21st International Conference on Sustainable Energy Technologies

12 to 14th August 2024, Shanghai, China

Sustainable Energy Technologies 2024
Conference Proceedings: Volume 4

WSSET

World Society of Sustainable Energy Technologies

Proceedings of the 21st International Conference on Sustainable Energy Technologies

SET2024

12 – 14th August 2024, Shanghai, China

Edited by
Professor Saffa Riffat

Supported by SET2024 Conference Organising Committee:

Chair: Professor Saffa Riffat
Co-Chair: Professor Zhongzhu Qiu
Editors: Dr Ziwei Chen
Dr Tianhong Zheng
Dr Yanan Zhang
Dr Yi Fan
Ms Zeny Amante-Roberts

Reference to a conference paper:
To cite a paper published in these conference proceedings, please substitute the highlighted sections of the reference below with the details of the article you are referring to:
Author(s) Surname, Author(s) Initial(s), 2024. 'Title of paper'. In: Riffat, Su. ed., **Sustainable Energy Technologies**: Proceedings of the 21st International Conference on Sustainable Energy Technologies, 12-14th August 2024, Shanghai, China. University of Nottingham: Buildings, Energy & Environment Research Group. Pp XX-XX. Available from: nottingham-repository.worktribe.com/ [Last access date].

ISBN-13 978-0-85358-356-1
Version: 01.12.2024

#530: The implementation of real-time weather forecasting system using internet of things and machine learning

Bhavesh PANDYA[1], Amin Al-HABAIBEH[2], Greg ROLLESTON[3], Michael FARNSWORTH[4], Drashya BANSAL[5]

[1] School of Science and Technology, Nottingham Trent University, United Kingdom, bhavesh.pandya@ntu.ac.uk
[2] Product Innovation Centre, Nottingham Trent University, United Kingdom, Amin.Al-Habaibeh@ntu.ac.uk
[3] Stormsaver Ltd., Newark, Nottinghamshire, United Kingdom, g.rolleston@stormsaver.com
[4] Stormsaver Ltd., Newark, Nottinghamshire, United Kingdom, m.farnsworth@stormsaver.com
[5] Indian Institute of Technology, Madras, India, drashyabansal@gmail.com

Abstract: This research explores the application of advanced technologies to enhance the precision, accessibility, and effectiveness of weather forecasts, presenting benefits to individuals, businesses, and communities. The research utilises data collected from IoT-enabled weather stations, employing appropriate IoT protocols for data acquisition. This data undergoes pre-processing and normalisation to facilitate further analysis. The primary objective is to monitor and predict weather changes, specifically focusing on the probability of rainfall in London. This is achieved through a detailed examination of various meteorological parameters such as temperature, humidity, wind speed, dew point, and wind gusts. Through feature engineering, critical predictors are identified and optimised by eliminating redundant elements, thus refining the model's efficiency. Key features such as temperature, humidity, wind speed, gusts, air pressure, and dew point are analysed alongside temporal variables like time of day, day of the week, and seasonal patterns. The weather conditions are classified into three categories: Cloudy, Fair, and Rain. The dataset spans from 2014 to 2023, with a 70% split for training and 30% reserved for testing. Upon evaluating 17 distinct classifiers, the Support Vector Machine (SVM) classifier emerged as the most effective, demonstrating an 88% recall, 70% precision, and a 78% F1-score. These findings highlight the potential of integrating machine learning with real-time weather monitoring to predict weather patterns accurately.

Keywords: Real Time Weather Monitoring; Iot; Machine Learning; Support Vector Machine; AI

1. INTRODUCTION

Weather forecasting is vital for agriculture, transportation, energy, and disaster management. Traditionally, it relies on data from satellites, radar systems, weather stations, and weather balloons (McIlveen, 1992). The advent of the Internet of Things (IoT) has enhanced weather forecasting by providing real-time, high-resolution data from interconnected devices (Balakrishnan, 2021). IoT devices include sensors and actuators that collect, transmit, and receive meteorological data such as temperature, humidity, air pressure, wind speed, and precipitation. These devices can be placed in various locations, including urban and rural areas, oceans, and the atmosphere. The data collected is sent to centralised servers for analysis and integration into machine learning-based weather prediction models (Banara, 2022). IoT offers advantages such as high spatial resolution, real-time data collection, and cost-effectiveness, improving the accuracy and reliability of forecasts. However, challenges such as data quality, calibration, security, privacy, interoperability, and connectivity must be addressed. Weather forecasting involves predicting future atmospheric conditions based on historical data and current observations. Recently, machine learning (ML) techniques have been increasingly utilised to complement traditional numerical weather prediction models, enhancing forecasting accuracy and efficiency. By using historical weather data, meteorological observations, and advanced ML algorithms, forecasters can develop models that capture the complex interactions and nonlinear dynamics of the atmosphere (Parmar, 2017). Integrating ML techniques can improve short-term and localised weather predictions, offering valuable insights for mitigating extreme weather events and climate change impacts. Techniques used include regression models, time series analysis, neural networks, ensemble learning, and deep learning. Support Vector Machine (SVM) is a powerful supervised learning algorithm popular in weather forecasting. It is useful for classification and regression tasks, making it applicable to various weather prediction aspects. This work explores the application of SVM techniques in weather forecasting and their advantages in capturing complex relationships within meteorological data. (Waqas, 2023).

2. CURRENT STATE-OF-THE-ART

Researchers have improved the accuracy of daily, monthly, and annual rainfall prediction by utilising machine learning algorithms, big data analysis, and data mining approaches. Their findings indicate that machine learning approaches are replacing data mining techniques in the prediction process and outperform conventional deterministic approaches in forecasting rainfall and weather (Kowar, 2012). This research examined various machine learning techniques to determine the best algorithms for accurately predicting rainfall in London. Below is a summary of related state-of-the-artwork to provide a better understanding and analysis of the research problem. Praba et al (2018) have developed a weather forecasting system leveraging IoT to provide real-time data, stored in the cloud, and used the Support Vector Machine method to forecast rainfall. Verma et al (2020) have implemented a real-time weather forecast system using various sensors and uploaded data to a ThingSpeak cloud server using an ESP8266-01 module and NodeMCU. They used a logistic regression model trained on pre-recorded sensor data, achieving an 84% prediction accuracy, or performing models using Artificial Neural Networks and Decision Trees. Sankarnarayan et al (2020) have proposed a deep neural network model to forecast flooding likelihood based on temperature and rainfall intensity. They found that deep neural networks had the best accuracy (89.71%) compared to other models using SVM, KNN, and NB.

In order to assess the effectiveness of eight statistical and machine learning models that use atmospheric synoptic patterns to predict long-term daily rainfall in a semi-arid climate (Tenerife, Spain), Diez-Sierra and Jesus (2020) assessed the effectiveness of eight statistical and machine learning models in predicting long-term daily rainfall in Tenerife, Spain. They found that NN was the most accurate method for daily rainfall prediction, with an average f-score of about 0.4, ahead of LR and SVM. Liyew et al. (2021) have examined machine learning techniques for rainfall prediction using data from the Bahir Dar City meteorological station in Ethiopia. They compared MLR, RF, and XGBoost, finding XGBoost performed bestin predicting daily rainfall using specific environmental data. Moreover, Tharun et al. (2018) have compared statistical modelling and regression techniques (SVM, RF, and DT) for rainfall prediction. They found that regression techniques outperformed statistical models, with the RF model showing the highest accuracy and performance. Garg and Pandey (2019) have compared SVM, SVR, and KNN for annual rainfall prediction, finding that SVM performed the best among the three methods. Subia and Ashwitha (2022) have proposed a hybrid algorithm combining CNN and LSTM with a first-order optimisation technique using gradient descent. This approach increased system accuracy to 87.3%, improving upon the separate CNN and LSTM frameworks. Sadhukhan et al. (2021) have used low-cost IoT devices equipped with GPS to forecast meteorological variables. They evaluated various computational tools, including SVM, KNN, DNN, Ridge, Linear Regression, and ANN, on provided parameters. (Sadhukhan, 2021). In addition, Kaushik et al. (2020) have examined KNN, ELM, and SVM algorithms for practical problem-solving, finding that SVM provided the best results with the lowest ET, MAE, and RMSE. ELM was found to be CPU-intensive, while KNN and ELM might have performed better with more parameters and data. From the above, it can be argued that further research is still needed to explore the utilisation of IoT and ML is predicting weather conditions.

3. MATERIALS AND METHODS

Figure 1 presents the flowchart of the methodology used. The process starts with data collection and processing. Following the data understanding, feature generation is developed. Created features are processed and used for model training. A tuning process is developed, and the best performing model is utilised for the system's operation and forecasting.

3.1. Location of the study

Geographical location for this study is city of London, which is situated in south-eastern England, within the United Kingdom. It is located along the River Thames, making it a prominent city in terms of both historical and contemporary significance. Geographically, London is located at approximately 51.5074° N latitude and 0.1278° W longitude.

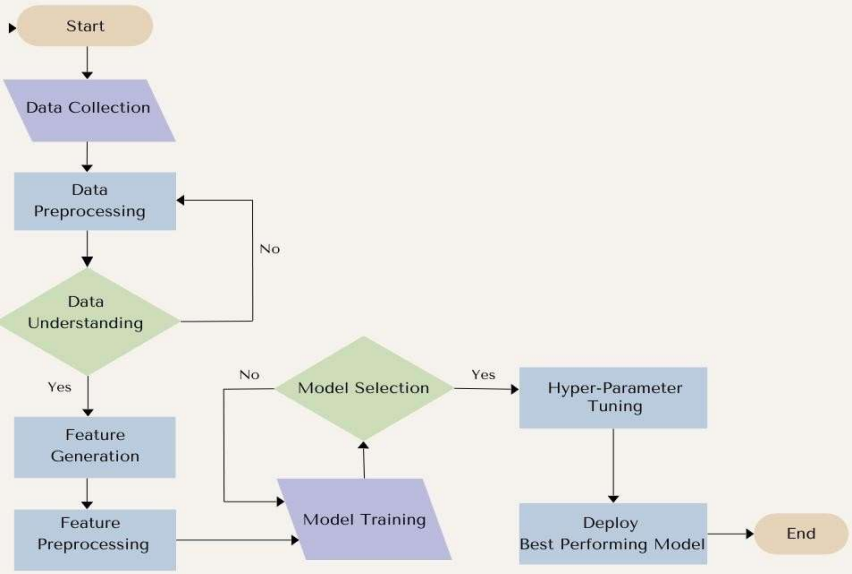

Figure 1: Flowchart of the proposed work

3.2. Data

Weather data sets are collected from https://www.wunderground.com/history/weekly/gb/london/EGLC/ during January 2014 and January 2024. The collected data sets include the following parameters:

***Precipitation Data**:* This includes historical rainfall measurements over time. It is collected in the form of rainfall amounts (in millimetres or inches) over specific time intervals (e.g., hourly, daily, monthly).

***Temperature Data**:* Temperature influences the likelihood and type of precipitation. Data on temperature, both historical and forecasted, are essential for understanding atmospheric conditions that can lead to rainfall.

***Humidity Data**:* Humidity levels in the atmosphere play a crucial role in determining the likelihood and intensity of rainfall. High humidity levels indicate moisture in the air, which can contribute to rainfall formation.

***Atmospheric Pressure Data**:* Changes in atmospheric pressure can affect weather patterns and influence rainfall. Monitoring atmospheric pressure provides insights into the movement and behaviour of weather systems.

***Wind Data**:* Wind direction and speed impact the movement and distribution of weather systems, including rainfall. Wind data helps in understanding how weather patterns propagate and where precipitation may occur.

***Dew Point Data**:* Dew point is the temperature at which air becomes saturated with moisture, leading to condensation and potentially precipitation. Monitoring dew point provides information about the moisture content of the air.

***Cloud Cover Data**:* Cloud cover affects incoming solar radiation and atmospheric dynamics, influencing rainfall patterns. Observing cloud cover helps in understanding the likelihood of precipitation.

***Topographical Data**:* Geographic features such as mountains, valleys, and bodies of water can significantly influence local weather patterns and rainfall distribution. Understanding the topography of the area helps in predicting rainfall accurately.

3.3. Data Mapping

Multiple conditions of weather data obtained from the above data source have been categorised into three categories namely: *Cloudy, Fair* and *Rain* as per Table 1.

Table 1: Mapping of weather conditions

Weather Condition	Mapped as	Weather Condition	Mapped as
Snow Grains	Rain	Rain / Windy	Rain
Light Rain		Thunder	
Light Rain / Windy		Light Snow	
Rain		Light Snow Shower	
Light Rain Shower		Rain and Snow	
Light Drizzle		Drizzle / Windy	
Rain Shower		Rain and Snow / Windy	
Showers in the Vicinity		Light Freezing Drizzle	
Drizzle		Heavy Rain Shower / Windy	
Heavy Rain Shower		Light Rain with Thunder	
Heavy Drizzle		Light Snow / Windy	
Light Rain Shower / Windy		Light Sleet	
Rain Shower / Windy		Heavy Rain / Windy	
Light Drizzle / Windy		Heavy T-Storm	
T-Storm		Light Freezing Rain / Windy	
Thunder in the Vicinity		Freezing Rain	
Heavy Rain		Snow	
Heavy T-Storm / Windy		Rain / Fog	
T-Storm / Windy			
Weather Condition	**Mapped as**	**Weather Condition**	**Mapped as**
Mostly Cloudy	Cloudy	Wintry Mix	Fair
Partly Cloudy		Fair	
Cloudy		Fair / Windy	
Cloudy / Windy		Wintry Mix / Windy	
Mostly Cloudy / Windy		Light Sleet / Windy	
Partly Cloudy / Windy		Snow Grains / Windy	
Haze		Haze / Windy	
Mist		------ This area is intentionally left blank -----	
Fog			
Hail			
Hail / Windy			
Thunder and Hailstorm			
Patches of Fog			
Small Hailstorm			
Thunder and Hail / Windy			
Thunder and Small Hailstorm			
Partial Fog			

3.4. Data Cleaning

The following method has been adopted for cleaning the data as shown in Figure 2. Data will normally need to be checked for missing (Null) data. Once the data is formatted based on the expected categories, columns with 80% or more of Null or missing data is removed. Data is then enhanced, and correct information is inserted in relevant columns, see Figure 2.

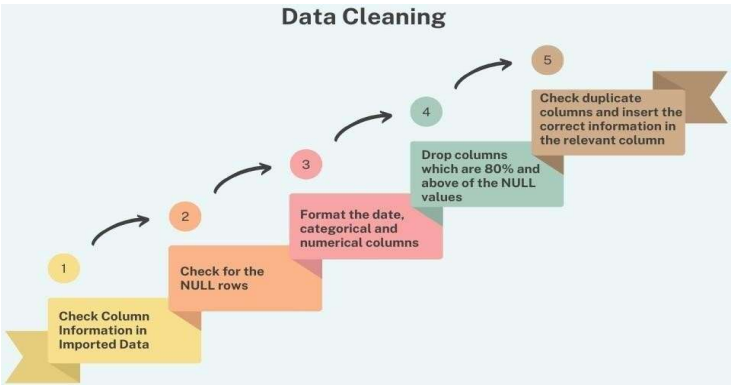

Figure 2: Flow Chart used for data cleaning

3.5. Feature Generation

The following methodology has been adopted for feature generation as presented in Figure 3. The three-year average for specific day, month and year is calculated. Following which, last year's daily average is calculated. Then the last three days average is calculated. Merge all information and past day data to create the required features.

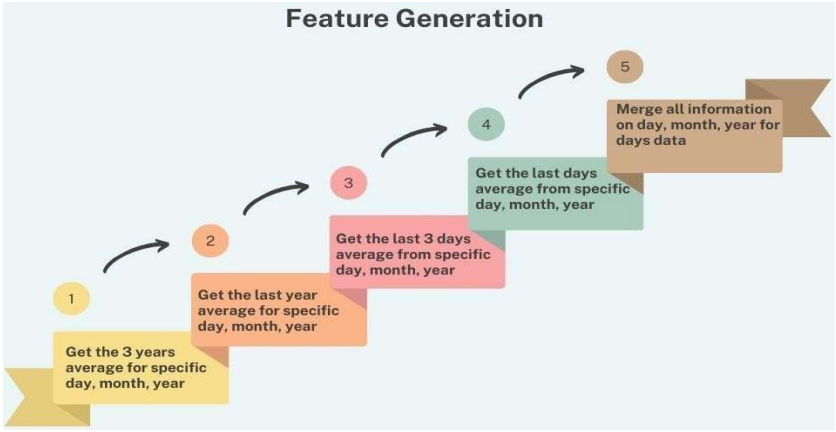

Figure 3: Flow Chart used for feature generation

3.6. Feature Creation

From the information collected from the data source and the flowchart as depicted above, the features are created. The average information for the current date from the past year has been collected from the source based on Table 2-a below.

Table 2(a): Feature creation (Past 1 year)

Feature	Description
LY1TEMP	Last year's average temperature
LY1DEWP	Last year's average dew point
LY1WINSP	Last year's average wind speed
LY1WINGT	Last year's average wind gust
LY1RAIN	Last year's average number of rainy days

Similarly, the average information for the current date for the last three years has been collected from the source as correlated in Table 2-b.

Table 2(b): Feature creation (Last 3 years)

Feature	Description
LY3TEMP	Last three year's average temperature
LY3DEWP	Last three year's average dew point
LY3WINSP	Last three year's average wind speed
LY3WINGT	Last three year's average wind gust
LY3RAIN	Last three year's average number of rainy days

Also the information about the rain during last year on the current date and number of times it rained in last 3 years are collected from the data source.

The information about the last day (*L1TEMP, L1DEWP, L1WINSP, L1WINGT, L1RAIN*) and last 3 days (*L3TEMP, L3DEWP, L3WINSP, L3WINGT, L3RAIN*) is collected specifically to get the recent weather information.

With the above historical information, the current date information is merged to get the current values as shown in Table 3-a.

Table 3(a): Descriptions of features (Average)

Feature	Description
CUR3TEMPAVG	Current temperature with respect to current three days' average temperature
CUR3DEWPAVG	Current dew point with respect to current three days' average dew point
CUR3WINSPAVG	Current wind speed with respect to current three days' average wind speed
CUR3WINGTAVG	Current wind gust with respect to current three days' average wind gust

Similarly, the above features with respect to last three years' current date average are also calculated as shown in Table 3-b.

Table 3(b): Descriptions of features (Last 3 years)

Feature	Description
CUR3TEMPY3AVG	Current temperature with respect to last three years' current date average temperature
CUR3DEWPY3AVG	Current dew point with respect to last three years' current date average dew point
CUR3WINSPY3AVG	Current wind speed with respect to last three years' current date average wind speed
CUR3WINGTY3AVG	Current wind gust with respect to last three years' current date average wind gust

The following stage of the work is to create the training and testing dataset, using train_test_split with test_size of 15% and stratify on the basis of target column. The min-max value of all the indicators is obtained from the training data as outlinedin Table 4.

Table 4: Min-Max values of the features

Feature #	Indicator	Min value	Max value
1	*CUR3TEMPAVG*	0.63	1.632
2	*CUR3DEWPAVG*	0.581	2.265
3	*CUR3WINSPAVG*	0.241	3.079
4	*CUR3WINGTAVG*	0.0	76.358
5	*CUR3TEMPY3AVG*	0.547	1.542
6	*CUR3DEWPY3AVG*	0.4	1.668
7	*CUR3WINSPY3AVG*	0.137	4.02
8	*CUR3WINGTY3AVG*	0.0	126.706
9	*CURTEMPAVG*	0.642	1.691
10	*CURDEWPAVG*	0.646	1.919
11	*CURWINSPAVG*	0.189	3.724
12	*CURWINGTAVG*	0.0	28.207
13	*CURTEMPY3AVG*	0.538	1.918
14	*CURDEWPY3AVG*	0.386	2.382
15	*CURWINSPY3AVG*	0.123	7.294
16	*CURWINGTY3AVG*	0.0	58.176

From the above, training and testing data are created as X_train, X_test, y_train, and y_test with indicators in X and targets in y. The simple imputer technique is applied to fill NULL rows in the test set, while NULL values are dropped in the training set to ensure training occurs on accurate and correct data. Standard Scalar was utilised for numerical columns and One Hot Encoder for categorical columns. Based on this setup, multiple prediction models as detailed below have been designed and tested:

***Dummy Classifier**:* A Dummy Classifier is a baseline model that helps assess the performance of more complex models by comparing them against simple strategies like random guessing (Bishop, 2006).

Gaussian NB: Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem. It's computationally efficient and works well with small datasets but assumes feature independence (Langley, 1992).

***Decision Tree Classifier**:* Decision trees are intuitive models that split data into branches to make predictions. They handle numerical and categorical data well but can be overfit without techniques like pruning (Kumar, 2013).

***SGD Classifier**:* The Stochastic Gradient Descent Classifier is a linear model efficient for large datasets. It updates model parameters iteratively, making it scalable and suitable for high-dimensional data (Paquin et al., 2023).

***XGB Classifier**:* XGBoost is a powerful gradient boosting algorithm known for its speed and performance. It builds trees sequentially, optimising them to minimise errors and prevent overfitting (Raihan, 2023).

***Extra Trees Classifier**:* Extra Trees is an ensemble method that creates multiple trees using random subsets of data and features. It trains quickly and is less prone to overfitting compared to traditional decision trees (Tina, 2022).

***Gradient Boosting Classifier**:* This classifier builds multiple weak models (usually trees) sequentially to improve predictions. Each model corrects errors from the previous one, enhancing overall accuracy (Rok, 2017).

***K Nearest Neighbors Classifier**:* KNN assigns class labels based on the majority vote of the nearest neighbors in the feature space. It's simple and effective but can be computationally intensive with large datasets (Taunk K., 2019).

***Random Forest Classifier**:* Random Forests build multiple decision trees on random data subsets and aggregate their predictions. This method reduces overfitting and improves accuracy over single decision trees (Gnuer, 2010).

***Tuned SVC Classifier**:* Support Vector Classifier (SVC) separates classes with a hyperplane. Tuning its parameters, such as the regularisation parameter and kernel type, enhances performance (Tang, 2012).

***Linear SVC**:* Linear SVC is a variant of SVM that finds a hyperplane for linearly separable data. It's efficient for large-scale datasets and robust to overfitting (Burges, 1998).

Tuned Linear SVC: Optimising Linear SVC by tuning hyperparameters like regularisation improves its performance, making it suitable for linearly separable datasets where efficiency and interpretability are crucial (Zhou, 2011).

***Logistic Regression**:* Logistic Regression models the probability of class membership using the logistic function. It's interpretable and effective for binary and multi-class classification tasks (Maalouf, 2011).

***SVC**:* Support Vector Classifier (SVC) aims to find the best separating hyperplane between classes, handling both linear and non-linear data with various kernel functions (Keerthi, 2000).

4. RESULTS

A total of 17 models have been designed and tested with the data mined from the data source previously mentioned. The performance matrix during training and testing sessions used are Accuracy, F1 Score, Recall and Precision. A brief about these parameters is provided below.

4.1. Accuracy

Training Accuracy: It refers to the percentage of correctly classified instances out of the total instances in the training dataset. This metric indicates how well the model has learned the patterns in the data that it was trained on (Angra and Alhuja, 2017). A high training accuracy suggests that the model has successfully captured the relationships in the training data.

$$\text{Training Accuracy} = \frac{\text{Number of correct predictions on training data}}{\text{Total number of training instances}} \times 100\% \quad (1)$$

Testing Accuracy: It refers to the percentage of correctly classified instances out of the total instances in a separate testing dataset that the model has not seen during training (Medar et al., 2017). This metric measures the model's ability to generalise new, unseen data. A high testing accuracy indicates that the model is likely to perform well on real-world data.

$$\text{Testing Accuracy} = \frac{\text{Number of corred predictions on testing data}}{\text{Total number of testing instances}} \times 100\% \quad (2)$$

4.2. Recall

It is also known as *sensitivity* or true positive rate which is a metric used to evaluate the performance of a classification model. It measures the ability of the model to correctly identify all relevant instances (i.e., all actual positives) (Feras, 2020). Recall for training data refers to the recall calculated based on the predictions that the model makes on the training dataset. This metric indicates how well the model identifies positive instances from the data it was trained on. Recall for testing data refers to the recall calculated based on the predictions that the model makes on the testing dataset. This metric indicates how well the model generalises its ability to identify positive instances to new, unseen data. A high training recall suggests the model effectively identifies positive instances in the training dataset. A high testing recall suggests the model maintains this capability with new data, indicating good generalisation.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

Where: True Positives (TP) are instances correctly predicted as positive and False Negatives (FN) are instances incorrectly predicted as negative (i.e., actual positives that were missed).

4.3. Precision

Also known as positive predictive value is a metric used to evaluate the performance of a classification model. It measures the accuracy of the positive predictions made by the model. Training precision refers to the precision calculated based on the predictions the model makes on the training dataset (Bansal and Singhrova, 2021). This metric indicates how many of the instances predicted as positive in the training data are actually positive. Testing precision (or validation precision) refers to the precision calculated based on the predictions the model makes on the testing dataset. This metric indicates how many of the instances predicted as positive in the testing data are actually positive. A high training precision suggests the model accurately identifies positive instances in the training dataset. A high testing precision suggests the model maintains this accuracy with new data, indicating good generalisation.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4)$$

4.4. F1-Score

F1 score is a metric used to evaluate the performance of a classification model, especially when dealing with imbalanced datasets. It is the harmonic mean of precision and recall, providing a single metric that balances both concerns (Narasimha Rao et al., 2023). The F1 score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst performance. F1 Score for training data refers to the F1 score calculated based on the predictions the model makes on the training dataset. This score indicates how well the model performs on the data it was trained on in terms of precision and

recall. F1 Score for testing data, also known as validation F1 score, refers to the F1 score calculated based on the predictions the model makes on the testing dataset. This score indicates how well the model is expected to perform on unseen data in terms of balancing precision and recall. A high training F1 score indicates the model performs well on the training data, capturing the relationships between features and labels effectively and a high testing F1 score indicates the model generalises well to new data, maintaining a good balance between precision and recall on unseen instances.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Table 5: Performance metrics of the various models with trained data set and testing data set

Model/Classifier	Training				Testing			
	Accuracy	F1 Score	Recall	Precision	Accuracy	F1 Score	Recall	Precision
Dummy	0.57	0.73	1.00	0.57	0.60	0.75	1.00	0.60
Gaussian NB	0.56	0.46	0.32	0.79	0.57	0.52	0.39	0.79
Decision Tree	0.62	0.67	0.67	0.67	0.57	0.64	0.63	0.65
XGB	0.67	0.71	0.70	0.72	0.63	0.69	0.67	0.70
SGD	0.67	0.71	0.72	0.71	0.60	0.69	0.75	0.66
Extra Trees	0.70	0.75	0.79	0.71	0.63	0.71	0.74	0.68
Gradient Boosting	0.68	0.74	0.78	0.70	0.63	0.72	0.77	0.67
KNN	0.68	0.73	0.75	0.71	0.65	0.73	0.77	0.69
RF	0.69	0.74	0.79	0.70	0.65	0.74	0.81	0.68
Tuned SVC	0.72	0.77	0.81	0.73	0.69	0.78	0.88	0.70
Linear SVC	0.71	0.76	0.79	0.72	0.68	0.75	0.79	0.71
Tuned Linear SVC	0.71	0.76	0.79	0.72	0.67	0.74	0.79	0.71
LR	0.71	0.76	0.79	0.73	0.69	0.76	0.81	0.72
SVC	**0.72**	**0.77**	**0.81**	**0.73**	**0.69**	**0.78**	**0.88**	**0.70**
Tuned LR	0.71	0.76	0.79	0.73	0.69	0.76	0.81	0.72
Tuned RF	0.69	0.75	0.81	0.70	0.65	0.74	0.83	0.67
Tuned XGB	0.68	0.72	0.73	0.71	0.65	0.72	0.73	0.71

Hence it can be concluded from the above table that based on the performance metrics, Support Vector Machine based classifier provides the best performance during testing offering a recall of 88%, a Precision of 70% and an F1 Score of 78%. Hence it exhibits its capability to be the chosen classifier compared to the other classifiers mentioned.

5. CONCLUSION

This study demonstrates the efficacy of Support Vector Classifier (SVC) in rainfall prediction for the selected geographical region. Through meticulous feature selection, model optimization, and rigorous performan ce validation, the SVC model exhibits remarkable accuracy in forecasting precipitation events. The model's success is predicated on the judicious selection of input features, encompassing historical meteorological data, barometric pressure readings, and seasonal variability. The research underscores the critical importance of high-resolution, comprehensive datasets and sophisticated data preprocessing techniques in ensuring robust model training. Our findings reveal that a well-calibrated SVC model, validated through appropriate cross-validation techniques, demonstrates exceptional discriminatory power in differentiating between precipitation and non-precipitation days, as corroborated by a suite of performance metrics.

6. FUTURE DIRECTIONS

While the current SVC model yields promising results, there remain avenues for further enhancement and investigation. Future research endeavours should prioritize comparative analyses with alternative machine learning paradigms to elucidate the relative strengths and limitations of the SVC approach. Exploration of ensemble methodologies and hybrid modelling techniques may yield incremental improvements in predictive accuracy. Furthermore, the integration of real-time data streams and the development of adaptive algorithms capable of responding to evolving climatic patterns represent promising areas for advancement. Addressing the computational challenges associated with large-scale datasets and improving the model's scalability will be crucial for operational deployment. Lastly, ongoing refinement and recalibration of the model in response to long-term climate trends will be essential to maintain its predictive power and relevance in the face of global environmental changes. These future directions hold the potential to further elevate the SVC model's utility in meteorological forecasting and decision-support systems.

7. ACKNOWLEDGEMENT

The authors would like to thank Innovate UK, Knowledge Transfer Partnership number 13246, for partially funding this work.

8. REFERENCES

Angra S. and Ahuja S., (2017) "Machine learning and its applications: A review," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, Andhra Pradesh, India, 2017, pp. 57-60, doi: 10.1109/ICBDACI.2017.8070809

Balakrishnan, Sivakumar & Nanjundaswamy, Chikkamadaiah. (2021). Weather monitoring and forecasting system using IoT. Global Journal of Engineering and Technology Advances. 8. 008-016. 10.30574/gjeta.2021.8.2.0109.

Banara S., Singh T. and Chauhan A., (2022) "IoT Based Weather Monitoring System for Smart Cities: A Comprehensive Review," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1 -6, doi: 10.1109/ICONAT53423.2022.9726106

Bansal A. and Singhrova A., (2021) "Performance Analysis of Supervised Machine Learning Algorithms for Diabetes and Breast Cancer Dataset," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 137-143, doi: 10.1109/ICAIS50930.2021.9396043.

Bishop, Christopher M, (2006) Pattern recognition and machine learning, Springer, NY, USA

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2:2, 121–167.

Feras A. Batarseh & Ruixin Yang (Eds), Data Democracy, 2020, Copyright © 2020 Elsevier Inc, DOI: https://doi.org/10.1016/C2018-0-04003-7, ISBN: 978-0-12-818366-3

Garg Arnav and Pandey Himanshu, Rainfall Prediction using Machine Learning, International Journal of Innovative Science and Research Technology, 2019, 4(5), 56-58

Genuer R., Poggi J. M., Tuleau-Malot C., 2010. Variable selection using random forests. Pattern Recognition Letters. 31.: 2225–2236.

Javier Diez-Sierra, Manuel del Jesus, Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods, Journal of Hydrology, Volume 586, 2020, 124789, ISSN 0022 - 1694, https://doi.org/10.1016/j.jhydrol.2020.124789.

Kaushik Sunil, Bhardwaj Akashdeep, Sapra Luxmi, “Predicting Annual Rainfall for the Indian State of Punjab Using Machine Learning Techniques”,2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) ,978-1-7281-8337-4/20/$31.00 ©2020 IEEE

Kavin R., Lakshmi K., Rani S. S. and Rameshkumar K., "Weather Monitoring System using Internet of Things," 2020, 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 26-29, doi: 10.1109/ICACCS48705.2020.9074332.

Keerthi, S., Shevade, C. B. S. K., & Murthy, K. R. K. (2000). A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Trans. Neural Networks, 11(1), 124–136.

Kowar Manoj Kumar, Shrivastava Gyanesh, Karmakar Sanjeev, Guhathakurta Pulak , 2012, “Application of artificial neural networks in weather forecasting: A comprehensive literature Review”, International Journal of Computer Applications (0975

- 8887), Volume 51– No.18.

Kumar Rajesh, Decision Tree for the Weather Forecasting, International Journal of Computer Applications 2013, 76(2):31-34, DOI:10.5120/13220-0620

Langley P., Iba W. and Thompson K., “An analysis of Bayesian Classifiers”, in Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, 1992

Liyew, C.M., Melese, H.A. Machine learning techniques to predict daily rainfall amount. Journal of Big Data 8, 153 (2021). https://doi.org/10.1186/s40537-021-00545-4.

Maalouf Maher, Logistic Regression in Data Analysis: An overview, International Journal of Data Analysis Techniques and Strategies, 2011, 3(3):281-299, DOI:10.1504/IJDATS.2011.041335

McIlveen, J. F. R. (J. F. Robin), 1992, Fundamentals of weather and climate, Pub: Chapman & Hall; New York, NY, USA: Van Nostrand Reinhold, Inc.

Medar R., Rajpurohit V. S. and Rashmi B., "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-6, doi: 10.1109/ICCUBEA.2017.8463779.

Narasimharao M., Swain B., Nayak P. P. and Bhuyan S., "Developing and Evaluating a Machine Learning Based Diagnosis System for Diabetes Mellitus using Interpretable Techniques," 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), Bhubaneswar, India, 2023, pp. 505-511, doi: 10.1109/APSIT58554.2023.10201753.

Paquin, A.L., Chaib-draa, B., Giguère, P., Stability analysis of stochastic gradient descent for homogeneous neural networks and linear classifiers, Neural Networks, Volume 164, 2023, Pages 382-394, ISSN 0893-6080,

https://doi.org/10.1016/j.neunet.2023.04.028.

Parmar, A.; Mistree, K.; Sompura, M. Machine learning techniques for rainfall prediction: A review. In Proceedings of the International Conference on Innovations in information Embedded and Communication Systems, Coimbatore, India, 17 – 18 March 2017

Praba M. S. Bennet, Martin Antony John, Srivastava Siddharth, Rana Ajay , Weather Monitoring System and Rainfall Prediction Using SVM Algorithm International Journal of Research in Engineering, Science and Management, 1 (10),2018, pp. 745-750

Preetum Nakkiran et al. SGD on Neural Networks Learns Functions of Increasing Complexity (2019), 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Raihan, M.J., Khan, M.AM., Kee, SH. et al. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. Sci Rep 13, 6263 (2023). https://doi.org/10.1038/s41598 -023-33525-0

Rok Blagus, Lara Lusa, Gradient boosting for high-dimensional prediction of rare events, Computational Statistics & Data Analysis, 2017, Vol. 113, pp.: 19-37, https://doi.org/10.1016/j.csda.2016.07.016.

Sadhukhan Mrinmoy, Dasgupta Sudakshina, Bhattacharya Indrajit, An Intelligent Weather Prediction System Based On IoT, 2021 Devices for Integrated Circuit (DevIC), 19-20 May, 2021, Kalyani, India

Sankarnarayan S, Prabhakar M, Satish S and Jain P , A Ramprasad and A Krishnan, Flood Prediction based on Weather Parameter using Deep Learning, Journal of Water & Climate Change, 2020, 11(4), pp. 1766 – 1783

Subia Salma, Ashwitha A, Hybrid CNN-LSTM Model: Rainfall Analysis and Prediction for Karnataka Region, Journal of Theoretical and Applied Information Technology, 2022. Vol.100. No 22: 6715-6727

Tang Y, Durand D. A tunable support vector machine assembly classifier for epileptic seizure detection. Expert Syst Appl. 2012 Mar 1;39(4):3925-3938. doi: 10.1016/j.eswa.2011.08.088. Epub 2011 Aug 30. PMID: 22563146; PMCID: PMC3341176.

Taunk K., De S., Verma S. and Swetapadma A., "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

Tharun VP, Prakash R, Devi SR, Prediction of Rainfall using Data Mining Techniques, in 2018 second international conference on inventive communication and Computational Techniques (ICCICCT), IEEE Explore, 2018, pp.1507-1512

TINA E. M., An Optimized Extremely Randomized Tree Model for Breast Cancer Classification, Journal of Theoretical and Applied Information Technology, 2022, Vol. 100(16), pp. 5234-5246

Verma Gaurav, Mittal Pranjul and Farheen Shaista, Real Time Weather Prediction System Using IOT and Machine Learning, Conference: 2020 6th International Conference on Signal Processing and Communication (ICSC), DOI:10.1109/ICSC48311.2020.9182766

Waqas, M.; Humphries, U.W.; Wangwongchai, A.; Dechpichai, P.; Ahmad, S. Potential of Artificial Intelligence -Based Techniques for Rainfall Forecasting in Thailand: A Comprehensive Review. Water 2023, 15, 2979. https://doi.org/10.3390/w15162979

Zhou, J., Shi J., Li G., Fine tuning support vector machines for short-term wind speed forecasting, Energy Conversion and Management, 2011, Vol. 2(4), PP.: 1990-1998, https://doi.org/10.1016/j.enconman.2010.11.007.