipur, Rajasthan using Machine Learning Models Based on Temperature, Vapor Pressure, and Relative Humidity" examines machine learning-based methods to rainfall forecasting in depth. This article discusses current rainfall prediction methods, state-of-the-art methodologies in different rainfall forecast applications, and a collection of commonly used meteorological datasets (temperature, vapor pressure, and relative humidity) from prior research. This article also includes details on experiments that were conducted to assess and compare the performance of meteorological datasets used in rainfall forecasting. They look at the accuracy of different machine learning techniques for rainfall prediction and highlight some future work that must be done with those methods and datasets since this is a survey study (geographical locations and datasets longer than a decade are suggested for more accuracy). The algorithm's running time is too lengthy owing to the increase of datasets, which is one of the existing methods' drawbacks. A review of various rainfall prediction techniques utilizing machine learning and deep learning is presented in this article. This study also discusses the limits of these techniques [1].

Adalya J. Patel's article "Weather Forecasts Based on Rainfall Prediction Using Machine Learning Methodologies" is a good example of this. Analyzes the various levels of accuracy of each machine learning model based on previous research publications. They put several models to the test, including logistic regression (LR), decision trees (DCT), random forest (RF), and support vector machines (SVM) (SVM). The Random Forest model is competent and reasonable for this rainfall prediction. The data utilized as feedback for classification and prediction has a major effect on the accuracy and prediction percentages. All models have benefits and drawbacks, and choosing which model is the best is the most difficult aspect. They discovered that the Random Forest (RF) classification model has a high degree of accuracy and acceptability for their meteorological dataset after evaluating all the supervised learning models mentioned above. The accuracy of the model's predictions may be improved by using a hybrid prediction model, which combines multiple machine learning algorithms. They plan to utilize the hybrid prediction model in future studies to improve accuracy [2].

K. Dutta and P. Gotham's article "Rainfall Prediction Using Machine Learning and Neural Network". The goal of this research is to forecast rainfall using machine learning and neural networks. The research compares machine learning and neural network techniques for rainfall

prediction and then shows which method is the most successful. It is explained how they utilize a machine learning technique to predict rainfall. The accuracy of the machine learning method is determined using two types of errors: MSE and RMSE. Rainfall forecasting may be done in two ways. The first is a LASSO regression-based machine-learning approach. The second method is to use a neural network. To discover an effective approach to forecast rainfall, two methods are being evaluated. The first employs the LASSO regression method, while the second employs an artificial neural network approach. The future development of this project will be a technique of lowering the percentage of mistakes. A decrease in the ratio of train data to test data would also be one of the most important gains [3].

S. B and J. K.S. published "Rainfall Prediction Using Machine Learning Techniques and an Analysis of the Outcomes of These Techniques". The goal of this study is to improve the accuracy of rainfall forecasting by optimizing and integrating Machine Learning techniques. The two main kinds of machine learning techniques that are utilized to build the prediction models in this instance are supervised learning and unsupervised learning. ANN, Logistic Regression, Nave Bayes, and Random Forest are the classification algorithms that are being evaluated and compared. Random Forest produces the most accurate rainfall forecasts when compared to other classification algorithms, with an accuracy of 87.76 percent and the highest recall and F-Measure values. They suggested that the model's prediction accuracy might be enhanced by combining several machine learning methods into a hybrid prediction model [4].

S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz published "Rainfall forecast in Lahore City using data mining methods". The use of different data mining methods for rainfall prediction in Lahore is investigated in this research, which contributes to the body of knowledge. Some of the methods utilized include Support Vector Machine (SVM), Naive Bayes (NB), k Nearest Neighbor (KNN), Decision Tree (DT), and Multilayer Perceptron (MLP). To guarantee accurate prediction, cleaning and normalization methods are used as part of a pre-processing strategy. The performance of data mining algorithms is measured in terms of accuracy, recall, and f-measure using different ratios of training and test data. Ten proportions (10:90-90:10) of training and test data are used to assess the performance of data mining algorithms. In the future, new classification algorithms and climatic features should be tested on a variety of meteorological data to provide more accurate forecasts[13].

I. Cholissodin and S. Sturine's article "Prediction of Rainfall Using Simplified Deep Learning-based Extreme Learning Machines". The exact regression equation model was identified in this research using an ELM-based Simplified Deep Learning prediction system based on the number of layers in the hidden node. The results of this research are expected to lead to the development of an optimal prediction model. The SDLCNN-ELM algorithm merges two types of features, namely the first feature extraction from CNN and the second feature extraction, namely the original features, on rainfall data with a limited number of features, resulting in the majority of the minimum MAD value being more dominant than when using conventional ELM, which only uses the original features. CNN focuses on helping in the identification of deeper underlying patterns that are difficult to quantify or describe using the original features. The SDLCNN-ELM improvement results may reduce errors from the average MAD value as compared to the ELM standard. The SDLCNN-ELM method is a collection of deep neural network families that have been proven to have lower error rates than pure ELM techniques when it comes to forecasting rainfall. This positive outlook on the future will be very helpful in addressing bigger and more complex problems. Future research should focus on using a variety of representative filtering approaches to uncover the hidden features of a characteristic that occurs in all cases and combining the hidden features with features that appear outside to uncover the hidden features of a characteristic that occurs in all cases [12].

S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis published "A comprehensive assessment of seven machine learning techniques for rainfall prediction in weather derivatives,". The main contribution of this study is to show how machine learning techniques, and more broadly intelligent systems, outperform current state-of-the-art rainfall prediction procedures in rainfall derivatives. They utilize Genetic Programming, Support Vector Regression, Radial Basis Neural Networks, M5 Rules, M5 Model trees, and k-Nearest Neighbors to compare the predictive performance of the present state-of-the-art (Markov chain with rainfall prediction) with six other prominent machine learning methods. Machine learning methods exceed the current state-of-the-art, according to this in-depth study. Another element of this research is the finding of links between various climates and forecast accuracy. As a result, the results show that machine learning-based intelligent systems may have a positive effect on forecasting rainfall with high prediction accuracy and minimal cross-climate correlations. Future work in this area will concentrate on increasing coverage by training and predicting on shorter, more specific time

frames, or developing a set of parameters that are more tailored to specific climates, a compromise between the "one-size-fits-all" approach used here and the much more computationally expensive individual tuning approach of tuning parameters for each data set. This kind of trade-off would enable computers to learn from more accurate data, reducing exposure to odd patterns and extreme values while simultaneously improving forecast accuracy [11].

Indrabayu, S. Aditama, A. A. Prayogi, S. Pallu, A. Achmad, and I. S. Areni published "Spatial-temporal method for forecasting rainfall in a tropical nation". To forecast rainfall, this system employs the Extreme Learning Machine (ELM) method, which combines five of the eight kinds of accessible metrological data, including rainfall, temperature, humidity, dew points, and visibility. The architectural model of a neural network on Makassar's weather condition consists of 15 units of input neurons on the input layer, 30 and 40 units of hidden neurons on the hidden layer, and 1 unit output (sunny, light rain, medium rain, and heavy rain) on the output layer. Rain forecasting sampling, such as that used by Indonesia's Badan Meteorologi, Klimatologi dan Geofisika (BMKG), is anticipated to utilize a short-term sample resolution in the future. A greater sample resolution will result in more accurate meteorological element correlation and improved prediction results [10].

M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss published "Improving the accuracy of rainfall rates from optical satellite sensors using machine learning-A random forest-based method applied to MSG SEVIRI". This investigation clouds physical characteristics derived from Meta Set Second Generation (MSG) Spinning Enhanced Visible and Infrared Imager (SEVIRI) data. Each RF model is being investigated to see how it may be modified. Because the amount of information on cloud features changes based on the time of day, daylight, nighttime, and twilight precipitation must all be handled separately. Second, the RF models are trained using the optimum values for the number of trees and randomly selected predictor variables from the tuning research. Finally, the completed RF models are used to predict rainfall rates using an independent validation data set, and the results are compared to co-located rainfall rates collected by a ground radar network. The goal of this research is to see whether the random forests ensemble classification and regression method can help with rain rate prediction throughout the day, night, and twilight (resulting in 24-hour precipitation estimates). To create a final operational retrieval mechanism, additional research will be required in the future. The rainfall rate assignment method will be used with a rain area detection



and process separation strategy. A combined assessment system comprising precipitation detection, process separation, and rainfall rate assignment is proposed and needed [9].

S. Poornima and M. Pushpalatha's article "Prediction of rainfall using enhanced LSTM based recurrent Neural Network with Weighted Linear Units". In this study, the deep learning method for rainfall prediction is utilized. An LSTM-based Intensified Recurrent Neural Network is used to predict rainfall. The proposed method is compared to current approaches such as Holt–Winter, ARIMA, ELM, RNN, and LSTM to demonstrate its improvement in rainfall prediction. Prior datasets must be considered for good accuracy. Because it can retain large quantities of data in its memory and avoids the vanishing gradient, the proposed Intensified LSTM network seems to have better prediction accuracy. The proposed Intensified LSTM increases accuracy somewhat, but only slightly, over current LSTMs. Although the proposed Intensified LSTM increases accuracy by a small margin over the current LSTM model based RNN, the new prediction model retains that accuracy for longer epochs while exhibiting reduced loss, RMSE, and learning rate. Any neural network that reaches an acceptable learning rate with loss depletion in fewer epochs may maintain its performance for any test data in the future without major fluctuations inaccuracy, error, and other performance metrics [8].

R. Abbas, and M. Raziullha's article "Rainfall prediction using a regression model". They utilized linear regression analysis in this research to provide new methods for predicting monthly rainfall. Rainfall predictions are based on the gathering of quantitative data about the current condition of the atmosphere. Several machine learning algorithms can learn complex mappings from inputs to outputs with just a few samples and little computing effort. Accurate rainfall forecasting is difficult due to the dynamic nature of the atmosphere. The difference in meteorological conditions from the previous year must be utilized to forecast future rainfall. They recommend utilizing linear regressions using inputs of temperature, humidity, and wind. This prediction will be very accurate since the researcher's theorized technique predicts rainfall based on historical data for a particular geographic region. The model's accuracy is superior to that of traditional rainfall forecast methods. In the future, the researchers want to use additional deep learning techniques to fix the model's security flaws [7].

S. Zainudin, D. S. Jasim, and A. A. Bakar's article "Comparative study of data mining methods for Malaysian rainfall forecast". Using Malaysian data, this study investigates a range of classifiers for rainfall prediction, including Nave Bayes, Support Vector Machines, Decision Trees, Neural Networks, and Random Forest. Because of their capacity to train on fewer data and forecast a greater quantity of data with a better F-measure, the Decision Tree and Random Forest perform well for rainfall prediction, according to the experimental results. They proposed a strategy for future work. Combining two or more prediction algorithms may enhance the accuracy of the prediction. When there is a direct link between rainfall and flow, using rainfall prediction has a significant influence on forecasting flow. More datasets and exploration of other locations and localities would be beneficial [5].

J. Refonaa, M. Lakshmi, A. C. S. Kiran, and A. R. Teja's article "Rainfall forecast using Apriori Algorithm". In this study, we utilized a variety of machine learning methods to anticipate rainfall, and we used it to forecast rainfall in the Chennai dataset. In this research, the Apriori Algorithm is utilized to display the feature set and determine the probability of rain. After the dataset has been preprocessed, the Apriori Algorithm is used to extract the dataset's features. When compared to previous models, the created model is expected to achieve 95 percent accuracy, potency, and time savings [6].

This proposed research uses machine learning models such as Support Vector Machine and Logistic Regression to predict rainfall using temperature maximum, relative humidity, and sea level pressure as attributes. This study only checks secondary weather factors' contributions in three different locations, such as Colombo (coastal and city), Vavuniya (countryside), and Katugastota (Hills) of Sri Lanka, causing rainfall by using location and time (day, month). At the end of the research, we can conclude a small factor's contribution to rainfall based on the accuracy difference of each location model or prediction error.

## 1.2 Research Gap

When we compared some exiting research to the proposed project

Research Authors	Data used	Models & Algorithms
Jitendra Shreemali, Praveen Galav, Gaurav Kumawat, Pankaj Chittora [4]	Temperature, Vapour pressure, Relative Humidity	Regression, Generalized regression, Linear –AS, LSVM, Random tree, Linear, XGBoost Linear, Tree-
S. B and J. K.S [2]	Temperature, humidity, Sea Level Pressure, Windy	Artificial neural network, Logistic Regression, Naïve Bayes, Random Forest
K. Dutta and P. Gouthaman, [1]	Temperature, humidity, Wind speed, Sunshine Duration	Neural network, LASSO Regression
<i>Adalya J.</i> , [3]	Min Temperature, Max Temperature, Rainfall, Sunshine, Evaporation & Pressure, Humidity, Cloud, Windspeed in the different period	Decision Tree, SVM, LR, RF
S. Zainudin, D. S. Jasim, and A. A. Bakar[5]	Temperature and humidity Mean of 24 hrs	SVM, RF, DCT, ANN, NB
J. Refonaa, M. Lakshmi, A. C. S. Kiran, and A. R. Teja [6]	Temperature and humidity	ANN using APRIORI ALGORITHM
J. Refonaa, M. Lakshmi, R. Abbas, and M. Raziullha[7]	Temperature and humidity	ANN using Linear Regression algorithm

S. Poornima and M. Pushpalatha[8]	Temperature and humidity	Neural Network with Weighted Linear Units
M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss[9]	Cloud top height, Cloud top Temperature, Cloud phase Cloud water path	RF
Indrabayu, S. Aditama, A. A. Prayogi, S. Pallu, A. Achmad, and I. S. Areni [10]	Temperature, humidity, Sea Level Pressure, dew point (DP), visibility, wind speed (WS), and wind direction	Neutral network Extreme Learning Machine (ELM)
S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis[11]	Dry days, longest dry spell Longest wet spell, Mean dry spell, Mean wet spell, Average daily rainfall Average annual rainfall Daily volatility, Highest intensity Median intensity	GP, SVR, RBF, M5F, M5P, KNN
I. Cholissodin and S. Sutrisno[12]	Image	convolution neural network (CNN) SDL-ELM

S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz[13]	Temperature, Atmospheric Pressure (weather station), Atmospheric Pressure (sea level) Pressure Tendency, Relative Humidity, Mean Wind Speed Minimum Temperature, Maximum Temperature, Visibility Dew Point, Temperature	SVM, MLP, DCT, ANN, NB
Proposed Research	Temperature, Humidity, Pressure, Location, Time (day, month)	Logistic Regression, Support vector Machine

*Table 1:Research Gap*

### 1.3 Research Problem

Each natural event is caused by many factors. Rainfall is also caused by many factors, but when we come to consider rainfall prediction using any type of method, including machine learning, we have some limitations such as availability, cost, and time, etc. So, we cannot use a lot of factors to predict rainfall because if we add many factors to the prediction, we need to gather large amounts of data and it will become more complicated, costly, and take a long time to process. But on the other hand, we need to ensure the accuracy of the rainfall prediction. So, we should select "time" base units as the replacement factor, which one can use to cover a wide range of small factors. Time (day, month) and location are natural factors that cover a wide range of small factors. If we select attributes with time and location, we can cover a lot of small factors such as monsoon, sunshine duration, sun direction, deforestation, height above sea level, etc.

Whether historical data availability in Sri Lanka creates a lot of limitations while collecting data for different locations for this research is still unclear. Such as limited attribute types available, data availability with limited time frames, attributes available for limited locations, some attributes not available in some locations, some attributes having considerable missing data in some locations, and limited availability of range-wide data (minimum, maximum, average) within an attribute within a time frame (hours).

When selecting locations, types of locations based on natural or human influence such as coastal, city, countryside, and hills are considered. Attributes were selected based on knowledge of earlier research and attribute range type minimum and maximum were selected because they are considered significant for the constant location with time. The 24hr (day) time frame is used because by selecting a day frame, one can avoid collecting attribute data in 2-time ranges for prediction. In a day frame, the attribute range is significant with constant location and specific time (day, month).

Finally, three locations in Sri Lanka such as Colombo (coastal and city), Vavuniya (countryside), and Katugastota (hills and countryside) were selected and the attributes of temperature maximum, relative humidity minimum, and sea level pressure average were selected for a 24hr time frame from 2015 to 2019.

## **1.4 Research Objectives**

### **1.4.1 Main Objective**

Rainfall is caused by many attributes, such as relative humidity, temperature, sea level pressure, etc. But in machine learning, only the best attributes are used for making rainfall prediction models. Select to find out the most effective data set for predicting rainfall based on machine learning and identify which dataset contributes to predicting flooding and building models for three different locations. By comparing the accuracy of each location model, we calculate the prediction error, which is influenced by other factors.

### **1.4.2 Specific Objectives**

1. Analysis of weather historical data (temperature maximum, relative humidity minimum, sea level pressure average) based on a 24hr (day) time frame and through predicting rainfall, understand each attribute's contribution.
2. Analysis of historical weather data (maximum temperature, minimum relative humidity, and average sea level pressure) for three locations: Colombo (coastal and city), Vavuniya (countryside), and Katugastota (hills) based on a 24hr (day) time frame. By predicting rainfall, we understand each attribute's contribution changes with location.
3. Historical weather data (maximum temperature, minimum relative humidity, and average sea level pressure) for three locations were examined over a 24-hour (day) period. Colombo (coastal and city), Vavuniya (countryside), and Katugastota (hills) are the three major cities in Sri Lanka. By predicting rainfall, we understand each attribute's contribution changes with location.
4. Checking accuracy is different between each model based on location, understanding the other factors influencing it, and selecting the model for predicting rainfall.

## 2 METHODOLOGY

### 2.1 Methodology

#### 2.1.1 System Overview

This rainfall prediction system was developed to predict rainfall, which one day will be used in flood decision-making models and provide predictions for people. Historical weather data is used to build machine learning models and weather data collected from IoT devices is used to predict rainfall.

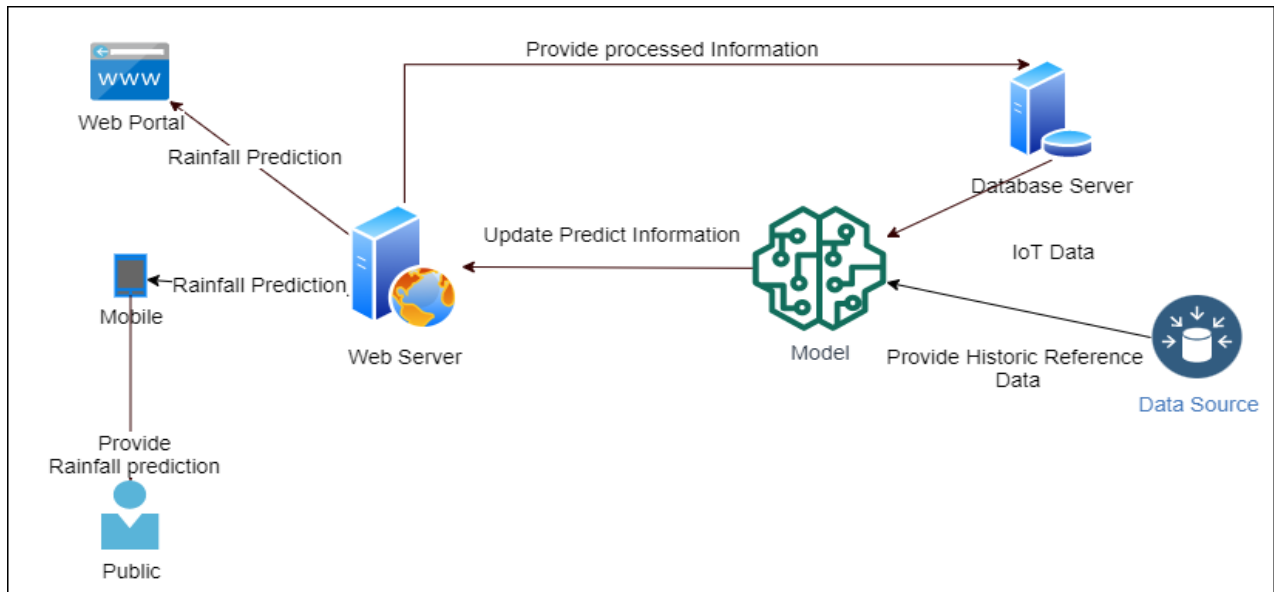


Figure 1: System Diagram

In this system, the machine learning methods of logistic regression and Support Vector Machine will be used to predict whether it will rain or not and the rainfall amount range (mm) by using data sets and checking accuracy differences. After the observation of the prediction accuracy, you can make assumptions and test them with the needed evaluation in the training model again and again. We can draw some conclusions based on the accuracy difference between each data set. At the end of the research, they found out the best dataset for the prediction model and predicted the rainfall and rainfall range.



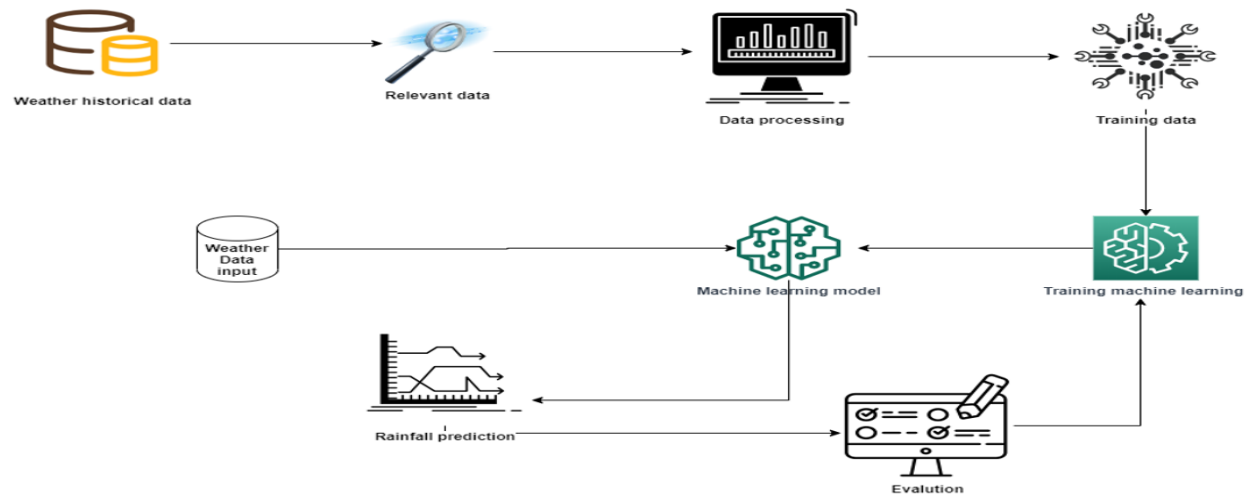


Figure 2:Process Diagram

### 2.1.2 Feasibility Study

A feasibility analysis is a research project or system that assesses a proposed project's or system's feasibility. It is a study or measurement of a software product's usefulness to the firm in terms of practicality. Feasibility studies are carried out for a variety of reasons, including establishing whether a software product is suitable for the company's growth, implementation, and project contribution. Following the completion of the feasibility study, a decision is taken about whether to proceed with the proposed project because it is technically viable or to avoid it because it is feasible to do a new analysis.

Rainfall has grown unpredictable due to the widespread use of machines around the world. Weather forecasters must seek new techniques to improve rainfall prediction, which is the connection that connects the correct place with the right model. To provide an alternate method, the "Location-based rainfall prediction" system predicts rainfall in a manner that is specific to the location. Based on location-based weather data, this algorithm forecasts the "Rain Range." To answer their difficulties, the framework employs machine learning and internet of things (IoT) technologies, which are both popular right now, to properly solve their challenges. The following technologies and tools were utilized to implement the proposed framework because it is an Android application.

### 2.1.2.1 Software Boundaries

- PyCharm



It is a popular IDE for computer programming that is specifically designed for the Python programming language. JetBrains, a Czech business, created and delivered it. Graphic debuggers, code analysis, testing, and version control have all been improved. It is cross-platform, featuring versions for Windows, Mac OS X, and Linux.

- GitLab



GitLab is a web-based DevOps lifecycle solution built by GitLab Inc. that includes a Git repository manager, issue tracking, and continuous integration and deployment pipeline capabilities under an open-source license. Ukrainian developers Dmitriy Zaporozhets and Valery Sizov built the program.

- Flask

The server-side calls in this application are built with Flask. The developed model was put into the flask framework to generate REST API calls.



- Scikit-learn

This is where the logistic regression and SVM algorithm is imported in this project. Techniques for classification, regression, and clustering are all included in this collection.



- Google's Colab

Colab is a cloud-based notebook, like Jupyter Notebook.

Unlike the jupyter notebook, you do not need to install any

dependencies. The colab had all the machine learning libraries installed. All that's left to do now is import and start using the relevant libraries.



- Firebase

For this project, Firebase was used to save prediction data for which one was used for the flood prediction model.



### 2.1.3 Data Collection & Processing

Historical weather data such as maximum temperature, minimum relative humidity, average pressure, and rainfall are collected daily from the meteorology department of Sri Lanka. The historical daily data from the 2015–2019 period was collected from three different locations of Sri Lanka, such as Colombo, Vavuniya, and Katugastota.

It is preferable to gather a large amount of data to improve the accuracy of the output of this rainfall prediction system. With four months of missing data, the collected historical data is substantially organized. Two extra attributes, "Rain\_Or\_Not" and "Rain Range" were created based on the collected rainfall attribute. The "Rain or Not" column was created based on a binary option if it rained (1) or not (0). The "Rain Range" column was created based on multi-classification tags such as (0.01mm as (0), (0.1-10)mm as (1), (10-20)mm as (2), (20-30)mm as (3), and (30+mm as (4).

<b>Attribute</b>	<b>Type</b>	<b>Description</b>
Relative Humidity	Numeric	Rh Min(percentage)
Temperature	Numeric	Max temperature (°C)
Pressure	Numeric	Sea Pressure mean(hpa)
Month	Numeric	month
Rain (targeted value)	Numeric	Boolean: 1 or 0
Rainfall range (targeted value)	Numeric	Ranges(0,1,2,3,4)

Table 2:Attributes

	Station_Name	yy	mm	dd	Tem_Max	Pressure	RH_Min	Rainfall(mm)	Rain_Or_Not	Rain_Range
0	Colombo	2015	1	1	30.3	1009.60	76	0.0	0	0
1	Colombo	2015	1	2	29.9	1011.50	72	0.0	0	0
2	Colombo	2015	1	3	30.2	1012.05	70	0.0	0	0
3	Colombo	2015	1	4	31.2	1011.75	68	1.5	1	1
4	Colombo	2015	1	5	31.0	1010.75	73	0.0	0	0
...	...	...	...	...	...	...	...	...	...	...
1793	Colombo	2019	12	27	31.9	1010.00	64	0.0	0	0
1794	Colombo	2019	12	28	31.2	1010.45	69	0.0	0	0
1795	Colombo	2019	12	29	33.7	1009.85	55	0.0	0	0
1796	Colombo	2019	12	30	32.2	1010.90	65	0.0	0	0
1797	Colombo	2019	12	31	31.1	1011.80	73	0.0	0	0

1798 rows × 10 columns

Figure 3:Colombo Data

	Station_Name	date	yy	mm	dd	Tem_Max	Pressure	RH_Min	Rainfall(mm)	Rain_Or_Not	Rain_Range
0	Katugastota	1/1/2015	2015	1	1	29.1	956.70	79	0.0	0	0
1	Katugastota	2/1/2015	2015	1	2	29.2	958.25	67	0.0	0	0
2	Katugastota	3/1/2015	2015	1	3	31.5	958.75	60	0.0	0	0
3	Katugastota	4/1/2015	2015	1	4	30.3	958.75	69	0.0	0	0
4	Katugastota	5/1/2015	2015	1	5	29.3	958.00	74	0.0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1791	Katugastota	27/12/2019	2019	12	27	29.8	957.05	70	0.0	0	0
1792	Katugastota	28/12/2019	2019	12	28	30.8	957.00	66	0.0	0	0
1793	Katugastota	29/12/2019	2019	12	29	29.7	957.50	63	0.0	0	0
1794	Katugastota	30/12/2019	2019	12	30	30.1	957.55	63	0.0	0	0
1795	Katugastota	31/12/2019	2019	12	31	30.6	958.45	69	0.0	0	0

1796 rows × 11 columns

Figure 4:Katugastota Data

	Station_Name	date	yy	mm	dd	Tem_Max	Pressure	RH_Min	Rainfall(mm)	Rain_Range	Rain_Or_Not
0	Vavuniya	2/1/2015	2015	1	2	31.6	1011.90	73	0.0	0	0
1	Vavuniya	3/1/2015	2015	1	3	32.5	1012.75	76	0.0	0	0
2	Vavuniya	4/1/2015	2015	1	4	32.5	1012.65	72	0.0	0	0
3	Vavuniya	5/1/2015	2015	1	5	32.2	1012.25	71	0.0	0	0
4	Vavuniya	6/1/2015	2015	1	6	31.5	1012.70	72	0.0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1789	Vavuniya	27/12/2019	2019	12	27	30.5	1011.70	70	0.0	0	0
1790	Vavuniya	28/12/2019	2019	12	28	30.4	1011.80	73	0.0	0	0
1791	Vavuniya	29/12/2019	2019	12	29	29.8	1012.20	77	0.0	0	0
1792	Vavuniya	30/12/2019	2019	12	30	30.9	1012.25	72	0.0	0	0
1793	Vavuniya	31/12/2019	2019	12	31	30.3	1013.00	78	0.0	0	0

1794 rows × 11 columns

Figure 5:Vavuniya Data

## **2.2 Commercialization Aspect of the Product**

As a start-up cost, the inventor is developing rainfall prediction models on his dime. Rainfall forecasting software is being developed as a personal effort at no expense to the developers. In a business venture, this cost is accounted for as capital. As a consequence, the venture's commercial operations decide the developers' return. Finally, the development team has decided to commence commercial operations. Many methods for delivering services to government organizations and other third-party customers have been developed, all of which will help to increase profit margins.

Forecasting API access is offered on a monthly or yearly subscription basis. Access to a one-time payment for a previously scheduled IoT gadget purchase. Dashboard data will be integrated into their systems with a subscription-based payment as a contract since the users are the government or customers. End users who have bought and installed Internet of Things (IoT) devices in the field.

The forecast will be accessible on the dashboards of the mobile and web applications for the chosen areas. Commercial and non-commercial users are the two types of end consumers. Commercial customers include government organizations and academics, for example. Researchers are non-commercial users, and nonprofit organizations concentrate on disaster assistance. Developers will be able to recover development expenses by utilizing the methods, while customers will see a return on investment depending on the number of benefits they get from employing system features.

## 2.3 Testing and Implementation

### 2.3.1 Testing

Software testing is a way of establishing the accuracy of software by considering all its attributes (reliability, scalability, portability, reusability, and usability) as well as reviewing the execution of software components to find bugs, mistakes, and defects. As shown in Figure 10, software testing levels provide an objective and impartial assessment of the software as well as a guarantee of its suitability. It comprises putting all components through their paces and providing the required services to determine whether they match the requirements. Testing is required because the software would be dangerous if it failed at some point owing to a lack of testing. As a result, programs can't be released to end-users unless they've been thoroughly tested. Without the assistance of automated technologies, manual testing is a means of evaluating an application's functionality according to the needs of the customer. White box testing, black-box testing, and grey-box testing are the three types.

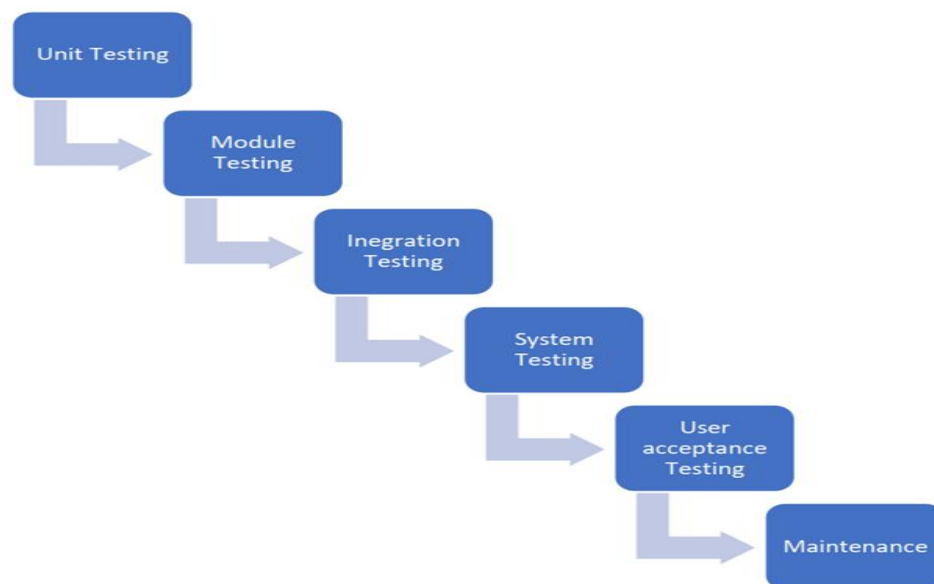


Figure 6:Level of Testing

### **2.3.1.1 Unit Testing**

Unit testing is the process of individually testing each unit or component of a software application. It's the initial step in the process of functional testing. Unit testing is used to confirm that unit components perform as expected. Unit testing is used to confirm that the isolated code is correct. A single application feature or piece of code is referred to as a unit component. This component of the testing is classified as white box testing. The disadvantage of unit testing is that it cannot discover any flaws in an application. Every software application's execution path cannot be assessed. The same can be said for the inspection of individual units.

### **2.3.1.2 Module Testing**

Module testing is a type of software testing in which the individual subprograms, subroutines, classes, or methods of a program are examined. Module testing offers to test the program's smaller building components rather than the full software package at once. As a result, other members of the study team looked through it.

### **2.3.1.3 Integration Testing**

After unit testing, integration testing is the next step in the software testing process. Units or independent parts of the program are examined as a group during this testing. The goal of integration testing is to find flaws that occur when integrated components or units interact. While all the components or modules are functioning independently, integration testing is done to check the data flow between them.



#### **2.3.1.4 System Testing**

System testing included the testing of a fully integrated software framework. The software is built-in modules that are then coupled with other software and hardware to make a complete computer system. System testing is a collection of tests meant to put an integrated software computer system through its paces and compare it to its specifications.

#### **2.3.1.5 User Acceptance Testing**

Acceptance testing is a type of formal testing that is carried out according to user requirements and feature processing. It determines whether the program complies with the requirements as well as the needs of the users. It's done similarly to Black Box testing, with a certain number of users required to determine the system's acceptability level. This monitoring was the responsibility of the client or end-user. It will be put to the test at the end of the testing process. Target customers that are searching for user happiness supply the reviews and user experience.

#### **2.3.1.6 Maintenance Testing**

This SDLC paradigm concludes with the maintenance phase. This is a vital component since the system is subject to changes during the software life cycle. Some of the functions carried out throughout this procedure include software upgrades, fixes, and improvements for the program.

All the testing phases for the created framework passed with flying colors. The entire gadget should be separated during the testing procedure, which would be a feasible approach to testing.

- The frontend of a mobile application
  - Test the Android application to ensure that it runs without errors.
- Backend API for the server
  - The developed machine learning model is put to the test to improve its accuracy.

### 2.3.1.7 Test Cases(Front end)

Test case ID	001
Test case scenario	Validate input field
Test steps	<ul style="list-style-type: none"><li>a. Users navigate to the prediction page</li><li>b. Fill required fields</li><li>c. Submit</li></ul>
Test data	data
Expected result	Re-direct to the prediction page screen
Actual result	As expected,
Pass/Fail	Pass

Table 3:Test Case1

Test case ID	002
Test case scenario	Validate input field
Test steps	<ul style="list-style-type: none"><li>a. Users navigate to the prediction page</li><li>b. Fill required fields except for the month field</li><li>c. Submit</li></ul>
Test data	data
Expected result	Re-direct to the login screen
Actual result	As expected
Pass/Fail	Pass

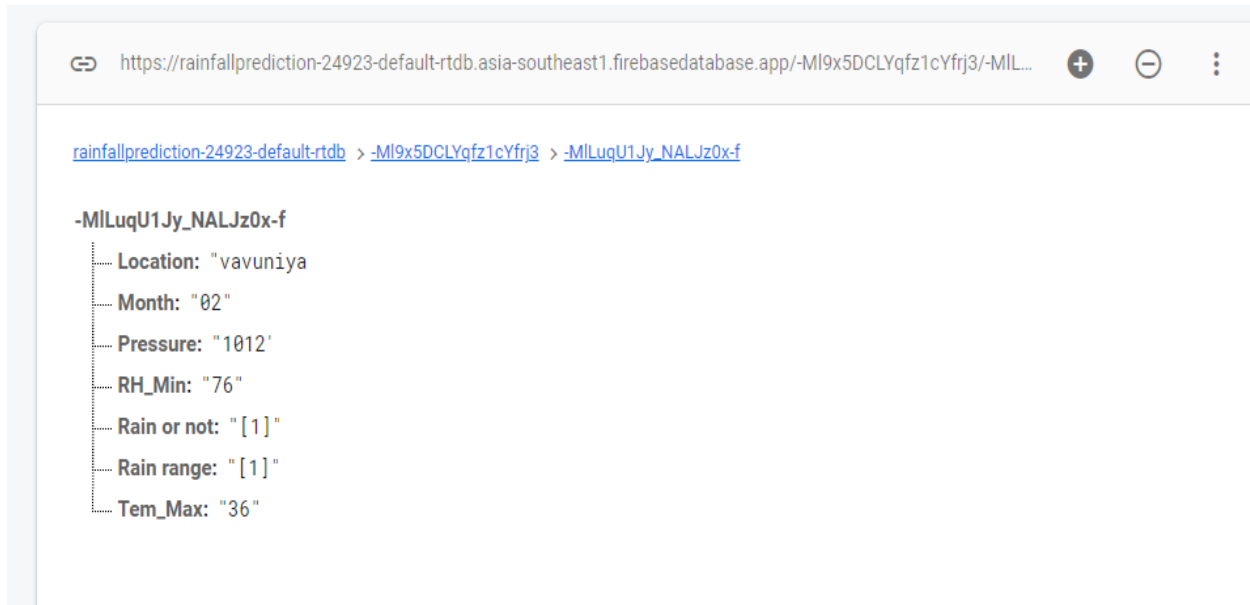
Table 4:Test Case2

Test case ID	003
Test case scenario	Validate input field
Test steps	<ol style="list-style-type: none"> <li>Users navigate to the prediction page</li> <li>Fill required fields expect the max temperature</li> <li>Submit</li> </ol>
Test data	data
Expected result	Re-direct to the login screen
Actual result	As expected
Pass/Fail	Pass

Table 5:Test Case3

Test case ID	004
Test case scenario	Validate input field
Test steps	<ol style="list-style-type: none"> <li>Users navigate to the prediction page</li> <li>Fill required fields expect minimum relative humidity</li> <li>Submit</li> </ol>
Test data	data
Expected result	Re-direct to the login screen
Actual result	As expected
Pass/Fail	Pass

Table 6:Test Case4



*Figure 7: Firebase Result*

### **2.3.2 Implementation**

The proposed framework was developed as a web application. This rainfall prediction system is where we input weather data manually. The model predicts whether there will be rainfall or not, and the rainfall amount ranges. Individuals can enter relevant weather data and receive a prediction.

#### **2.3.2.1 Data Used**

This research mainly considers two main data sets as the daily basis for predict rainfall. Such as

Data set 1

1. Temperature maximum
2. Relative humidity minimum
3. Sea level pressure Average

Data set 2

1. Dataset 1
2. Time (Month)
3. Locations (Colombo(Coastal and City), Vavuniya (Countryside), and Katugastota(Hills))

Data set one is the main factor that is causing rainfall and that is already being used in exciting research. Based on location, data set two, data set one, and replacement factor time (month) is used. This weather historical data for the past 5 years was collected from the metrology department of Sri Lanka.

### 2.3.2.2 Train the Model

This is a crucial stage of the study because the results of this phase have a big impact on the system's output. During this technique, we must divide the dataset into two pieces. Analyzed data is split into 20% test data and 80% trained data used in the trained model. Accuracy checks with each attribute change and understands the attribute's contribution. The right answer must be included in the training data, often known as an aim or target attribute. The learning algorithm looks for patterns in the training data that map the input data attributes to the goal and then produces a machine learning model that captures these patterns. We used two machine learning algorithms to train the model. As an example,

- Logistic Regression(LR)
- Support Vector Machine(SVM)

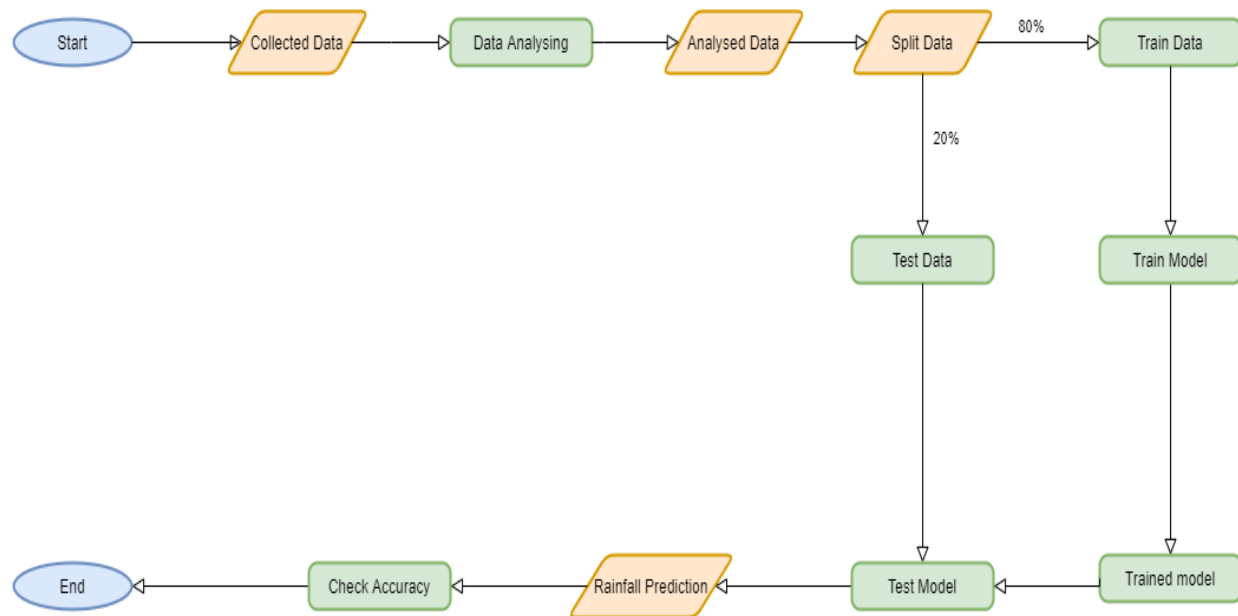


Figure 8: Model Training

### 2.3.2.3 Logistic Regression

The technique of logistic regression is one of the ways of doing a regression analysis on binary input parameters. The sigmoidal function, which is at the basis of the entire approach, is the focus of the logistic regression function represented in Fig 9. It uses an S-shaped format to map out the features of any given input data between 0 and 1. The LR model is a linear one that employs a sigmoid function, as indicated in equation (1). This equation is used to undertake classifications with results ranging from 0 to 1, giving the LR the capacity to do the probabilistic interpretation. In this proposed system, logistic regression is used to predict for binary classification whether it will rain or not.

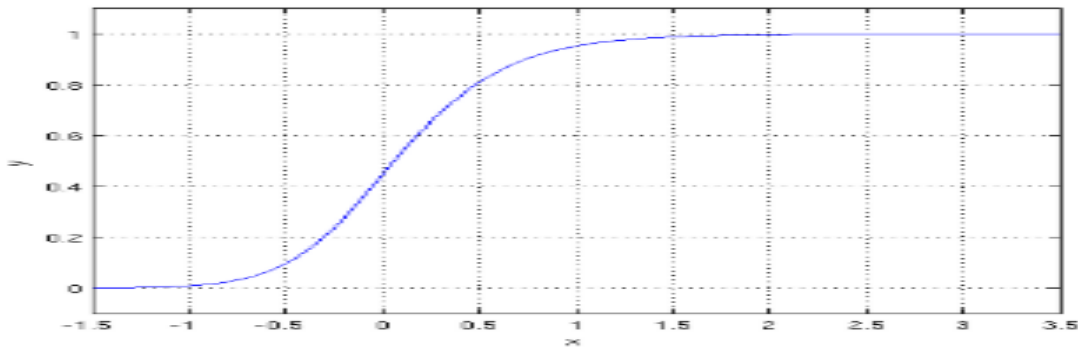


Figure 9: Logistic Function

$$F(x) = \frac{1}{(1 + e^{-x})}$$

Equation1: Sigmoid Function

In most cases, the regression coefficients are evaluated using maximum probability estimation. Most of the writers used statistical models to predict rainfall, such as univariate or multivariate binary logistic regression.

### 2.3.2.4 Support Vector Machine

The SVM model was created in 1995 by Cortes and Vapnik. The SVM model is a well-known model that uses a hyperplane to distinguish two classes. The best element of the SVM model is the kernel function, which converts primary input into high-dimensional data and then finds a hyperplane to separate the groups. As a result, SVM is a useful technique for categorization. Furthermore, bit determination has a major impact on categorization execution. It's difficult to decipher the RBF and other non-linear kernels. The linear kernel does not provide a great prediction for non-separable datasets, either.

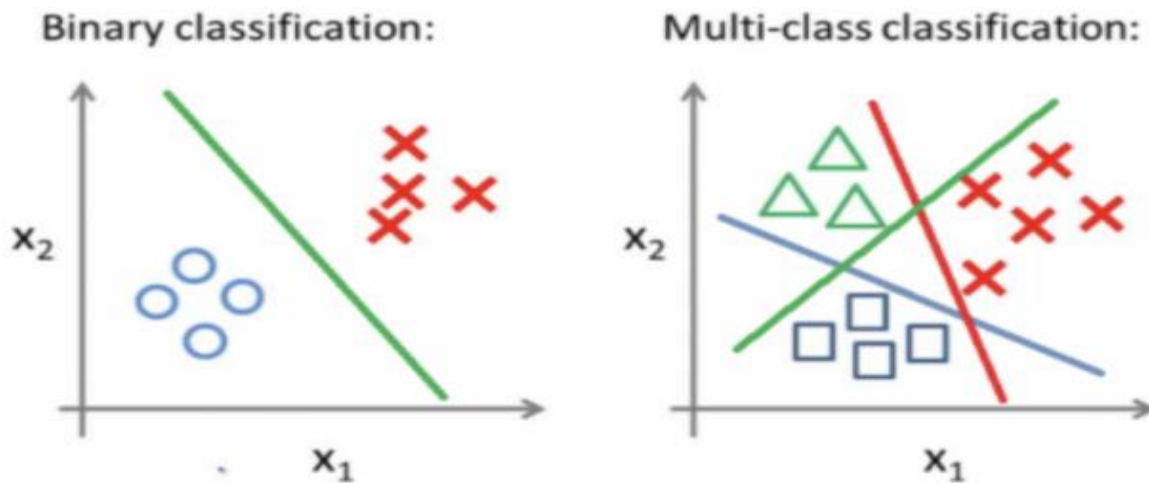


Figure 10:SVM Binary & Multi classification

In this system, SVM is used to predict for binary classification “rain or not” behavior and for multi-classification divide, the rain range into five categories (No Rain (0.1mm), Drizzle (0.1-10mm), Normal (10-20) mm), Strong (20-30) mm), and Heavy (30+ mm)).



### 2.3.2.5 Rainfall Prediction logical view

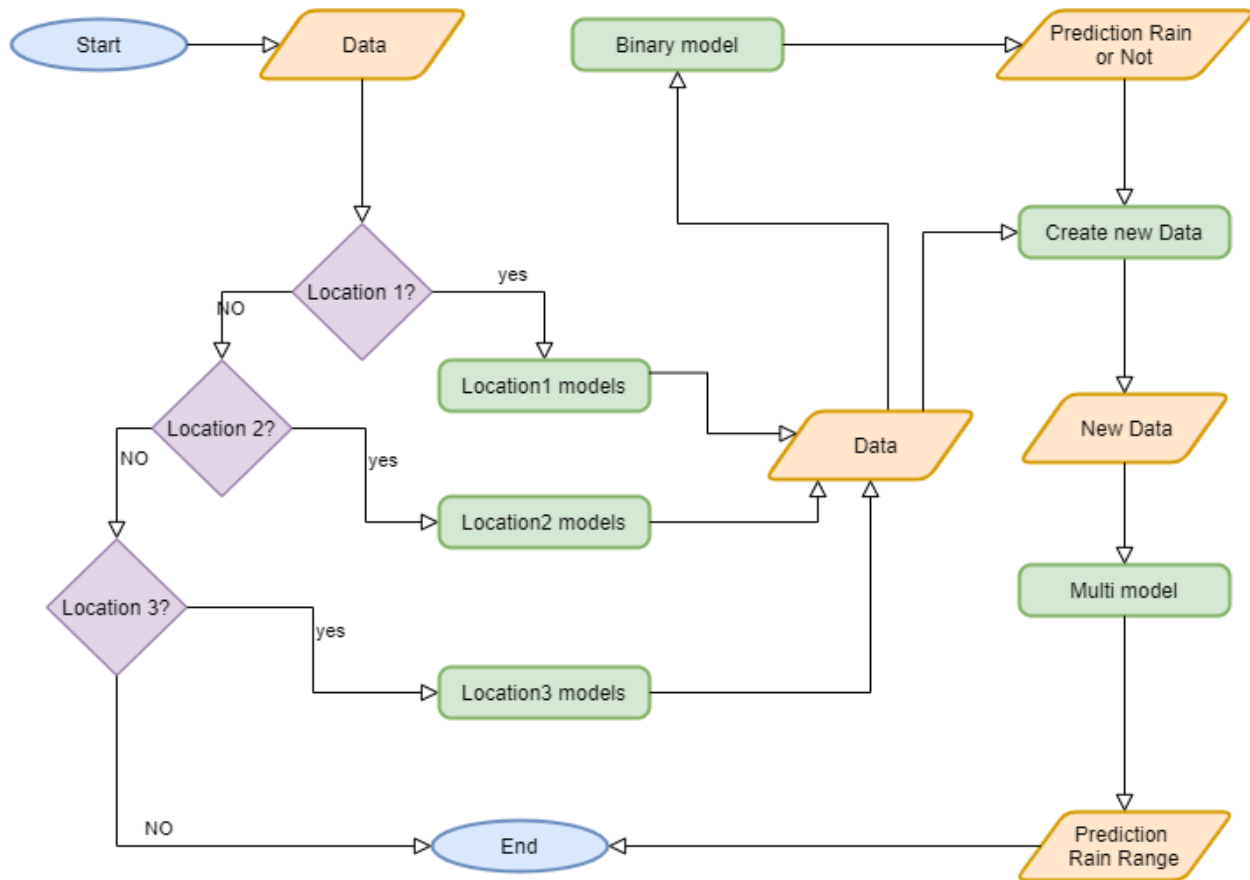


Figure 11: Prediction Flow Chart

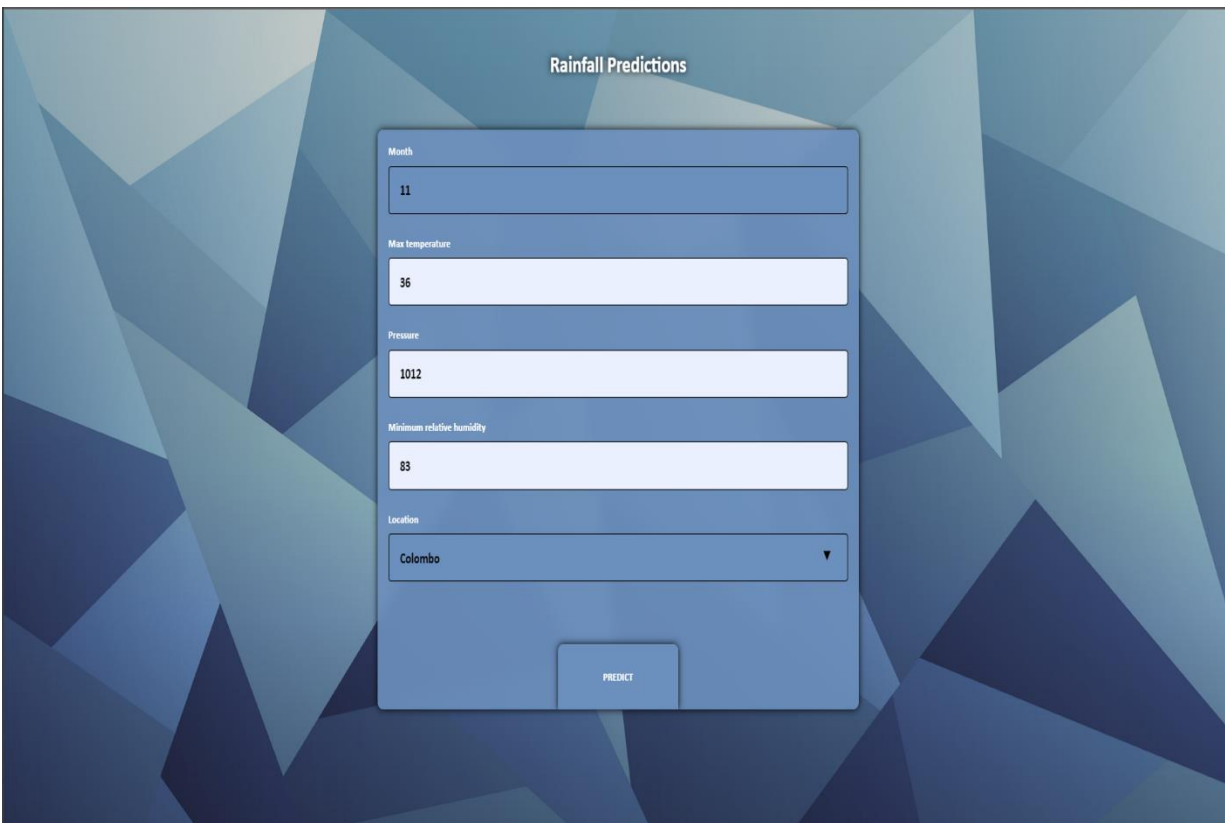
The final prediction system selects a prediction model based on location. Once the model has selected the input data to use to predict the binary classification "Rain or Not", This output adds to input data and new input data will be created. new data used as input for the multi-classification "Rain Range"

In this final rainfall prediction model, binary classification model output (rain or not) is used as input in the multi-classification model to increase the accuracy of multi-classification. The final Rainfall prediction system makes a prediction based on IoT data and a prediction model for specific IoT device locations.

### 2.3.2.6 Designs

In the proposed rainfall prediction system, machine learning methods are used to create prediction models. This is the simplest method for entering data from the input form. By developing location-based rainfall prediction models, one of the key goals of this study topic is to increase the accuracy of predicting rainfall and rain range.

- Input weather data from (figure 11).
- Prediction page of Rainy day (figure 12)
- Prediction page of Sunny day (figure 13)

The image shows a web application interface titled "Rainfall Predictions". It features a central form with a blue header and a light blue background. The form contains five input fields: "Month" with the value "11", "Max temperature" with the value "36", "Pressure" with the value "1012", "Minimum relative humidity" with the value "83", and "Location" with a dropdown menu showing "Colombo". A "PREDICT" button is located at the bottom right of the form. The background of the entire page is a dark blue geometric pattern.

*Figure 12:Weather data input form*



Figure 14: Prediction Page(Rainy Day)

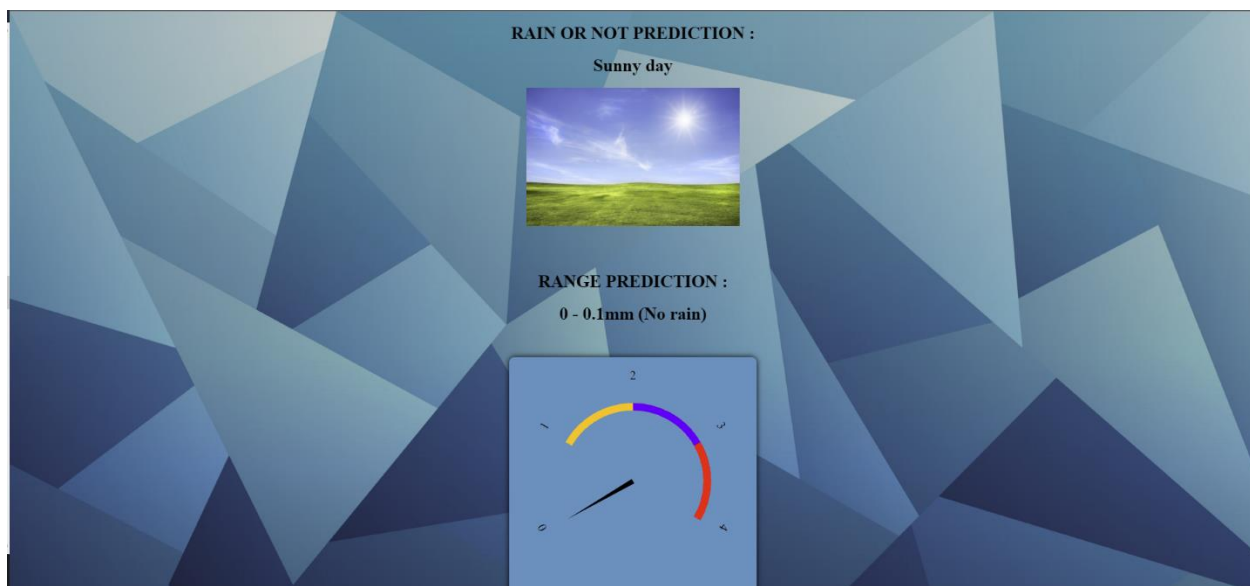


Figure 13: Prediction Page(Sunny Day)

## 3 RESULTS & DISCUSSION

### 3.1 Results

Figure 14 shows the result of Support Vector Machine Colombo binary classification model accuracy.

```
✓ [22] svm_binary_test_MAE = mean_absolute_error(y_test,y_pred)
0s svm_binary_test_MSE = mean_squared_error(y_test,y_pred)
svm_binary_test_RMSE = np.sqrt(mean_squared_error(y_test,y_pred))
svm_binary_test_R2 = r2_score(y_test,y_pred)

✓ print('MAE',svm_binary_test_MAE)
0s print('MSE',svm_binary_test_MSE)
print('RMSE',svm_binary_test_RMSE)
print('R2',svm_binary_test_R2)

MAE 0.28888888888888886
MSE 0.28888888888888886
RMSE 0.5374838498865699
R2 -0.15612648221343872

✓ acc = predict_range_new1.score(x_test,y_test)
0s print("Colombo",round(acc*100, 2), '%')

Colombo 71.11 %
```

Figure 15:SVM (Binary-Colombo)

Figure 15 show the result of Logistic Regression Katugastota binary classification model accuracy

```
✓ [16] LR_binary_test_MAE = mean_absolute_error(y_test,y_pred)
0s LR_binary_test_MSE = mean_squared_error(y_test,y_pred)
LR_binary_test_RMSE = np.sqrt(mean_squared_error(y_test,y_pred))
LR_binary_test_R2 = r2_score(y_test,y_pred)

✓ print('MAE',LR_binary_test_MAE)
0s print('MSE',LR_binary_test_MSE)
print('RMSE',LR_binary_test_RMSE)
print('R2',LR_binary_test_R2)

MAE 0.2111111111111111
MSE 0.2111111111111111
RMSE 0.45946829173634074
R2 0.15343915343915338

✓ [18] acc = Katugastota_predict.score(x_test,y_test)
0s print("Katugastota",round(acc*100, 2), '%')

Katugastota 78.89 %
```

Figure 16:LR Katugastota model

### 3.1.1 Models Accuracy Comparison with Data Set Changes - 1

Figure 19 shows the training and testing accuracy comparison between the Logistic Regression binary classification models for each dataset. The first graph shows the comparison of the accuracy of the combined data(Temperature maximum, relative humidity minimum, month) set without a specific location. The second graph shows the comparison of the accuracy of the dataset (Temperature maximum, relative humidity minimum, month, and station). The third graph shows the comparison of the accuracy of the dataset (Temperature maximum, Relative humidity minimum, month) with the specific location of Colombo. The fourth graph shows the comparison of the accuracy of the dataset (Temperature maximum, Relative humidity minimum, month) with the specific location of Katugastota. The fifth graph shows the comparison of the accuracy of the dataset (Temperature maximum, Relative humidity minimum, month) with the specific location of Vavuniya.

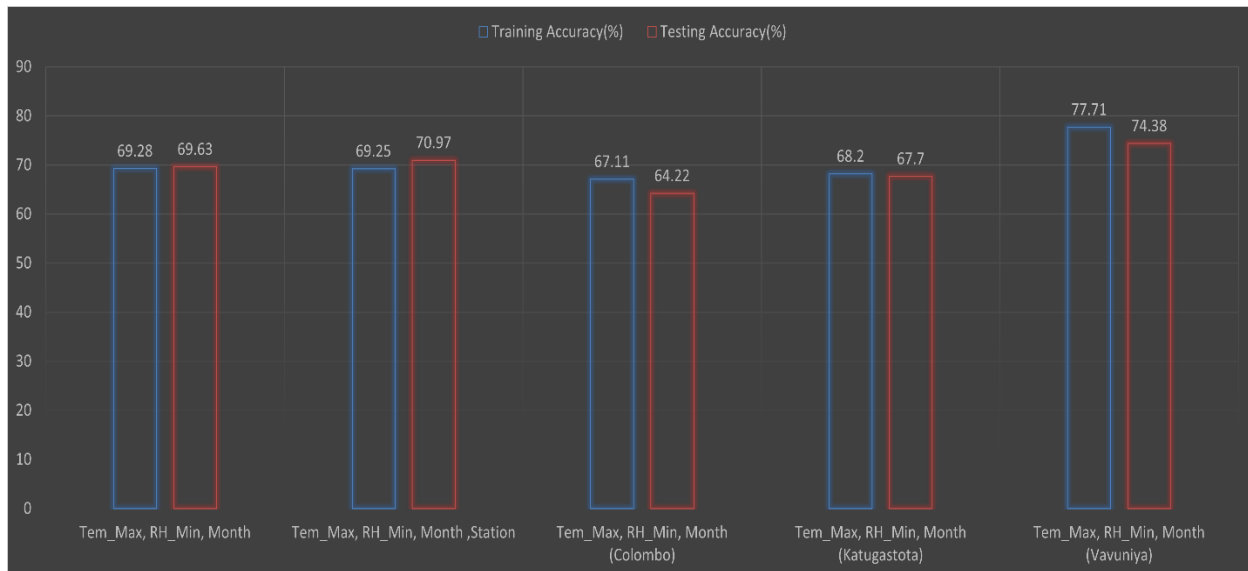


Figure 17:LR Model Comparison with Data Set Changes - 1

### 3.1.2 Model Accuracy Comparison with Data Set Changes - 2

Figure 18 shows the training and testing accuracy comparison between the Logistic Regression binary classification models for each dataset. The three graphs show the comparison of the accuracy of the combined data (Temperature maximum, relative humidity minimum, month, and sea level pressure) set with specific locations in Vavuniya, Katugastota, and Colombo. The fourth and fifth graphs show the comparison of the accuracy of the dataset (Temperature maximum, relative humidity minimum, month, sea level pressure) and the dataset (Temperature maximum, relative humidity minimum, month, sea level pressure) for a specific location in Vavuniya.

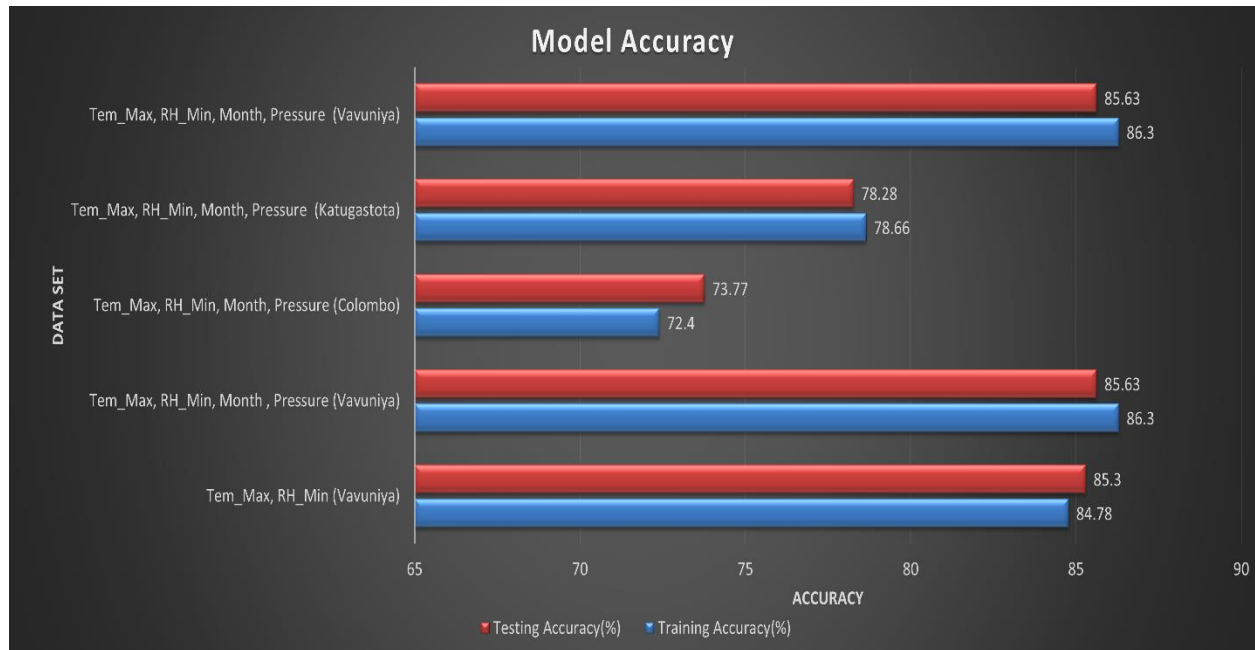


Figure 18:LR Model Comparison with Data Set Changes - 2

### 3.1.3 All Models Accuracy Comparison

Figure 17 shows an accuracy comparison between the Logistic Regression (LR) binary classification model, Support Vector Machine (SVM) binary classification model, Support Vector Machine (SVM) multi-classification model, Auto-ML best model (Random Forest) binary classification, and Auto-ML best model (Random Forest) multi-classification for the dataset (Temperature maximum, relative humidity minimum, month, and sea level pressure) for specific locations in Colombo, Katugastota, and Vavuniya.

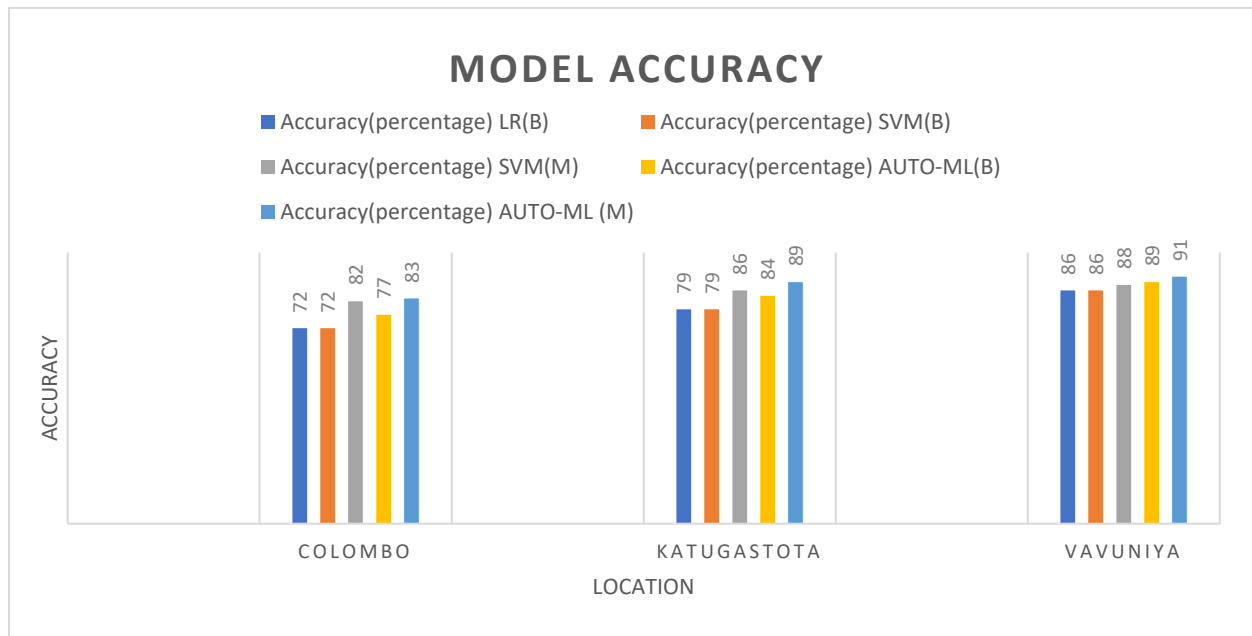


Figure 19: All Models Accuracy Comparison

### 3.2 Research Findings

The goal of this research is to better understand the attribution contributions of rainfall based on three different locations in Sri Lanka and to improve the process' overall performance. The data was gathered in three different locations: Colombo (coast and city), Vavuniya (countryside), and Katugastota (hills).

The contribution of temperature and relative humidity is higher in causing rainfall, but the contribution of the attribute "month" can be neglected, and the attribute pressure gives less contribution to rainfall when compared to temperature and relative humidity. Maximum temperature and minimum relative humidity with a time frame of the day (24 hours) and location as a constant are significant values. While using the same attribute data set for each specific location selected from Sri Lanka, such as Colombo (coastal and city), Vavuniya (countryside), and Katugastota (Hills), the prediction error or accuracy of each location model is different. Coastal and city location prediction errors seem higher than in hills or rural areas.

SVM and Logistic Regression are roughly similar in terms of accuracy. While using the Binary model output "Rain or Not" attribute as an extra input to the multi-classification model, its accuracy has increased. Auto sci-kit learns to select random forests as the best accuracy model. Random forest gives better accuracy than support vector machine or logistic regression. Because random forest uses many decision trees and prioritizes the model with the highest accuracy, it provides better accuracy.



### 3.3 Discussion

The contribution of temperature and relative humidity is higher in causing rainfall as expected, but the contribution of the attribute "month" can be neglected while using specific locations, and the attribute pressure gives less contribution to rainfall. While using the same attribute data set for each specific location selected from Sri Lanka, such as Colombo (coastal and city), Vavuniya (countryside), and Katugastota (Hills), the Vavuniya and Katugastota models give better accuracy than the Colombo model.

SVM and Logistic Regression are roughly similar in terms of accuracy. While using the Binary model output "Rain or Not" attribute as an extra input to the multi-classification model, its accuracy has increased. Auto sci-kit learn was used to compare the best classifier accuracy and our model accuracy for binary and multi-classification. Auto sci-kit learns to select random forests as the best accuracy model. Random forest gives better accuracy than support vector machine or logistic regression. Because random forest uses many decision trees and prioritizes the model with the highest accuracy, it provides better accuracy. While the split data set for each decision tree is important, the misclassified data from the earlier decision trees get more priority.

## 4 BUDGET & JUSTIFICATION

Component	Amount (LKR)
Cost for daily weather data for the month per parameter	40.00 Rs per month
Parameters(Relative humidity Minimum, Temperature Maximum, Sea level pressure, Rainfall)	x 4 parameters
5 years data	x12 months x 5 years
3 locations	x 3 locations
Total	= 28800.00 Rs

*Table 5: Budget*

## 5 SUMMARY OF STUDENT CONTRIBUTION

Member	Component	Task
Vinobaji. S	Development of a machine learning algorithm for rainfall prediction based on locations based which one uses in the Decision-making model as input.	<p>A feasibility study is conducted to determine the aspect's requirements.</p> <p>Identify the best dataset based on attributes contribution inaccuracy which will be used in the rainfall prediction model.</p> <p>Create a framework for obtaining weather data as an input.</p> <p>Test the generated machine learning model on a dataset with known results to verify accuracy.</p> <p>To increase accuracy, use binary classification model output as input for Multi classification model.</p>

*Table 6: Student Contribution*

## 6 CONCLUSIONS

The contribution of temperature and relative humidity is higher in causing rainfall as expected because of the maximum temperature and minimum relative humidity with time frame day (24hrs) and location as constant being significant values. but the contribution of the attribute "month" shows less contribution can be neglected while using specific locations, and the attribute pressure gives less contribution to rainfall. While using the same attribute data set for each specific location selected from Sri Lanka, such as Colombo (coastal and city), Vavuniya (countryside), and Katugastota (Hills), the Vavuniya and Katugastota models give better accuracy than the Colombo model. Location Colombo can be considered as another factor that influences place. This may be due to coastal areas or artificial factors (Global Warming, deforestation).

I've found that SVM and Logistic Regression are roughly similar in terms of accuracy. Even though the auto-machine learning model auto sci-kit learn shows the best rainfall forecast as Random Forest. The data used as input for prediction and classification has a significant impact on the percentage of accuracy and prediction. All models have advantages and disadvantages, and the most difficult part is deciding which model is the best. Logistic Regression for binary classification and SVM for multi-classification models indicate a proficient and appropriate model for this rainfall prediction.

Auto sci-kit learn was used to compare the best classifier accuracy and our model accuracy for binary and multi-classification. The best accuracy model, according to Auto Sci-kit Learn, is random forest. Because random forest uses numerous decision trees and prioritizes the best accuracy models, it provides improved accuracy. While dividing the data set for each decision tree, the misclassified data from the previous decision tree is given higher priority.

Future work can be considered as comparing another rural coastal location's accuracy with a city coastal location, considering artificial factors that influence rainfall (air pollution level of a location, etc.), planning to use a deep learning model to predict rainfall, and finding another time factor which can be used to replace the month.

## REFERENCE LIST

- [1] “Rainfall Prediction for Udaipur, Rajasthan using Machine Learning Models Based on Temperature, Vapour Pressure, and Relative Humidity,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 6S, pp. 133–137, 2020, doi: 10.35940/ijrte.f1024.0386s20.
- [2] “Weather Forecasts Based on Rainfall Prediction Using Machine Learning Methodologies,” *Adalya J.*, vol. 9, no. 6, 2020, doi: 10.37896/aj9.6/009.
- [3] K. Dutta and P. Gouthaman, “Rainfall Prediction using Machine Learning and Neural Network,” *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 1954–1961, 2020, doi: 10.35940/ijrte.a2747.059120.
- [4] S. B and J. K.S, “Rainfall Prediction Using Machine Learning Techniques and an Analysis of the Outcomes of These Techniques,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 04, no. 09, pp. 365–371, 2020, doi: 10.33564/ijeast.2020.v04i09.047.
- [5] S. Zainudin, D. S. Jasim, and A. A. Bakar, “Comparative analysis of data mining techniques for Malaysian rainfall prediction,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1148–1153, 2016, doi: 10.18517/ijaseit.6.6.1487.
- [6] J. Refonaa, M. Lakshmi, A. C. S. Kiran, and A. R. Teja, “Rainfall prediction using Apriori Algorithm,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 3, pp. 593–596, 2019, doi: 10.35940/ijrte.B1109.0782S319.
- [7] J. Refonaa, M. Lakshmi, R. Abbas, and M. Raziullha, “Rainfall prediction using regression model,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special issue 3, pp. 543–546, 2019, doi: 10.35940/ijrte.B1098.0782S319.
- [8] S. Poornima and M. Pushpalatha, “Prediction of rainfall using intensified LSTM based recurrent Neural Network with Weighted Linear Units,” *Atmosphere (Basel)*, vol. 10, no. 11, 2019, doi: 10.3390/atmos10110668.
- [9] M. Kühnlein, T. Appelhans, B. Thies, and T. Nauss, “Improving the accuracy of rainfall rates from optical satellite sensors with machine learning - A random forests-based approach applied to MSG SEVIRI,” *Remote Sens. Environ.*, vol. 141, pp. 129–143, 2014, doi: 10.1016/j.rse.2013.10.026.
- [10] Indrabayu, S. Aditama, A. A. Prayogi, S. Pallu, A. Achmad, and I. S. Areni, “Spatial-temporal approach for predicting rainfall in a tropical country,” *ICIC Express Lett.*, vol. 13, no. 2, pp. 113–118, 2019, doi: 10.24507/icicel.13.02.113.

- [11] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, “Kent Academic Repository Full-text document (pdf) Enquiries Citation for published version Link to recording in KAR An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,” *Expert Syst. Appl.*, vol. 85, pp. 1–27, 2017.
- [12] I. Cholissodin and S. Sutrisno, “Prediction of Rainfall using Simplified Deep Learning-based Extreme Learning Machines,” *J. Inf. Technol. Comput. Sci.*, vol. 3, no. 2, p. 120, 2018, doi: 10.25126/jitecs.20183258.
- [13] S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali, and Z. Nawaz, “Rainfall prediction in Lahore City using data mining techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 4, pp. 254–260, 2018, doi: 10.14569/IJACSA.2018.090439.

## APPENDICES

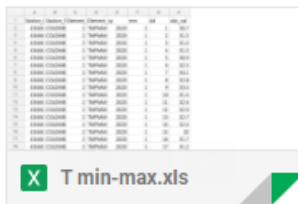
Data collection from the meteorology department of Sri Lanka.



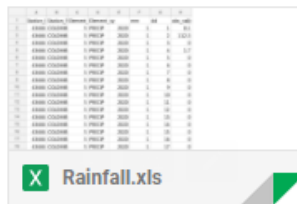
Computer Division\_DOM <computermeteo@gmail.com>

to me ▾

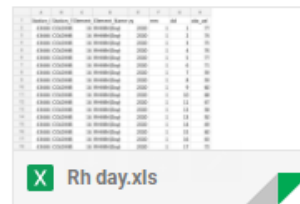
### 3 Attachments



X T min-max.xls



X Rainfall.xls



X Rh day.xls

Figure 20:Data Collection1



Computer Division\_DOM <computermeteo@gmail.com>

to me ▾

Wed, 31 Mar, 14:57

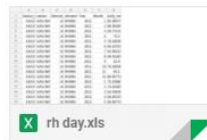
### 7 Attachments



X Monthly RH\_2010....



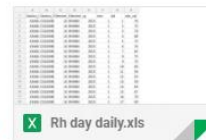
X Rf.xls



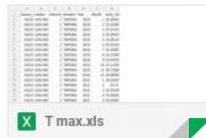
X rh day.xls



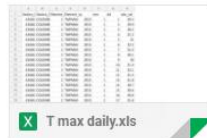
X rf daily.xls



X Rh day daily.xls



X T max.xls



X T max daily.xls

Figure 21:Data Collection 2



**climate division** <climatedivision123@gmail.com>  
to me ▾

Dear Sir/Madam,  
Sorry for the delay in releasing the data you requested.  
Please find the attachment.  
B/regards  
Climate Division

\*\*\*

Location	Month	Year	Daily Average Mean Sea Level Pressure
A	January	2010	1013.2
A	February	2010	1013.2
A	March	2010	1013.2
A	April	2010	1013.2
A	May	2010	1013.2
A	June	2010	1013.2
A	July	2010	1013.2
A	August	2010	1013.2
A	September	2010	1013.2
A	October	2010	1013.2
A	November	2010	1013.2
A	December	2010	1013.2
B	January	2011	1013.2
B	February	2011	1013.2
B	March	2011	1013.2
B	April	2011	1013.2
B	May	2011	1013.2
B	June	2011	1013.2
B	July	2011	1013.2
B	August	2011	1013.2
B	September	2011	1013.2
B	October	2011	1013.2
B	November	2011	1013.2
B	December	2011	1013.2
C	January	2012	1013.2
C	February	2012	1013.2
C	March	2012	1013.2
C	April	2012	1013.2
C	May	2012	1013.2
C	June	2012	1013.2
C	July	2012	1013.2
C	August	2012	1013.2
C	September	2012	1013.2
C	October	2012	1013.2
C	November	2012	1013.2
C	December	2012	1013.2
D	January	2013	1013.2
D	February	2013	1013.2
D	March	2013	1013.2
D	April	2013	1013.2
D	May	2013	1013.2
D	June	2013	1013.2
D	July	2013	1013.2
D	August	2013	1013.2
D	September	2013	1013.2
D	October	2013	1013.2
D	November	2013	1013.2
D	December	2013	1013.2
E	January	2014	1013.2
E	February	2014	1013.2
E	March	2014	1013.2
E	April	2014	1013.2
E	May	2014	1013.2
E	June	2014	1013.2
E	July	2014	1013.2
E	August	2014	1013.2
E	September	2014	1013.2
E	October	2014	1013.2
E	November	2014	1013.2
E	December	2014	1013.2
F	January	2015	1013.2
F	February	2015	1013.2
F	March	2015	1013.2
F	April	2015	1013.2
F	May	2015	1013.2
F	June	2015	1013.2
F	July	2015	1013.2
F	August	2015	1013.2
F	September	2015	1013.2
F	October	2015	1013.2
F	November	2015	1013.2
F	December	2015	1013.2
G	January	2016	1013.2
G	February	2016	1013.2
G	March	2016	1013.2
G	April	2016	1013.2
G	May	2016	1013.2
G	June	2016	1013.2
G	July	2016	1013.2
G	August	2016	1013.2
G	September	2016	1013.2
G	October	2016	1013.2
G	November	2016	1013.2
G	December	2016	1013.2
H	January	2017	1013.2
H	February	2017	1013.2
H	March	2017	1013.2
H	April	2017	1013.2
H	May	2017	1013.2
H	June	2017	1013.2
H	July	2017	1013.2
H	August	2017	1013.2
H	September	2017	1013.2
H	October	2017	1013.2
H	November	2017	1013.2
H	December	2017	1013.2
I	January	2018	1013.2
I	February	2018	1013.2
I	March	2018	1013.2
I	April	2018	1013.2
I	May	2018	1013.2
I	June	2018	1013.2
I	July	2018	1013.2
I	August	2018	1013.2
I	September	2018	1013.2
I	October	2018	1013.2
I	November	2018	1013.2
I	December	2018	1013.2
J	January	2019	1013.2
J	February	2019	1013.2
J	March	2019	1013.2
J	April	2019	1013.2
J	May	2019	1013.2
J	June	2019	1013.2
J	July	2019	1013.2
J	August	2019	1013.2
J	September	2019	1013.2
J	October	2019	1013.2
J	November	2019	1013.2
J	December	2019	1013.2
K	January	2020	1013.2
K	February	2020	1013.2
K	March	2020	1013.2
K	April	2020	1013.2
K	May	2020	1013.2
K	June	2020	1013.2
K	July	2020	1013.2
K	August	2020	1013.2
K	September	2020	1013.2
K	October	2020	1013.2
K	November	2020	1013.2
K	December	2020	1013.2
L	January	2021	1013.2
L	February	2021	1013.2
L	March	2021	1013.2
L	April	2021	1013.2
L	May	2021	1013.2
L	June	2021	1013.2
L	July	2021	1013.2
L	August	2021	1013.2
L	September	2021	1013.2
L	October	2021	1013.2
L	November	2021	1013.2
L	December	2021	1013.2
M	January	2022	1013.2
M	February	2022	1013.2
M	March	2022	1013.2
M	April	2022	1013.2
M	May	2022	1013.2
M	June	2022	1013.2
M	July	2022	1013.2
M	August	2022	1013.2
M	September	2022	1013.2
M	October	2022	1013.2
M	November	2022	1013.2
M	December	2022	1013.2
N	January	2023	1013.2
N	February	2023	1013.2
N	March	2023	1013.2
N	April	2023	1013.2
N	May	2023	1013.2
N	June	2023	1013.2
N	July	2023	1013.2
N	August	2023	1013.2
N	September	2023	1013.2
N	October	2023	1013.2
N	November	2023	1013.2
N	December	2023	1013.2
O	January	2024	1013.2
O	February	2024	1013.2
O	March	2024	1013.2
O	April	2024	1013.2
O	May	2024	1013.2
O	June	2024	1013.2
O	July	2024	1013.2
O	August	2024	1013.2
O	September	2024	1013.2
O	October	2024	1013.2
O	November	2024	1013.2
O	December	2024	1013.2
P	January	2025	1013.2
P	February	2025	1013.2
P	March	2025	1013.2
P	April	2025	1013.2
P	May	2025	1013.2
P	June	2025	1013.2
P	July	2025	1013.2
P	August	2025	1013.2
P	September	2025	1013.2
P	October	2025	1013.2
P	November	2025	1013.2
P	December	2025	1013.2
Q	January	2026	1013.2
Q	February	2026	1013.2
Q	March	2026	1013.2
Q	April	2026	1013.2
Q	May	2026	1013.2
Q	June	2026	1013.2
Q	July	2026	1013.2
Q	August	2026	1013.2
Q	September	2026	1013.2
Q	October	2026	1013.2
Q	November	2026	1013.2
Q	December	2026	1013.2
R	January	2027	1013.2
R	February	2027	1013.2
R	March	2027	1013.2
R	April	2027	1013.2
R	May	2027	1013.2
R	June	2027	1013.2
R	July	2027	1013.2
R	August	2027	1013.2
R	September	2027	1013.2
R	October	2027	1013.2
R	November	2027	1013.2
R	December	2027	1013.2
S	January	2028	1013.2
S	February	2028	1013.2
S	March	2028	1013.2
S	April	2028	1013.2
S	May	2028	1013.2
S	June	2028	1013.2
S	July	2028	1013.2
S	August	2028	1013.2
S	September	2028	1013.2
S	October	2028	1013.2
S	November	2028	1013.2
S	December	2028	1013.2
T	January	2029	1013.2
T	February	2029	1013.2
T	March	2029	1013.2
T	April	2029	1013.2
T	May	2029	1013.2
T	June	2029	1013.2
T	July	2029	1013.2
T	August	2029	1013.2
T	September	2029	1013.2
T	October	2029	1013.2
T	November	2029	1013.2
T	December	2029	1013.2
U	January	2030	1013.2
U	February	2030	1013.2
U	March	2030	1013.2
U	April	2030	1013.2
U	May	2030	1013.2
U	June	2030	1013.2
U	July	2030	1013.2
U	August	2030	1013.2
U	September	2030	1013.2
U	October	2030	1013.2
U	November	2030	1013.2
U	December	2030	1013.2
V	January	2031	1013.2
V	February	2031	1013.2
V	March	2031	1013.2
V	April	2031	1013.2
V	May	2031	1013.2
V	June	2031	1013.2
V	July	2031	1013.2
V	August	2031	1013.2
V	September	2031	1013.2
V	October	2031	1013.2
V	November	2031	1013.2
V	December	2031	1013.2
W	January	2032	1013.2
W	February	2032	1013.2
W	March	2032	1013.2
W	April	2032	1013.2
W	May	2032	1013.2
W	June	2032	1013.2
W	July	2032	1013.2
W	August	2032	1013.2
W	September	2032	1013.2
W	October	2032	1013.2
W	November	2032	1013.2
W	December	2032	1013.2
X	January	2033	1013.2
X	February	2033	1013.2
X	March	2033	1013.2
X	April	2033	1013.2
X	May	2033	1013.2
X	June	2033	1013.2
X	July	2033	1013.2
X	August	2033	1013.2
X	September	2033	1013.2
X	October	2033	1013.2
X	November	2033	1013.2
X	December	2033	1013.2
Y	January	2034	1013.2
Y	February	2034	1013.2
Y	March	2034	1013.2
Y	April	2034	1013.2
Y	May	2034	1013.2
Y	June	2034	1013.2
Y	July	2034	1013.2
Y	August	2034	1013.2
Y	September	2034	1013.2
Y	October	2034	1013.2
Y	November	2034	1013.2
Y	December	2034	1013.2
Z	January	2035	1013.2
Z	February	2035	1013.2
Z	March	2035	1013.2
Z	April	2035	1013.2
Z	May	2035	1013.2
Z	June	2035	1013.2
Z	July	2035	1013.2
Z	August	2035	1013.2
Z	September	2035	1013.2
Z	October	2035	1013.2
Z	November	2035	1013.2
Z	December	2035	1013.2

Figure 22:Data Collection 3