Predicting Meteorological Data using Machine Learning

Syed Hamid Ali Shah
Department of Electrical and Electronics Engineering
University of Bolton, Greater Manchester, UK

Rameez Asif, PhD
Department of Electrical and Electronics Engineering
University of Bolton, Greater Manchester, UK

ABSTRACT
Understanding the nature and behavior of weather has always been an essential task for humans, as it has a significant impact on the property and economy of a country. Machine Learning algorithms can predict the patterns of the weather nature i.e. floods, hurricanes, storms, cyclones, and rain. This paper looks at applying machine learning techniques to predict target variables i.e. precipitation to independent variables i.e. wind speed, temperature, pressure, and soil temperature for three years of weather data collected from the Manchester region that is publicly available on the OpenMeteo website. The objective of this research is to evaluate and compare the accuracy, precision, F1 score, and recall of linear regression, support vector machine regression, k-nearest neighbor regression, and random forest regression algorithms using Python,starting with preprocessing the data, developing the algorithms, training, and finally testing it. The low mean squared error (MSE), R^2 score, and mean absolute error (MAE) illustrate the ability of these algorithms for prediction. Further analysis of these algorithms shows linear and support vector machine regression with 92.2% and 92.5% accuracy.

Keywords
Machine Learning, Weather Prediction, Weather Data, Linear Regression, Support Vector Machine Regression, K- K-Nearest Neighbor Regression, Random Forrest Regression

1. INTRODUCTION
Weather patterns play an important role in a country's economy, and forecasting them in advance is always challenging for humans. Weather forecasting is done by collecting relevant characteristics of an atmosphere from meteorological observations[1]. It is important to predict the weather as agriculture, aviation, health, and tourism are greatly affected by it. Earlier, weather forecasting was done through physical observation. In the mid-20^{th} century, with the advancement of computer technologies, Traditional Numerical Weather Prediction (NWP) techniques were used. This method uses long differential and integral equations based on mass, momentum, and energy principles to develop a model and assess the atmospheric condition[2].Despite NWP methods having been used well in weather prediction for several years, the physical model has limitations in accurately predicting the output of the weather.Physical models are effective in predicting the weather for short distances i.e., ranging from 1km to 10km.Due to this limitation, and limits in human resources as well as the high cost of tools and sensors in collecting weather data, a need for more accurate techniques is needed to overcome this problem.

With the advancement and development of research in computer science, new trends in solving weather prediction with more accuracy and precision are introduced, called machine learning techniques. Machine learning algorithms can identify the relevant pattern from past weather data and learn it for future prediction without human intervention.Linear regression, Support Vector Machine (SVM) regression, KNN regression, and Random Forest (RF) regression are promising algorithms to develop a machine-learning model for accurate prediction [3]. These algorithms can discover the hidden patterns of relation between input and output to improve prediction accuracy.

Fluctuations in climate make it complex to accurately predict the weather. This dynamic pattern of weather complicatesadopting the accurate weather prediction model for meteorologists and scientists, which necessitates the exploration of supervised machine learning i.e. LR, SVM regression, KNN regression, and RF regression.

This study aims to develop models that can predict weather accurately using LR, SVM regression, KNN regression, and RF regression. This involves training models with the data and testing them. From evaluation metrics, such as accuracy, precision, F1 score, and recall, it becomes evident which model is performing well among all.

The remainder of this paper is organized as follows. Section 2 provides an overview of related research on weather prediction using supervised machine learning algorithms. Section 3 describes the methodology for the proposed work. Results and discussion are addressed in section 4. Finally, Section 5 presents a conclusion and future recommendations.

2. RELATED STUDY
Recently, researchers have shown great interest in predicting weather conditions using ML techniques for accuracy.

Certain approaches are available in literature for weather prediction, including the physical methods, persistence method, statistical method, spatial correlation techniques, artificial intelligence, and hybrid methods and so on [4]- [11].

García-Vázquez et al.,in [12] suggested four supervised machine learning algorithms with a prime focus on linear regression and support vector machine regression for predicting the internal temperature of a greenhouse. For internal weather data (dewpoint, temperature, and humidity) and external data (temperature, humidity, and solar radiation) collection, a meteorological station is installed. A one-year dataset is collected, comprising a season division for better analysis. Sixteen experiments were involved in the model's performance evaluation using MAE, R^2, RMSE, and MAPE metrics. The results show that LR and SVR have the highest prediction accuracy among all algorithms.

Because of simplicity and efficiency, [13] evaluated the performance of the KNN algorithm for weather prediction. With different values of k, KNN was used for three cities in Nigeria to predict humidity, temperature, and rainfall. KNN's performance was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of

determination (R^2). A satisfactory performance has been shown in temperature prediction by KNN, but it may vary for specific weather variables being predicted and the data being used. This makes it useful in dynamic classification problems.

Suman in [14] demonstrated four models,i.e.,linear regression (LR), support vector regression (SVR), multivariate adaptive regression splines (MARS), and random forest (RF), for the prediction of daily and mean weekly rainfall at Ranichauri station, located in the district of Utarakhand. Meteorological variables, i.e., minimum and maximum temperature, the relative humidity for morning and evening, afternoon and morning vapor pressure, wind speed, wind direction in the morning and afternoon, evaporation, solar radiation, and rainfall collected from the Agromet Forest Unit (AFU) Ranchauri. Models were validated through statistical parameters. Overall performance for daily and mean weekly rainfall, RF was ranked first. Hence, the RF model is the best for daily and weekly rainfall prediction at Ranichauri station.

The reviewed work shows great potential in employing machine learning algorithms like linear regression, support vector machine regression, k-nearest neighbor regression, and random forest regression. Each of these algorithms highlights capturing the challenging nonlinear relationship, effectively classifying data, and predicting tasks with high accuracy. As continued climate change challenges, it is crucial to explore the capabilities of these diverse machine learning methods.

3. MATERIALS AND METHODS

As illustrated in **Figure 1** Raw historical data were initially collected from Open-Meteo over three years. This data is preprocessed and split into training and testing sets for machine learning model implementation. The working hypothesis for this research experiment is set as precipitation as a target variable,withthe independent variables,i.e.,wind speed, temperature_2m, pressure, and soil temperature. These models are evaluated with different performance matrices for result analysis and comparison.

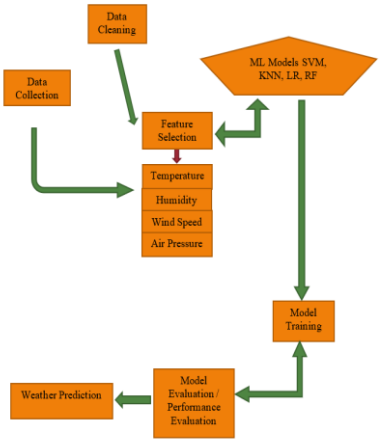

Figure 1: WorkflowMethodology

3.1 Historical Weather Data Collection

Historical weather data is collected from the Open-Meteo website [15] starting from 7/8/2021 to 7/8/2024 for Manchester, UK (Coordinates: 53.47, -2.24). This dataset includes weather parameters i.e. temperature, relative humidity, precipitation, rain, snowfall, pressure, wind speed at 10 and 100m, and soil temperature. The**Table *1***includes 26,328 rows and 11 columns. 23 values are missing in this dataset.

Table 1: Weather Data Parameters

Column Name	Non-Null Count	Data Type	Missing Values
Time	26,328	Object	0
Temperature_2m (°C)	26,305	Float64	23
Relative_humidity_2m (%)	26,305	Float64	23
Precipitation (mm)	26,305	Float64	23
Rain (mm)	26,305	Float64	23
Snowfall (cm)	26,305	Float64	23
Pressure_msl (hPa)	26,305	Float64	23
Wind_speed_10m (km/h)	26,305	Float64	23
Wind_speed_100m (km/h)	26,305	Float64	23
Soil_temperature_0_to_7 cm(°C)	26,305	Float64	23

3.2 Data Preprocessing

Raw data is preprocessed into a real-time machine-learning compatible format, i.e. time column is converted into digits. Missing values are filled with the forward-fill technique, also called the last observation carried forward approach, which is necessary for the time series continuity. The standard scaler method is used for data scaling for continuous variables with a zero mean. In feature selection, the correlation matrix is created to discover the close relationship between the features for effective results.

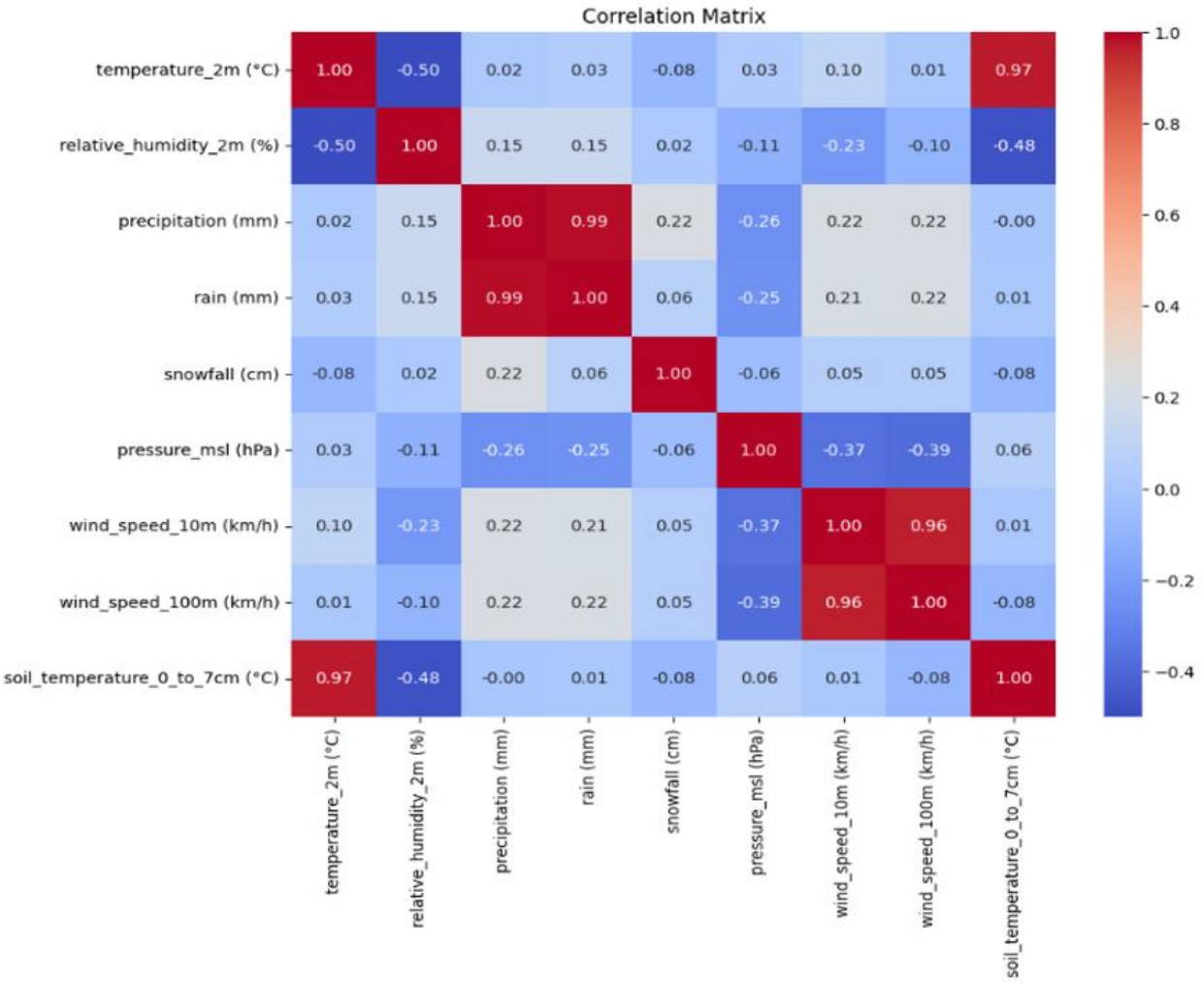

Figure 2: Correlation Matrix

In **Figure 2**,The correlation matrix shows that four independent variables have a great impact on the precipitation as a target variable i.e. temperature_2m, wind speed_10m, pressure, and soil temperature. The time series plot of the selected features and predicted variable from 7/8/2021 to 7/8/2024 is shown in **Figure 3**.

3.3 Machine Learning Algorithms

In this research, the following machine learning algorithms are developed, trained, tested, and their results are compared.

3.3.1 Linear Regression

This machine-learning algorithm describes the relationship between the independent variables with the dependent variable. The dependent variable is calculated as a linear combination of independent variables, as shown in (1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

3.3.2 KNN Regression

This algorithm averages the values of the k nearest points for calculating the output quantity as depicted in (2).

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i \quad (2)$$

In (2),y_i are the outputs of the k-nearest neighbors.

3.3.3 Support Vector Machine Regression

This algorithm is used to deal with non-linear data patterns. An optimal hyperplane is drawn to create the largest distance between data. Mathematically, SVM regression is shown in (3).

$$f(x) = w \cdot x + b \quad (3)$$

In equation (3), x is the input vector, b is the bias term, w is the weight vector.

3.3.4 Random Forest Regression

This algorithm combines multiple decision trees to make a single model. Each tree in the forest builds from its subset of data and makes its prediction. The final prediction is based on the average of all the decision tree predictions. Mathematically, it is given in equation (4).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t \quad (4)$$

Where $T$is the number of trees, and $\hat{y}_t$ shows prediction from each tree.

3.3.5 Performance Evaluation

Four metrics are used as a performance measure for the models developed. First comes the accuracy score computed as

$$\frac{TN + TP}{TP + FP + TN + FN}$$

Followed by Precision, computed as

$$\frac{TP}{FP + TP}$$

The third one is Recall, given by

$$\frac{TP}{TP + FN}$$

Fourth is the F1 score, computed as

$$2 \times \frac{Precesion \times Recall}{Precision + Recall}$$

According to the confusion matrix, TP stands for True positives, TN stands for True negatives, FP stands for False positives, and FN stands for False negatives.

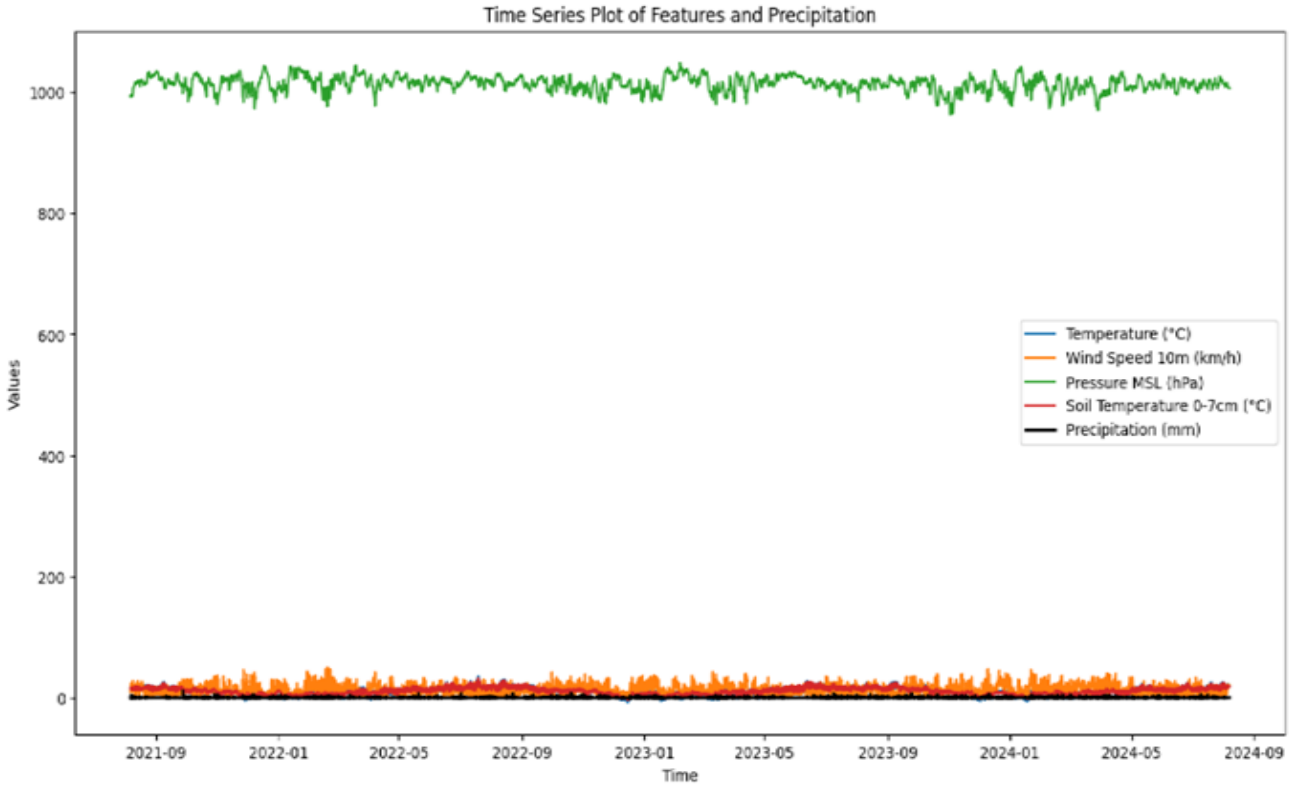

Figure 3: Time Series plot for the dependent variable and the independent variable

4. RESULTS AND DISCUSSION

Figure 4shows a scatter graph of a linear regression model in which actual precipitation (mm) is compared with the predicted precipitation (mm). The red line shows the perfect prediction of actual values with the predicted value. Scattered blue dots show the actual and predicted data points. From the result, it is evident that precipitation with low values isbetter predicted than with high values of precipitation. This depicts that linear regression finds difficulty in this dataset. The complex relation between the input and output variables needs more tuning for the correlation.

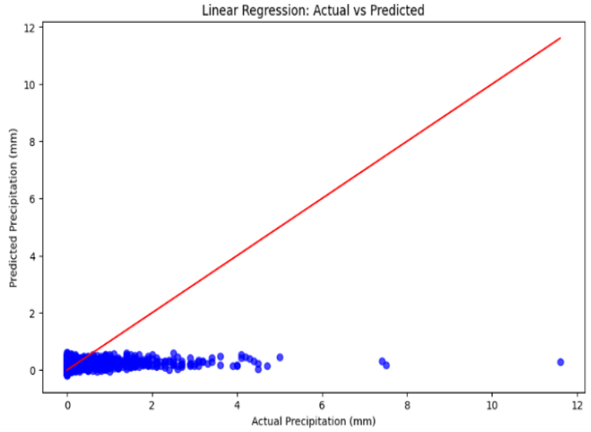

Figure 4: Linear Regression Results

Figure 5 below describes the results of the KNN model for the prediction of precipitation with actual values of precipitation. The green scattered data points show the actual and predicted values for precipitation. The red line shows the exact scenario where the predicted values should match the actual values. This model works fine for lower values of precipitation with the independent variables relationship but

for high values of precipitation, this model shows considerably satisfactory results.

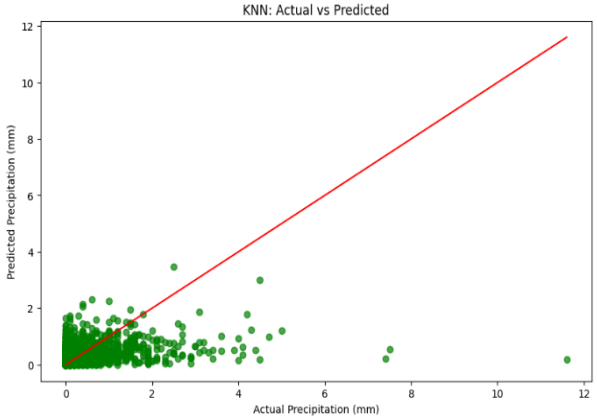

Figure 5: KNN Regression Results

Figure 6 shows the scattered graph of the SVM regression model. This model performs well for the values from 0 – 2mm of precipitation, and from 2 – 3mm, moderate performance is recorded, but with higher values, the model underperforms. The need for hyperparameter tuning is much needed for better results.

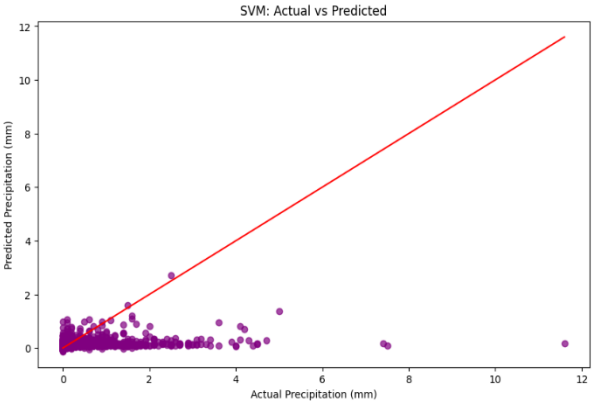

Figure 6: SVM Regression Results

The results of the random forest regression model are shown in **Figure 7**. This model performs well in predicting precipitation for the actual values. This means that the independent variables somehow correlate a strong relationship with the target variable. Results can be further improved by collecting large datasets and hyperparameter tuning of the features for the best relationship.

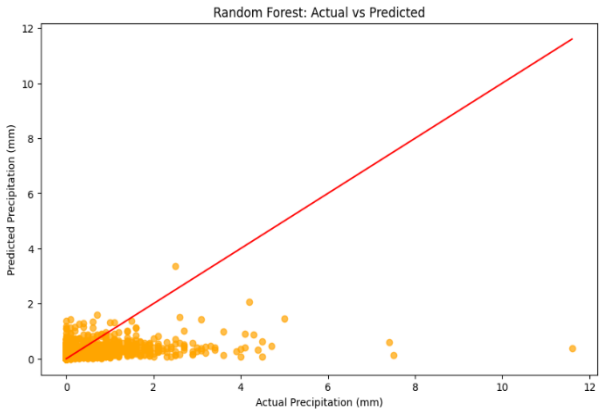

Figure 7: Random Forest Regression Results

Table 2and **Figure 8** show the summary of four models with three key metrics for the prediction of precipitation. KNN has the lowest MSE and MAE values (0.177 & 0.150), which showcase the lowest prediction error as compared to other models. In terms of prediction errors, RF regression also performed well with MSE (0.185) and MAE (0.174). Despite the lowest MSE (0.203) for SVM, it failed to perform as KNN and RF. Among all models, linear regression has the highest MSE (0.205) and lowest R^2 (0.070).

Table 2: Models Performance comparison

Model	**MSE**	**R^2 Score**	**MAE**
Linear Regression	0.205888	0.070245	0.198003
KNN	0.177875	0.196746	0.150809
SVM	0.203173	0.082505	0.172193
Random Forest	0.185492	0.162348	0.174671

Table 3 and **Figure 9** Show the classification matrix comparison. Despite SVM's high accuracy, it does not mean that this model has the highest number of samples classified correctly, especially with the imbalanced dataset, because it does not consider false negatives and false positives. SVM has the highest precision (0.63), which is greatly important to reduce the false positive predictions. KNN has an accuracy of 91% with moderate precision, F1 score, and recall.

Table 3: Class score comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Linear Regression	0.922522	.411765	0.034826	0.064220
KNN	0.911508	0.417526	0.402985	0.410127
SVM	0.925940	0.636364	0.069652	0.125561
Random Forest	0.918534	0.432836	0.216418	0.288557

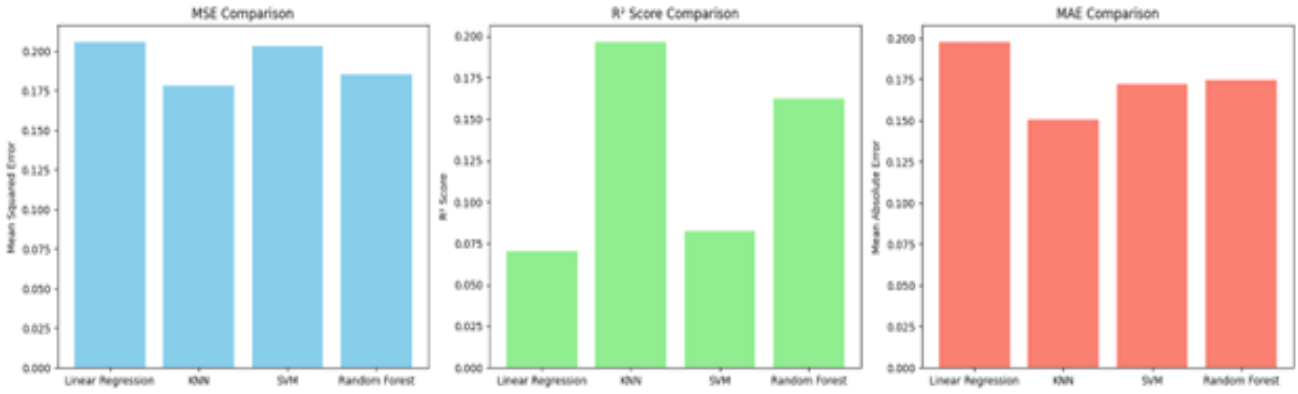

Figure 8: Models Performance plot

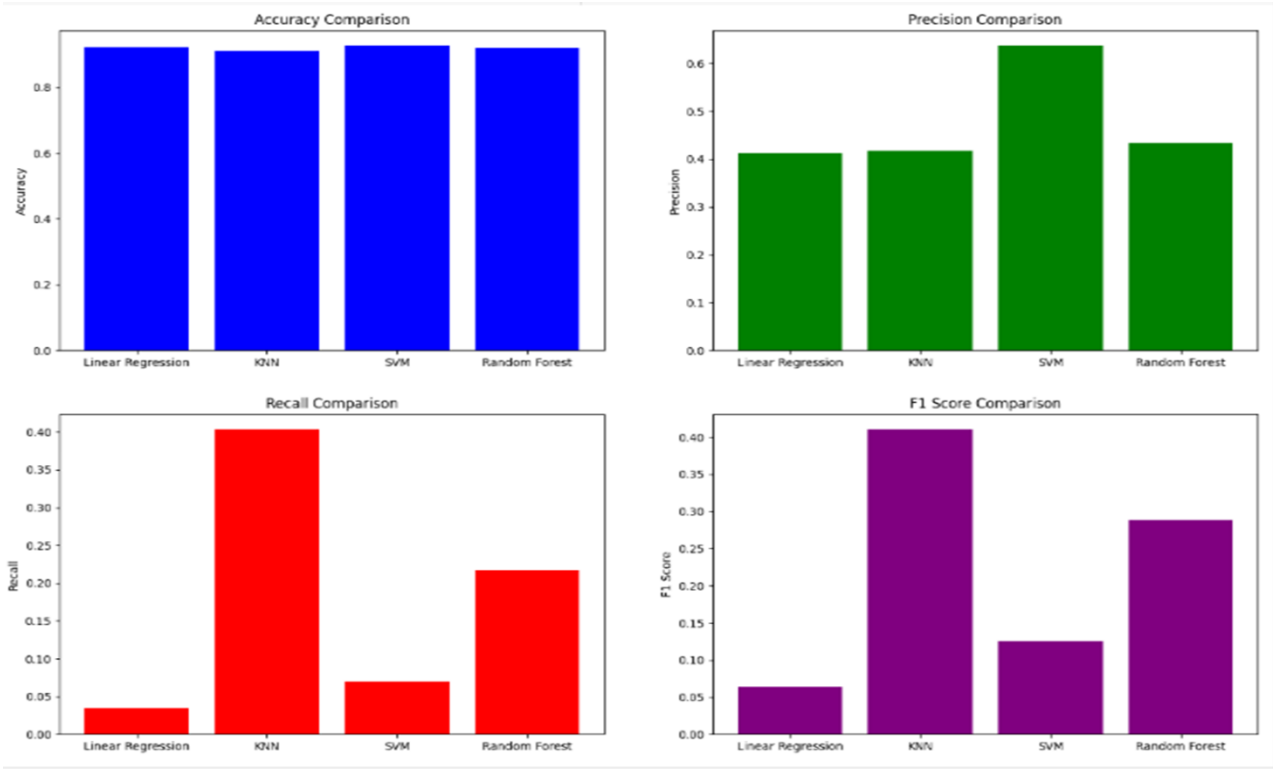

Figure 9: Class score comparison plot

5. CONCLUSION

In this paper, four machine learning models are evaluated for weather prediction for Manchester City. Precipitation is predicted with respect to pressure, temperature, wind speed, and soil temperature, by implementing linear regression, support vector machine regression, k-nearest neighbor regression, and random forest regression for the dataset of three years. Historical data for three years is collected, preprocessed, and split for training and testing machine learning models. Different performance indicators like MSE, R^2, and MAE are measured in order to assess performance of model. All model performances are evaluated with SVM and LR topping in accuracy (92.59% and 92.25%).

5.1 Future Recommendations

In this study, several improvements can be made. Increasing the number of meteorological input variables and increasing the dataset size for more than five years can help the model predict extreme events. Hyperparameter tuning can also improve the prediction by finding the best correlation of the input variables. Further ML algorithm hyperparameters can be optimized to improve performance. Depending upon the type of dataset, Deep learning techniques like Convolutional NeuralNetwork (CNN) and Recurrent Neural Network (RNN) can also be implemented for better results.

Acknowledgements: The authors would like to acknowledge the valuable contributions of Mr. Yijun Wang in supporting the revision of this manuscript.

The authors would also like to acknowledge the University of Bolton for providing a supportive academic environment that facilitated the development and completion of this research.

6. REFERENCES

[1] Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting” Stanford University, 2016.

[2] Young, M.V. and Grahame, N.S. (2022). The history of UK weather forecasting: the changing role of the central

guidance forecaster. Part 2 : the birth of operational numerical weather prediction. *Weather*.

[3] Zhao, Q., Liu, Y., Yao, W. and Yao, Y. (2022). Hourly Rainfall Forecast Model Using Supervised Learning Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 60, pp.1–9.

[4] Chen, L., Han, B., Wang, X., Zhao, J., Yang, W. and Yang, Z. (2023). Machine Learning Methods in Weather and Climate Applications: A Survey. *Applied Sciences*, 13(21), p.12019.

[5] Zhang, H., Liu, Y., Zhang, C. and Li, N. (2025). Machine Learning Methods for Weather Forecasting: A Survey. *Atmosphere*, [online] 16(1), pp.82–82.

[6] Wang, J., Wang, Z., Ye, J., Lai, A., Ma, H. and Zhang, W. (2023). Technical Evaluation of Precipitation Forecast by Blending Weather Radar Based on New Spatial Test Method. *Remote Sensing*, 15(12), p.3134.

[7] Ashok, S.P. and Pekkat, S. (2022). A systematic quantitative review on the performance of some of the recent short-term rainfall forecasting techniques. *Journal of Water and Climate Change*, 13(8), pp.3004–3029.

[8] Verma, S., Srivastava, K., Tiwari, A. and Verma, S. (2023). *Deep Learning Techniques in Extreme Weather Events: A Review*. [online] arXiv.org.

[9] Mukkavilli, S.K., Civitarese, D.S., Schmude, J., Jakubik, J., Jones, A., Nguyen, N., Phillips, C., Roy, S., Singh, S., Watson, C., Ganti, R., Hamann, H., Nair, U., Ramachandran, R. and Weldemariam, K. (2023). *AI Foundation Models for Weather and Climate: Applications, Design, and Implementation*. [online] arXiv.org.

[10] Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Morteza Mardani, Kurth, T., Hall, D.H., Li, Z., Kamyar Azizzadenesheli, Hassanzadeh, P., Karthik Kashinath and Animashree Anandkumar (2022). FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *arXiv (Cornell University)*.

[11] Mounia El Hafyani, Khalid El Himdi and Salah-Eddine El Adlouni (2024). Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique. *Frontiers in Water*, 6.

[12] García-Vázquez, F., Ponce-González, J.R., Guerrero-Osuna, H.A., Carrasco-Navarro, R., Luque-Vega, L.F., Mata-Romero, M.E., Martínez-Blanco, M. del R., Castañeda-Miranda, C.L. and Díaz-Flórez, G. (2023). Prediction of Internal Temperature in Greenhouses Using the Supervised Learning Techniques: Linear and Support Vector Regressions. *Applied Sciences*, [online] 13(14), p.8531.

[13] Abdulraheem, M.; Awotunde, J.B.; Abidemi, E.A.; Idowu, D.O.; Adekola, S.O. Weather prediction performance evaluation on selected machine learning algorithms. IAES Int. J. Artif. Intell. 2022, 11, 1535

[14] Suman Markuna, Kumar, P., Ali, R., Dinesh Kumar Vishwkarma, Kuldeep Singh Kushwaha, Kumar, R., Vijay Kumar Singh, Chaudhary, S. and Alban Kuriqi (2023). Application of Innovative Machine Learning Techniques for Long-Term Rainfall Prediction. *Pure and applied geophysics*, 180(1), pp.335–363.

[15] Open-meteo.com. (2022). *Docs | Open-Meteo.com*. [online] Available at: https://open-meteo.com/en/docs#latitude=53.4809&longitude=-2.2374 [Accessed 1 Sep. 2024.

7. AUTHOR'S PROFILE

Dr. Rameez Asif holds a Ph.D. in RF and Microwave Engineering and an M.Sc. with Distinction from the University of Bradford. As a KTP Antenna Engineer, he developed a patented electromagnetic radiation shield featured in Defense Online for its defense-grade metamaterial applications.

Following his Ph.D., he worked as a Senior RF Design Engineer at Visibility Asset Management Ltd., leading the deployment of RFID systems tracking over £200M in industrial assets, and achieving read ranges beyond 16m on metallic surfaces. He also served as an Operational Consultant for Slymm Engineering.

In 2020, he returned to academia as a Postdoctoral Research Associate in the FUN Research Group, contributing to the Horizon 2020 SINAPSE project on secure aeronautical communications. He now lectures at the University of Greater Manchester and is a Fellow of HEA with a Distinction in PGCert.

Dr. Asif's research spans metasurfaces, RIS, AI-driven wireless systems, and secure biomedical sensing. He is a Technical Editor for Future Internet, IEEE reviewer, and mentor to students now employed at Intel, BAE Systems, NHS, and KPMG. In 2024, he received the Outstanding Lecturer – School of Engineering Award.

Syed Hamid Ali Shah earned a BSc in Electrical Power Engineeringfrom COMSATS University Abbottabad (2014), an MSc in Electrical Engineering (Power & Control) from CECOS University Peshawar (2019), and an MSc in Electrical and Electronic Engineering from the University of Bolton, UK (2024). His research bridges machine learning and power system applications, focusing on weather prediction models to enhance forecasting for grid operation and renewable energy integration. His work reflects a strong interest in data-driven solutions for resilient and intelligent power systems