Paper No: SE-08 **Systems Engineering**

A novel approach for weather prediction for agriculture in Sri Lanka using Machine Learning techniques

J. S. A. N. W. Premachandra*
Department of Computer Science
Gen. Sir John Kotelawala Defence University, Sri Lanka
nishadiwasana833@gmail.com

P. P. N. V. Kumara
Department of Computer Science
Gen. Sir John Kotelawala Defence University, Sri Lanka
nandana@kdu.ac.lk

***Abstract* - Climate variability in recent years has critically affected the usual aspects of human lives, where the agriculture sector can be considered as one of the most vulnerable. Sri Lanka is also facing these climate changes over the past few decades. It has resulted in rainfall pattern changes where the expected rain may not occur during the expected time and amount. The mismatch between the rainfall pattern and traditional seasonal cultivation schedule has critically affected the agricultural sustainability. Even with the current technological advancements, weather prediction is one of the most technically and scientifically challenging tasks. This paper presents a novel machine learning-based approach for predicting rainfall for precision agriculture in Sri Lanka and it can be recognized as the first attempt to validate machine learning models to predict the weather in Sri Lankan context for precision agriculture. By analyzing the nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. For the prediction model, cross-validated data are trained and tested with four machine learning algorithms: Multiple Linear Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest. Currently, Support Vector Machine, K-Nearest Neighbors models have achieved accuracies of 88.57%, 88.66%. Random Forest has been recognized as the best-fitted model with 89.16% accuracy. The results depict a significant accuracy in this novel approach for Sri Lankan weather prediction.**

Keywords - data mining, machine learning, precision agriculture, weather prediction

I. INTRODUCTION

As a developing country in the Asian region, Sri Lanka has an economy based on agriculture while emphasizing that the agricultural sector is playing a significant role in the country's current development in both economic and social aspects[1]. Climate variability in recent years has critically affected the usual aspects of human lives, where the agricultural sector can be considered as one of the most vulnerable. According to the report "Sustainable Sri Lanka 2030 Vision and Strategic Path", as a developing country, Sri Lanka is facing potential agricultural risks due to unpredictable climatic changes[2]. The discrepancy between the rainfall pattern and traditional seasonal cultivation due to climatic variabilities is the main problem which is addressed in this research. According to the Intergovernmental Panel of Climate Change, among the sub-regions of Asia, South Asia is facing the most vulnerable climate changes. Sri Lanka has also been facing these changes during the past few decades, which has been resulted in rainfall pattern changes where the expected rain may not occur during the expected time as well as with the expected amount and intensity. As a result, a mismatch between the rainfall pattern and traditional seasonal cultivation schedule will happen. This problem indicates the current necessity of an advanced weather prediction model that can be used to guide farmers on their cultivation schedules based on weather and make them ready to handle the issues that occurred with the uncertain climate changes.

The climate of Sri Lanka consists of a variety of different conditions which depend on the geographical existence of different locations on the island. Generally, Sri Lanka has been divided into three main climatic zones: wet, dry, and intermediate. This research aims to propose a weather prediction model to predict daily rainfall in Kandy district, Sri Lanka, which belongs to both wet and intermediate climate zones.

As a result of the modern technological advancements in data analysis, variations in weather-related atmospheric conditions such as precipitation/rainfall, humidity, wind speed, wind direction, temperature, etc. are now accessible for any person. Weather can be demonstrated as an atmospheric state based on the above-mentioned parameters at a particular time and location. As Wiston [3] have mentioned in their research article, the scientific estimation of weather conditions for a specific future time can be performed with the following three steps,

1. Observing and collecting the required data related to weather
2. Processing and analyzing collected data
3. Extrapolating for future state prediction of the atmosphere

Combining the above observations analyzed data with designed models integrated with computer systems will produce a prediction model. All these three steps are significant for improving the accuracy of weather prediction. Most of the existing approaches are based on the weather data related to a particular geographical region on which the research has focused. Therefore, when developing a weather prediction model for Sri Lanka, it is vital to identify the most appropriate weather conditions with higher reliability. Also, processing and analyzing collected data is highly affected for obtaining the most accurate results. Required data preprocessing techniques are different based on the nature of the collected data. Therefore, to obtain high-quality predicted weather results through this research, data preprocessing techniques are identified and applied while ensuring that the originality of

raw data is not changed. When selecting the machine learning algorithms, the nature of the input data and the expected output has to be considered[4]. For the weather prediction model developed through this research, the machine learning algorithms have been selected by considering the size, nature of the distribution of the input dataset, speed, and accuracy of the output.

In this study, a historical weather data set has been received from the Central Environment Authority, including hourly data of different weather conditions such as rain gauge, average temperature, etc. By analyzing the nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. Daily data has been generated from the collected hourly data by averaging. A sequence of data preprocessing techniques has been used to assure the quality of the predicted output. A Cross-Validation has been done for the preprocessed data by partition the data set as 70% for model training and 30% for testing purposes [5]. Four machine learning algorithms are used for the weather prediction model: Multiple Linear Regression, Support Vector Machine, K Nearest Neighbors, and Random Forest. Based on the performance and accuracy, the best-fitted model for weather prediction is recognized.

The organization of the paper is as follows. Literature Review. Methodology and Results have been demonstrated in sections II, III, and IV, respectively. Finally, section V discusses the conclusion of this study.

II. Literature Review

A comparative study conducted by Medar [6] have stated different weather-predicting techniques as below,

- Synoptic Weather Prediction- weather parameters are observed within a specific time
- Numerical Weather Prediction includes advanced computer programs based on physical and mathematical equations or algorithms related to weather. Variations occur within the weather over time for deriving meteorological predictions
- Statistical Weather Prediction- identified as a part of objective weather prediction and it is generally focused on least square regression procedures

There are numerous existing weather prediction approaches proposed by researchers through their studies about statistical models and data analytic techniques for predicting future weather in terms of different weather-related variables. Some of these attempts are on identifying the most accurate and efficient techniques for data analytics to predict weather are based on statistical models, while some are based on regressions, decision trees, clustering, neural networks, and many other data mining techniques[3].

Data preprocessing can be identified as an integral step in machine learning-based weather prediction. Research on rainfall prediction conducted by Mohapatra [7] recognizes the importance of data preprocessing because of the difficulty of dealing with the existing outliers and inconsistencies of raw data.

Rainfall Prediction based on data mining approaches can be identified as data models that are more data-intensive than compute-intensive. Bayesian prediction model supports in reducing compute overhead while efficiently working with large data sets. In addition, the Bayesian classifier demonstrates a supervised learning methodology and a statistical methodology for the classification process[8].

An application developed for atmospheric temperature prediction based on Support Vector Regression has been able to recognize the better performances of Support Vector Machines in weather Prediction. It is a compulsory practice to select the most suitable parameters for the application since parameter selection significantly affects the overall system performance[9].

Sequential Patterns-based classification for time series and numeric data from multiple sources has become a significant method in the field of data mining. Yasmin [10] has been able to recognize the importance of processing numeric data and classifying the identified sequential patterns in data to mine data with high accuracy. The system has the ability to maintain a good accuracy in terms of not eliminating the original meaning of raw data but the use of limited parameters to reduce the system complexity has a possibility to affect the accuracy of the system.

Air Pollution data has also been used in weather forecasting approaches. One such system has been proposed by Chakraborty [11] to forecast weather with an Incremental K-means clustering algorithm. However, though the accuracy is considerably high, the insights provided by the output results of this system are minimal.

A similar approach based on clustering analysis has been proposed by Kalyankar [12] for analyzing meteorological data. Clustering can be considered as one of the most useful data mining techniques that can be used to identify hidden patterns in large data sets.

Uncertainty can be a significant aspect of weather prediction because it is really difficult to forecast the future without having certainty in data. As Shahi [13] have indicated in their research, Fuzzy C-Mean clustering can improve the accuracy of weather predicting systems based on data mining techniques such as regression models and decision trees.

A rainfall prediction model developed by Joseph [14] based on Artificial Neural Networks is an empirical method-based prediction approach. In these types of approaches, since the amount of time required for model training excessively increased with the number of neurons, it is necessary to carefully determine the number of hidden layer neurons required for the model.

Shah [15] has provided a rainfall prediction model which enhances the accuracy by using a combination of machine learning and data mining techniques. According to their study, the best accuracy was given by Neural Networks and ARIMA models. In contrast, the Random Forest model has given the best accuracy in classification out of several machine learning algorithms used.

All these researches are conducted in different countries based on the relevant geographical context. However, none of them are based on Sri Lankan Agriculture domain and never validated regarding the Sri Lankan context for precision agriculture purposes.

TABLE I: SUMMARY OF LITERATURE REVIEW

No:	Application	Technologies Used	Attributes	Data Set	Remarks
01	Rainfall Prediction By: Mohopatra [7]	• Linear Regression • K-fold Cross Validation	• Precipitation • Wet day frequency	• Monthly • 100 years	Accuracy: 70% Pros: • Discrepancies in raw data have been removed successfully during data preprocessing. Cons: • Accuracy will be decreased due to the use of limited attributes.
02	Rainfall Prediction By: Nikam [8]	• Bayes Method	• Pressure • Relative • Humidity • Wind Speed • Rainfall	• Daily • 16000 instances	Accuracy: 81% - 96% Pros: • Simplicity • Efficient Performance Cons: • Accuracy depends on the size of the training data set • Missing values in an attribute category
03	Temperature Prediction By: Radhika [9]	• Support Vector Regression	• Maximum Temperature	• Daily • 5 years	Accuracy: Not mentioned Pros: • A better performance by SVM Cons: • System performance depends on the parameter selection
04	Extreme Weather Prediction By: Yasmin [10]	• Sequential Pattern Mining • Progressive Sequence Tree(PS Tree)	• Precipitation • Wind direction • Wind Speed	• 10 min. intervals	• Accuracy: Not Mentioned Pros: • Reduces the data complexity through data categorization. • Fast performance with high scalability. Cons: • Accuracy will be decreased due to the use of limited attributes
05	Weather Category Forecasting By: Chakraborty [11]	• Incremental K-means Clustering	Air Pollution elements • NOx • CO2 • SO2 • RPM	• Daily • 10 months	Accuracy: 83.3 % Pros: • Good Accuracy with a small data set. Cons: • Not compared with other existing incremental algorithms for clustering. • Predicted output is insufficient to make insights on the weather.
06	Analyzing Meteorological Data By: Kalyankar [12]	• K-means Clustering	• Rainfall • Pressure • Temperature	• Daily • 4 yrs	Accuracy: Not Mentioned Pros: • Can be used to build dynamic and adaptive prediction models. Cons: • Not compared with other existing incremental algorithms for clustering. • Predicted output is insufficient to make insights on the weather.
07	Temperature Prediction By: Shahi [13]	• Type-1 Fuzzy Logic System Fuzzy C Mean Clustering	• Temperature • Humidity	• 15 min. intervals • 4600 instances	Accuracy: 1.6590 RMSE Pros: • Higher accuracy by detecting outliers in data Cons: • Accuracy depends on the size of the data set
08	Rainfall Prediction By: Joseph [14]	• Artificial Neural Networks	• Humidity • Temperature • Pressure • Precipitable water • Wind speed	• Daily • 370 instances	• Accuracy: 87% Pros: • ANN can be used with both linear and non-linear data. Cons: • Model training time increase with the number of hidden layer neurons

09	Rainfall Prediction By: Shah [15]	• ARIMA model • Holt Winter method • Simple Moving Average model • Seasonal Naive method • Neural Networks	• Max. and Min. temperature • Relative Humidity • Wind Speed	• Daily (Jun. to Dec.) • 35 yrs	• Accuracy: 70.5% Pros: • Good accuracy through few parameters. Cons: • Dataset includes only half of every year (Jun to Dec). • Predict the rainfall only for months with a possibility to rain.

According to the literature review, several limitations exist in the currently available weather prediction approaches. Among them, the problems that occurred during the data collection process can be identified as major issues. In addition, inaccuracy in data where the collected data are not related to the problem domain, a high amount of missing data has affected the accuracy of the existing systems.

Inefficient data preprocessing has also affected accuracy reductions in current weather predicting systems. As a result of not carefully handling the incomplete and inconsistent data, most systems have been unable to obtain a high-quality output.

Weather prediction systems based on a single machine learning algorithm have faced the problem of selecting the best algorithm. However, the systems that have used multiple machine learning algorithms have not focused on selecting the most appropriate algorithms according to the research domain.

The main problem identified through the literature review is that even different features have been used by different researchers to ensure the accuracy and performance of their systems. Therefore, those proposed approaches have not consolidated those advanced features and techniques into a single system. For example, even a system has considered using a large data set for its model train, it does consider systematic data-preprocessing techniques. As a result, even the data set is adequately large, due to insufficient data preprocessing techniques, the expected accuracy and performance of the system will not be achievable. Also, the systems that give a considerable accurate level cannot provide valuable insights through the predicted results. Therefore it is important to carefully recognize the nature of the intended output given through the model while thinking about whether that output can fulfill the purpose of requirements.

III. METHODOLOGY

The proposed architecture of the weather prediction approach comprises a set of interrelated steps such as data collection, data preprocessing, exploratory data analysis, application of machine learning algorithms, evaluation and identification of the best ML algorithm, and analysis of results. This research mainly focuses on identifying the most appropriate technology-based solution for weather prediction for precision agriculture in Sri Lanka. Even numerous advanced technologies are emerging continuously, it is important to select the most appropriate by identifying the nature of the context in which we are trying to apply them.

In Sri Lanka, rainfall is one of the most significant weather conditions required in agriculture-based decision-making. However, due to the high expensiveness of the available weather data in Sri Lanka, we have to perform the predictions based on a small dataset that does not comprise more than a few thousand records. By considering the nature of the requirement of predicted weather for Sri Lankan Agriculture basis and the available data on different weather parameters, we have proposed a machine learning-based weather prediction approach.

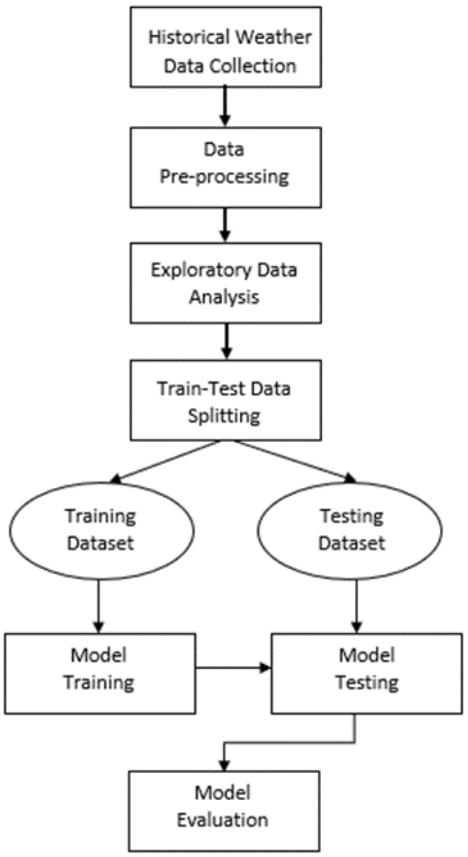

Fig 1. Proposed architecture

As we emphasized in the literature review, it is important to apply different techniques for each proposed architecture step to obtain a better accuracy level. Therefore, the aim of this research is to follow the identified effective practices in the reviewed literature while overcoming the issues that exist within the current approaches in order to build up a better solution for weather prediction using machine learning.

A. Data Collection and pre-processing

In the first part of the proposed weather prediction model, a data set of historical weather data in Sri Lanka is retrieved from the meteorological department and the Central Environmental Authority. Hourly data from 01.012019 to 28.02.2021 is collected. . By analyzing the

nature of the weather in Sri Lanka, the relationship of weather attributes with agriculture, availability, and accessibility, seven attributes are selected, including rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context. Daily data is generated from the collected hourly data by averaging.

In order to assure the quality of the predicted results, the collected and structured data are pre-processed through a sequence of data preprocessing techniques as follows,

- **Data Consolidation –** Required data were collected from different sources and therefore, it is required to integrate them into a single table.
- **Data Reduction-** To maintain the prediction model's efficiency, redundant and unnecessary data were removed from the data set.
- **Data Cleansing-** Since the dataset consists of null values and noises, it is very important to handle them carefully. As concluded in the literature review, they are filled with average values instead of replacing missing values with zero.
- **Data Discretization**- To utilize data within machine learning algorithms, rainfall data values are segregated into two intervals: Rain (1) where Rain Gauge is greater than 0mm, and No Rain(0) where Rain Gauge is 0mm.

In order to preprocess data efficiently and accurately, we use python with its libraries including NumPy, Matplotlib, Pandas.

B. Exploratory data analysis

In order to identify the nature of weather condition distributions and correlations, distribution graphs and correlation matrices are used[16]. Correlation matrices can be used to recognize the weather conditions that are most affected by rainfall. In addition to data summarization, data visualization is useful in discovering insights in data effectively and efficiently. In this study, R ggplot2 is used for the exploratory data analysis because it provides better visualization features through its default plots with magnificent graphics.

C. Train-test data splitting

Weather data are usually time series but to prevent unnecessary bias to the machine learning model, we used the Train_Test_Split module. The Train_Test_Split approach, a common cross-validation technique, is done for the pre-processed data by partitioning the data set as 70% for model training and 30% for the testing purpose.

D. Training and testing model

After analyzing the nature of the input dataset and the expected requirements of the output results four supervised machine learning algorithms are used. The purpose of using multiple algorithms instead of a single algorithm is to predict rainfall at a highly accurate level through an evaluation comparison of the results. Multiple Linear Regression has been used as a regression model while Support Vector Machine, K-Nearest Neighbors and Random Forest Models have been used as classification models[17].

a) Multiple Linear Regression (MLR)

Multiple Linear Regression is a machine learning regression approach, which attempts for the relationship modeling between two or more independent variables and response through fitting a linear equation for the observed data. Homogeneity invariance, independence of observations, multi-variate normality, and linearity are the assumptions of the regression model[18].

b) Support Vector Machine (SVM)

In this algorithm, it tries to identify a hyperplane within an x-dimensional space that has the ability to classify the data points in a distinct manner where ***x*** means the number of features. Out of all the possibilities, the hyperplane with the maximum margin is selected where the distance between the classes is maximum[19].

c) K-Nearest Neighbors (KNN)

This supervised machine learning algorithm also can be used for both classification and regression problems. K denotes the number of neighbors whose nearest to an unknown new variable is required to predict[20].

d) Random Forest (RF)

Random Forest is a famous and straightforward machine learning algorithm and it is based on ensemble learning that creates an effective model by combining multiple classifiers. This algorithm provides a combination of multiple decision trees and therefore, accuracy is high as well; it reduces overfitting up to a large content[21].

For each algorithm, default parameters are used without performing any modifications. After the model training process, it is used for predicting daily rainfall, based on the data available within the testing dataset. In this study, the weather prediction approach is based on supervised machine learning, including both regression and classification. For the implementation of the proposed solution, sci-kit-learn which is a Python-based module in machine learning also supported by pandas which is a Python library of statistical tools and data structures are used.

E. Model evaluation

For the evaluation of the above machine learning models, a confusion matrix and classification reports are used[22]. Since regression models give a continuous output, before computing the confusion matrix, the predicted output is classified into two categories as below;

- Rain Gauge > 0: Output= 1
- Rain Gauge < 0: Output=0

Accuracy, precision, and recall are the three metrics considered for the model evaluating process. Through the evaluation, the acceptable algorithms for weather prediction are recognized and then the most accurate approach is selected.

IV. RESULTS AND DISCUSSION

In this study, the gathered dataset includes 14000 records and 7 weather attributes were selected from the collected dataset. They are Rain Gauge, Relative Humidity (RH), Average Temperature (AT), Wind Speed (WS Raw),

Wind Direction (WD Raw), Solar Radiation(Solar Rad), Ozone Concentration (O3 Conc).

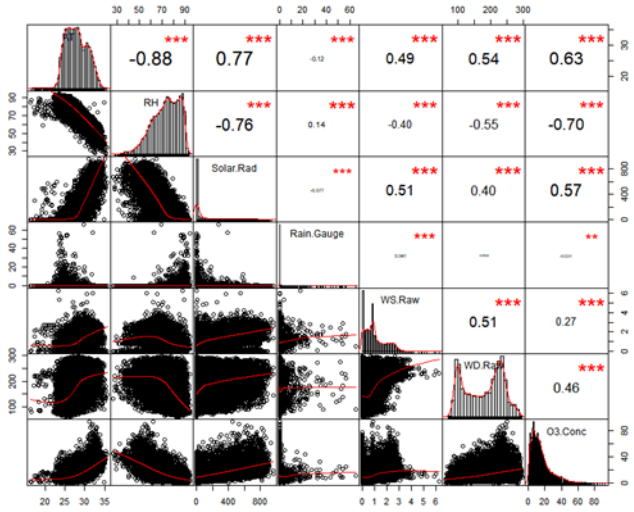

Fig 2. Correlation matrix chart

According to the correlation matrix chart represented in Fig. 2, computed through R, Rain Gauge and Solar Rad are not normally distributed.

	AT	RH	Solar.Rad	Rain.Gauge	WS.Raw	WD.Raw	O3.Conc
AT	1.00	-0.88	0.77	-0.12	0.49	0.54	0.63
RH	-0.88	1.00	-0.76	0.14	-0.40	-0.55	-0.70
Solar.Rad	0.77	-0.76	1.00	-0.08	0.51	0.40	0.57
Rain.Gauge	-0.12	0.14	-0.08	1.00	0.04	0.00	-0.02
WS.Raw	0.49	-0.40	0.51	0.04	1.00	0.51	0.27
WD.Raw	0.54	-0.55	0.40	0.00	0.51	1.00	0.46
O3.Conc	0.63	-0.70	0.57	-0.02	0.27	0.46	1.00

Fig. 3. Correlation matrix

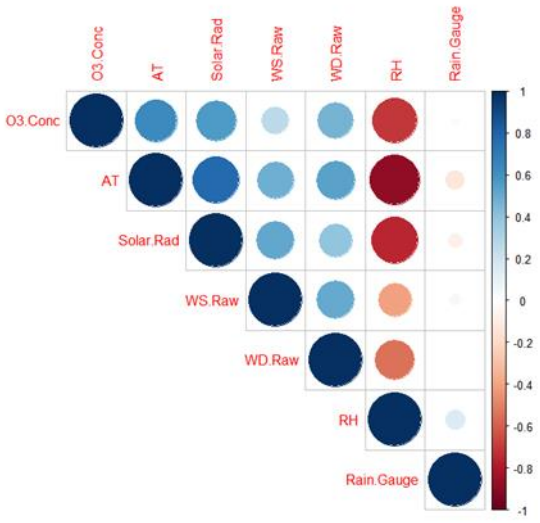

Fig. 4. Correlogram

As represented in Fig. 3 and Fig. 4, correlations among Rain Gauge and other weather parameters are slightly weak, it has computed correlations between Rain Gauge and multiple weather parameters as shown in Figure 5. The correlation between Rain Gauge and the combination of AT, RH, Solar Rad, WS Raw, WD Raw, O3 Conc. is 0.4949 which is a considerable value.

	Rain Gauge
RH + WSRaw	0.1494476
RH + WSRaw + O3	0.1538214
RH + WSRaw + O3 + AT	0.1794619
RH + WSRaw + O3 + AT + Solar Rad	0.1808611
RH + WSRaw + O3 + AT + Solar Rad + WDRaw	0.4949962

Fig. 5. Multiple Correlation

Also when the dataset is large, it is statistically significant even with a weak correlation[23].

A. Evaluation of MLR Model

According to the confusion matrix in Fig. 6 and the classification report in Fig. 7, the accuracy of the predicted output is 44% which is a considerable low accuracy. The accuracy of linear regression is often affected by the normal distribution nature of the data. Since the weather parameters are slightly weak, it is difficult to increase the accuracy of this model.

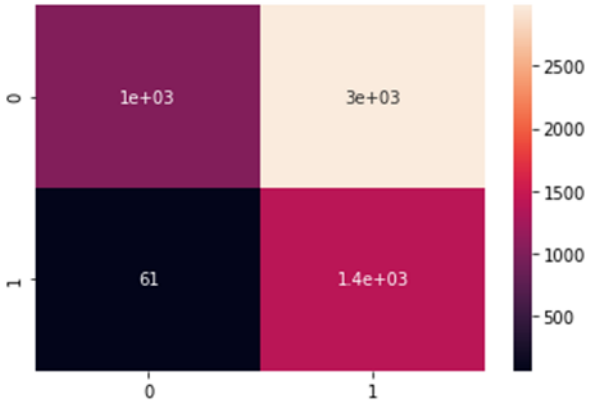

Fig. 6. Confusion matrix- MLR model

	precision	recall	f1-score	support
0.0	0.94	0.26	0.40	4018
1.0	0.32	0.96	0.48	1469
accuracy			0.44	5487
macro avg	0.63	0.61	0.44	5487
weighted avg	0.78	0.44	0.42	5487

Fig. 7. Classification Chart- MLR Model

As we concluded in the literature review, the accuracy of regression models depends on the number of variables used. The linear regression model proposed by Mohopatra [7] has acquired 70% accuracy with 2 attributes. In this research, we attempted to predict rainfall using 7 attributes but due to the weaknesses in the normal distribution of the data set which we used, we could reach 44% of accuracy.

B. Evaluation of SVM Model

According to the conclusions made through the literature review, most of the machine learning models including SVM required proper selection of weather parameters. Therefore, in this research, we highly focused on selecting the most suitable weather parameters by studying the domain and performing effective data analysis techniques [9-10].

As depicted in the classification report in Fig. 8, the SVM model has achieved 89% accuracy. This accuracy has been taken by rounding off the value 88.57%. This model has offered a high accuracy compared to the linear regression model. The main reason for achieving good accuracy is the ability of SVM to handle input spaces with non-linear features. Both precision and recall also have achieved greater than 80% where precision is 83% and recall is 89%.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.02	0.02	0.02	5487
weighted avg	0.83	0.89	0.85	5487

Fig. 8. Classification chart- SVM Model

C. Evaluation of KNN Model

KNN is a supervised machine learning model which can learn from already labeled data. As we previously mentioned, Rain Gauge values are appropriately labeled as either 1 or 0, and the dataset is properly preprocessed. Since our dataset is considerably small, we selected KNN which will be more applicable in these scenarios. For example, Chakraborty [11] has used a small data set with K-means clustering to forecast weather category and their model has secured 83% accuracy.

In this research, as depicted in the classification report in Fig. 9, the KNN model has achieved 89% accuracy when k=7. This accuracy has been taken by rounding off the value 88.66%. Thus, the precision is 83% and recall is 89%, similar to the KNN model.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.02	0.02	0.02	5487
weighted avg	0.83	0.89	0.86	5487

Fig. 9. Classification chart: KNN Model

D. Evaluation of RF Model

Random Forest was selected in this approach since it can be used in both regression and classification problems as well it has a simplified methodology to measure the relative importance of every feature on prediction. As depicted in the classification report in Fig. 10, the RF model has achieved 89% accuracy. This value has been taken by rounding off the value 89.16%. It has an 89% of recall which is similar to both SVM and KNN. However, the precision is 84% which is slightly higher than SVM and KNN models. Since the Random Forest model gives the best overall accuracy compared to the other models, Radom Forest can be recognized as the best-fitted model.

	precision	recall	f1-score	support
accuracy			0.89	5487
macro avg	0.04	0.03	0.03	5487
weighted avg	0.84	0.89	0.86	5487

Fig. 10. Classification chart: RF Model

E. Comparison of Machine Learning models

According to the evaluation, the summary showed in TABLE. II, the MLR model has achieved the lowest accuracy at 44%. However, it has been able to achieve a 78% precision. The highest accuracy, 89.16% has achieved by the RF model with the highest precision of 84%. Both SVM and KNN also have been able to achieve high accuracies as 88.57% and 88.66% as respectively.

TABLE II. EVALUATION RESULTS OF FOUR ML MODELS

Evaluation Criteria	Accuracy	Precision	Recall
MLR	44%	78%	44%
SVM	88.57%	83%	89%
KNN	88.66%	83%	89%
RF	89.16%	84%	89%

V. CONCLUSION AND FUTURE WORK

In this research, we comprehensively addressed that, weather plays a significant role in the field of agriculture. However, climate variability is always beyond human control. Sri Lanka is also struggling with the mismatch between weather pattern variations and traditional cultivational schedules. Accurate weather forecasts enable farmers to schedule their cultivation tasks while minimizing weather-based agricultural damages. The proposed architecture attempts to introduce a novel machine learning-based approach for predicting rainfall for precision agriculture in Sri Lanka. Since the weather conditions in Sri Lanka are not perfectly matched with other countries, it is very important to identify the most related weather conditions to predict the weather.

First, we concluded that seven weather attributes could be used to predict rainfall in Sri Lanka for precision agriculture. The selected attributes are rain gauge, relative humidity, average temperature, wind speed, wind direction where solar radiation and ozone concentration are uniquely selected for Sri Lankan context.

Secondly, through the exploratory data analysis, we concluded that the multiple correlation of the weather attributes is 0.4949 which is a good value compared to the correlations observed within existing.

Thirdly, we concluded that several data preprocessing techniques are required to enhance the quality of the prediction. Therefore, data consolidation, reduction, cleansing, and discretization were performed on the data carefully.

Fourthly, by studying and analyzing the problem background and the nature of obtained dataset to improve the accuracy, four supervised machine learning algorithms were selected. For the prediction, model cross-validated data were trained and tested with Multiple Linear Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest.

Finally, with the model evaluations, Random Forest was recognized as the best-fitted model that achieved 89.16% accuracy. This can be considered as a better level of accuracy compared to the prevailing weather prediction approaches.

As for future work is expected to increase the size of the dataset and apply more data preprocessing techniques such as feature engineering to enhance the quality of the

dataset. Since SVM and KNN models have also given better accuracy levels, it is important to build and evaluate a hybrid ensemble learning model which combines these machine learning models for this weather prediction approach. Deep learning is a member of the broader community of machine learning and it is based on artificial neural networks with representation learning. It is expected to apply deep learning for predicting the weather with a large dataset and evaluate the accuracy improvement.

REFERENCES

[1] T. B. Adhikarinayake, "Methodical design process to improve income of paddy farmers in Sri Lanka," [publisher not identified], Wageningen, 2005.

[2] "Sri Lanka tackles challenges to rice production to end reliance on imports," oxford business group.

[3] M. Wiston and M. Km, "Weather Forecasting: From the Early Weather Wizards to Modern-day Weather Predictions," J Climatol Weather Forecasting, vol. 06, no. 02, 2018, doi: 10.4172/2332-2594.1000229.

[4] L. H. S. De Silva, N. Pathirage, and T. M. K. K. Jinasena, "Diabetic Prediction System Using Data Mining," presented at the Proceedings in Computing, 9th International Research Conference-KDU, Sri Lanka, 2016.

[5] Berrar, "Cross-Validation," in Encyclopedia of Bioinformatics and Computational Biology, Elsevier, 2019, pp. 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X.

[6] R. Medar, A. B. Angadi, P. Y. Niranjan, and P. Tamase, "Comparative study of different weather forecasting models," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, Aug. 2017, pp. 1604–1609. doi: 10.1109/ICECDS.2017.8389719.

[7] Mohopatra Sandeep, "Rainfall Prection using 100 years of Meteorological Data." 2017.

[8] B. Nikam and B. B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach," in 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation, Seoul, Korea (South), Sep. 2013, pp. 132–136. doi: 10.1109/CIMSim.2013.29.

[9] R. Yalavarthi and M. Shashi, "Atmospheric Temperature Prediction using Support Vector Machines," 2009. doi: 10.7763/IJCTE, 2009.V1.9.

[10] R. Y. Yasmin, A. E. Sakya, and U. Merdijanto, "A classification of sequential patterns for numerical and time series multiple source data — A preliminary application on extreme weather prediction," in 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Nov. 2017, pp. 1–5. doi: 10.1109/ICODSE.2017.8285845.

[11] S. Chakraborty, N. K. Nagwani, and L. Dey, "Weather Forecasting using Incremental K-means Clustering," p. 6.

[12] Meghali A. Kalyankar, "Data Mining Technique to Analyse the Metrological Data," IJARCSSE.

[13] Shahi, R. B. Atan, and N. Sulaiman, "Detecting Effectiveness of Outliers and Noisy Data on Fuzzy System Using FCM," p. 13.

[14] J. Joseph and Ratheesh T K, "Rainfall Prediction using Data Mining Techniques," 2013.

[15] U. Shah, S. Garg, N. Sisodiya, N. Dube, and S. Sharma, "Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques," in 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, Dec. 2018, pp. 776–782. doi: 10.1109/PDGC.2018.8745763.

[16] T. Pham-Gia and V. Choulakian, "Distribution of the Sample Correlation Matrix and Applications," *OJS*, vol. 04, no. 05, pp. 330–344, 2014, doi: 10.4236/ojs.2014.45033.

[17] Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," IEEE Access, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[18] R. Bevans, "An introduction to multiple linear regression." https://www.scribbr.com/statistics/multiple-linear-regression/

[19] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," Medium, Jul. 05, 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed Jun. 02, 2021).

[20] "KNN - The Distance Based Machine Learning Algorithm," Analytics Vidhya, May 15, 2021. https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/ (accessed Jun. 02, 2021).

[21] S. Awasthi, "Random Forests in Machine Learning: A Detailed Explanation," datamahadev.com, Dec. 05, 2020. https://datamahadev.com/random-forests-in-machine-learning-a-detailed-explanation/ (accessed Jun. 02, 2021).

[22] M. Goonathilake and P. Kumara, "SherLock 1.0: An Extended Version of 'SherLock' Mobile Platform for Fake News Identification on Social Media," Sri Lanka, p. 7, 2020.

[23] "The Correlation Coefficient (r)." https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html (accessed May 28, 2021).