LSTM and XGBoost Ensemble Model: An Approach for assessment and monitoring of Weather Prediction

Shimaila[1], Dr. Sifatullah Siddiqi[1]

[1]Department of CSE, Integral University, Lucknow, India.
shimailaphd@gmail.com, sifatullah.siddiqi@gmail.com

Cite this paper as: Shimaila, LSTM and XGBoost Ensemble Model: An Approach for assessment and monitoring of Weather Prediction. *Frontiers in Health Informatics*, 13 (4), 979-986

Abstract
Managing the effects of climate change and extreme weather events that affect public safety, transportation, and agriculture etc requires accurate weather forecasting. This study offers a novel method for enhancing weather forecasting by utilizing Principal Component Analysis (PCA) for dimensionality reduction, Recursive Feature Elimination (RFE) for feature selection, Long Short-Term Memory (LSTM) networks for sequence modeling, and eXtreme Gradient Boosting (XGBoost) for predictive modeling. Comprehensive data collection, thorough preprocessing, feature selection, and the development of an ensemble model using LSTM and XGBoost are all part of the suggested methodology. This approach improves the ability to identify intricate, nonlinear relationships in meteorological data, leading to more accurate and reliable predictions. The model was evaluated using several metrics, achieving a high accuracy of 95.9% and an AUC of 0.83, demonstrating its effectiveness.

Keywords: Weather Forecasting, Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), eXtreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM).

1. Introduction

Weather forecasting is a vital tool for human life and societal activities since it can solve a variety of issues in a number of industries, such as agriculture, transportation, construction, and disaster assistance [1]. It is essential in many areas and helps with planning, preparation, and decision-making [2]. The socioeconomic impact of weather forecasting and climate change cannot be underestimated [3][4]. Weather forecasting has been transformed by Numerical Weather Prediction (NWP) models [5]. These physics-based models incorporate observations from a variety of sources, including satellite data, using data assimilation techniques [6][7]. However, data thinning is necessary due to the high density of satellite observations. Higher geographic resolution is generally thought to produce more accurate weather forecasts in NWP models, particularly when it comes to rainfall prediction [8–12]. Nevertheless, convection-permitting spatial resolutions in global NWP models are not achievable with conventional computing techniques. Consequently, research looked into alternative computer systems that make use of Graphical Processing Units (GPUs) [13].

Weather forecasting frequently uses a number of AI algorithms. Based on patterns and trends, supervised machine learning methods such as Support Vector Machine (SVM), Random Forest (RF), and neural networks may evaluate historical weather data and predict current and future conditions. Because it can handle data with several dimensions and nonlinear relationships, RF has been used in weather prediction [14][15]. However, it has difficulties when predicting beyond the observed data. SVM is used to categorize meteorological conditions and forecast rainfall. It exhibits favorable results with data that cannot be separated linearly, despite being computationally costly and requiring careful hyper-parameter tuning for optimal performance [16][17]. Deep Learning (DL) [18] is becoming popular as a successful neural network development. Deep learning is a data-driven technique that reveals complex correlations between input and output data by using reference input sets and comparable labelled output data. It can potentially surpass conventionally constructed standard models [19]. The Earth system research domain, which includes satellite remote sensing and weather forecasting, seems to be a good fit for using DL techniques [20]. The growing quantity of Earth Systems data sets from many sources, such as crowdsourcing sensors and EO satellites [21][22], offers DL algorithms the chance to gain important insights and enhance weather forecasting models [23][24]. Neural networks, such as deep learning models, have gained a lot of popularity due to their capacity to recognize intricate patterns and make precise predictions. In weather forecasting, they have been helpful in predicting temperature, precipitation, and humidity. Neural network training requires a lot of labeled data, which can be computationally expensive. This study suggests a novel strategy for utilizing these cutting-edge DL techniques to increase weather forecasting efficiency in order to overcome these difficulties.

The proposed methodology enhances weather forecasting by leveraging advanced machine learning techniques, thereby providing more reliable and precise weather predictions. This research addresses the growing demand for improved forecasting accuracy, which is essential for planning and decision-making processes across various industries and communities.

This study introduces several novel contributions to the field of weather forecasting:

• Utilizes the advantages of both time-series and gradient-boosting techniques, LSTM and XGBoost models in an ensemble framework to improve overall predictive accuracy.

•The proposed model captures essential patterns in the data while reducing computational complexity by employing PCA for dimensionality reduction and advanced feature selection methods like recursive feature elimination.

•The proposed system can dynamically adapt to new patterns and improve its predictive performance, making weather forecasts more reliable and timelier by incorporating mechanisms for continuous learning and real-time feedback.

2. Research Methodology

Together with the suggested framework, this section describes the methods and procedures used in the study. The following are the particular approaches and strategies used:

(i) Principal Component Analysis (PCA)

PCA is a method of extracting features from data sets without the need for supervision. It transforms the data into a lower-dimensional representation while minimizing the error as much as possible [25]. The PCA technique is employed in this study for weather forecasting due to its ability to offer a statistically robust technique for simplifying the data and producing an entirely novel set of variables known as principal components. By employing this approach, every produced element is directly linked to the initial variables in a proportionate way. PCA offers the benefit of having orthogonal principal components for each data set, ensuring that there remains no duplicate information after preprocessing [26]. Using the following equation, PCA can be usefully represented:

$$X_{new} = X.W \qquad (1)$$

Where:

- X_new is the transformed feature set.
- X is the original feature set.
- W represents the transformation matrix obtained from PCA.

(ii) Recursive Feature Elimination (RFE)

Irrelevant features are common in large datasets. Recurring features impact the inefficiency of the classification algorithm. This could lead to less accurate predictions. Determine which variables are crucial for making accurate forecasts using RFE. This preserves the valuable characteristics acquired from the more refined feature sets while reducing the dataset's dimensionality. A method that iteratively searches for a target number of attributes is "recursive". RF classifiers rank attributes in descending order of priority after establishing their significance. Next, the model is retrained with the updated feature set to improve classification accuracy and remove less important features. The loop continues as long as there are additional features to be included [27].

(iii) Long Short-Term Memory (LSTM)

LSTM is a specific kind of Recurrent Neural Network (RNN) that is explicitly built for modeling and forecasting time series data. Conventional RNNs encounter difficulties in capturing long-term dependencies, mostly due to problems such as gradient vanishing. In order to overcome this difficulty, LSTM introduced gate operations that regulate data flow. Because LSTMs can learn from past weather patterns and use this information to predict future circumstances, they are especially useful in weather forecasting. [28].

(iv) eXtreme Gradient Boosting (XGBoost)

XGBoost, a sophisticated machine learning algorithm, has gained prominence for its performance and efficiency in predictive modeling. Meteorologists can use XGBoost's advantages to create more precise short- and long-term forecasts by integrating it into a weather forecasting system. [29]. It can be particularly useful in ensemble models, where its predictions are combined with those from other models, such as LSTM neural networks, to improve overall forecast accuracy and robustness.

Figure 1 illustrates the suggested architecture for enhancing weather forecasting effectiveness. By incorporating intelligent tools into the process, the suggested methodology seeks to improve the accuracy of weather forecasting. It begins with comprehensive data collection, encompassing both historical weather data and real-time meteorological inputs from sensors and IoT devices. This data undergoes rigorous processing to ensure quality and consistency, followed by feature selection and extraction to identify the most relevant information. LSTM and XGBoost, are then developed and trained on this data. The methodology attempts to provide extremely precise predictions by integrating their strengths using an ensemble approach. Models are evaluated using various metrics and deployed for real-time forecasting, with continuous monitoring and feedback loops in place to ensure ongoing improvement and adaptation to new data.

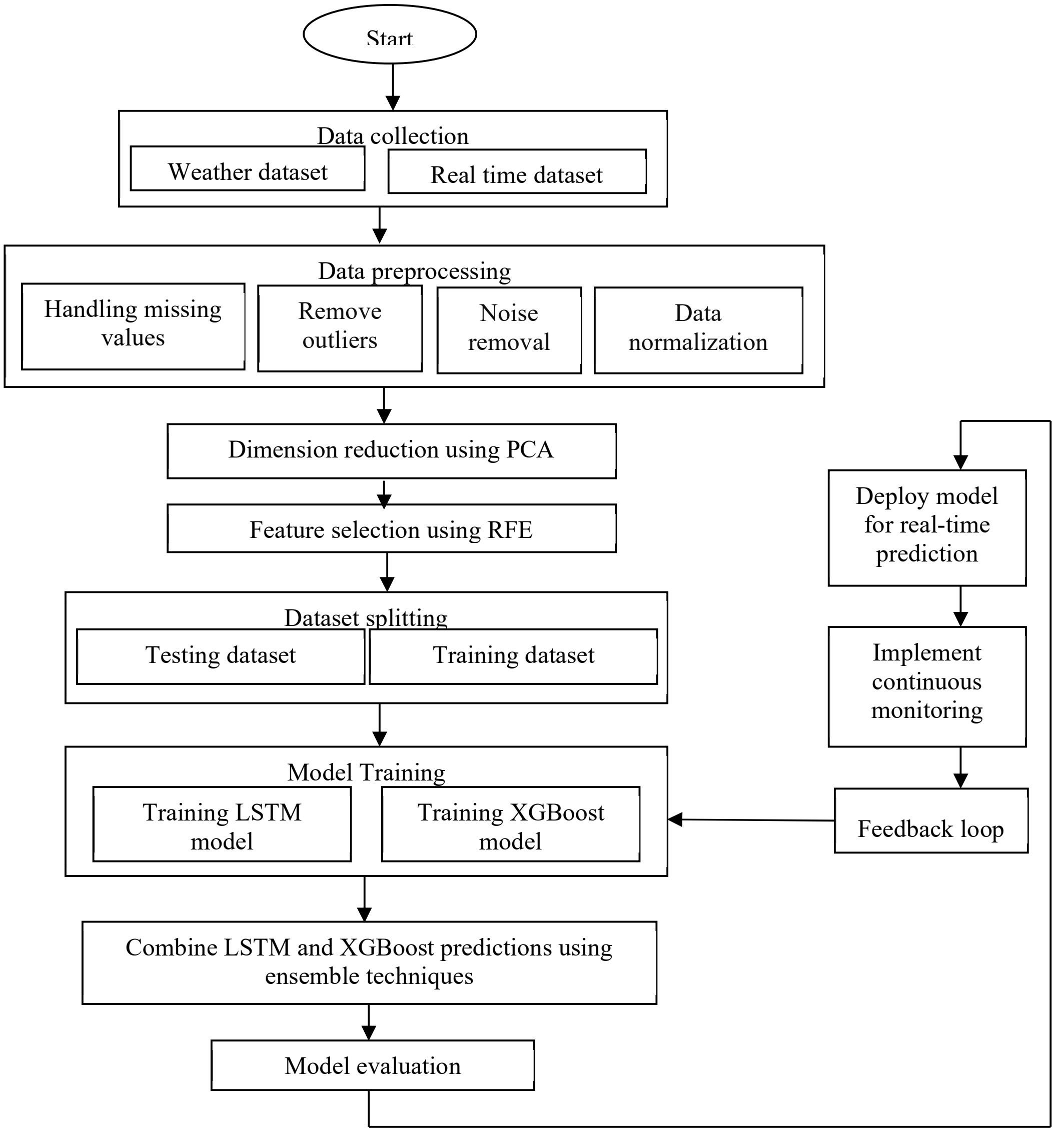

Figure 1. Proposed methodology

3. Results and Discussion

The dataset description and result analysis are discussed in this section:

(i) Dataset Description

The dataset is a comprehensive weather dataset [30], containing 142,193 entries and 24 columns, detailing various meteorological observations across multiple locations. The variables are recorded across different dates and locations, allowing for temporal and spatial analysis of weather patterns. The dataset is well-suited for various types of analyses, including predictive modeling, trend analysis, and classification tasks related to weather prediction.

(ii) Result analysis

This section presents a detailed analysis of the results obtained by the proposed algorithms:

- **Based on Confusion matrix for Ensemble model**

The confusion matrix in Figure 2 illustrates the evaluation of an ensemble model's performance in predicting two classes, denoted as "0" and "1". The matrix shows that out of 33,133 actual instances of Class 0, the model correctly predicted 31,151 cases but incorrectly classified 1,982 as Class 1. For Class 1, out of 9,525 actual instances, the model correctly identified 3,153 cases, while 6,372 instances were misclassified as Class 0. The confusion matrix offers valuable information regarding the model's classification accuracy by identifying both accurate and inaccurate predictions for each class.

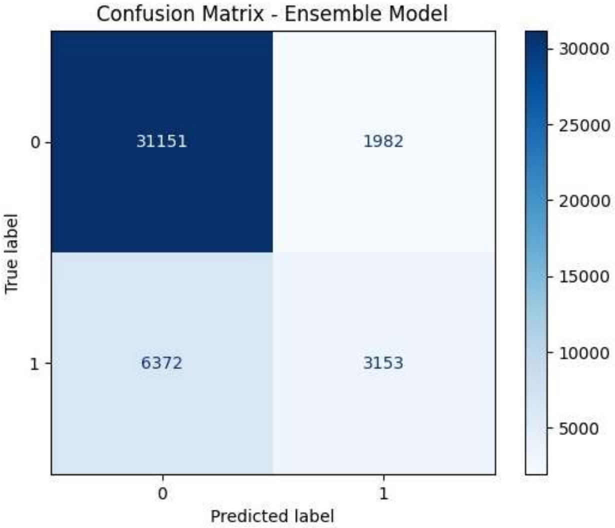

Figure 2. Confusion matrix for Ensemble model

- **Based on Accuracy**

Figure 3 shows a line graph depicting the model accuracy over 50 epochs, comparing the training accuracy and validation accuracy. The x-axis and y-axis represents number of epochs and accuracy respectively. The training accuracy, indicated by the blue line, shows a gradual increase, starting from around 93.7% and reaching approximately 95.8% by the end of the 50 epochs. The validation accuracy, represented by the orange line, also increases over time and generally remains slightly higher than the training accuracy, with some fluctuations, eventually reaching around 95.9%. The graph indicates that the model is improving its performance with each epoch, with both training and validation accuracies converging and showing consistent improvement.

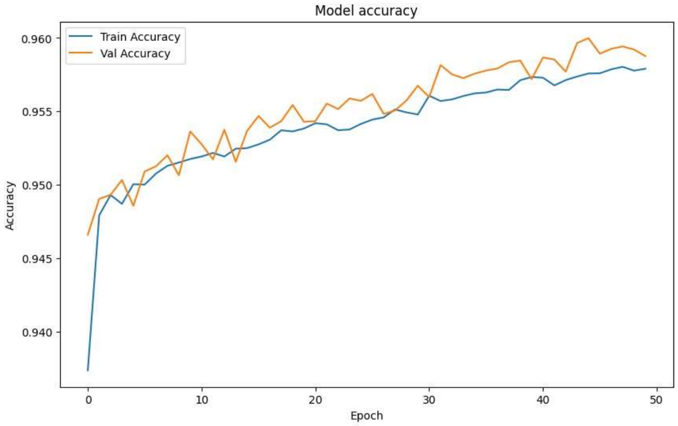

Figure 3. Accuracy

• Based on loss

Figure 4 displays a line graph representing the model's loss over 50 epochs, comparing the training loss and validation loss. The x-axis and y-axis represents the number of epochs and loss value respectively. The training loss, shown by the blue line, starts at around 0.415 and decreases steadily throughout the training process, reaching approximately 0.390 by the end of the 50 epochs. The validation loss, indicated by the orange line, follows a similar downward trend, starting slightly below the training loss at about 0.405 and also decreasing to about 0.390 by epochs end. The plot suggests that the model is effectively learning and reducing error over time, with both the training and validation loss values converging and showing consistent improvement.

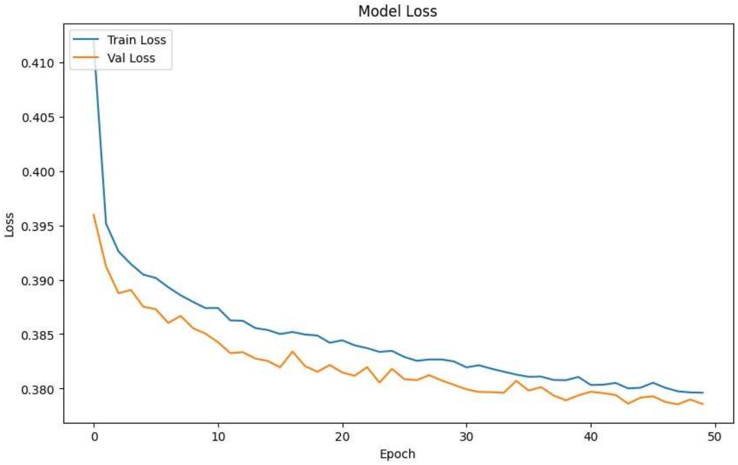

Figure 4. Model Loss Over Epochs

• Based on ROC Curve

A ROC curve, which is used to evaluate a binary classification model's effectiveness, is shown in Figure 5. The True Positive Rate is shown on the y-axis, and the False Positive Rate is shown on the x-axis. The trade-off between sensitivity (the capacity to reliably identify positive cases) and specificity (the capacity to accurately identify negative cases) at different threshold levels is shown by the ROC curve, which is depicted in orange. A random classifier's performance is shown by the diagonal blue dashed line, which has an area under the curve (AUC) value of 0.5. With a good balance between true positive and false positive rates, the model's AUC of 0.83 shows that it performs noticeably better than random guessing.

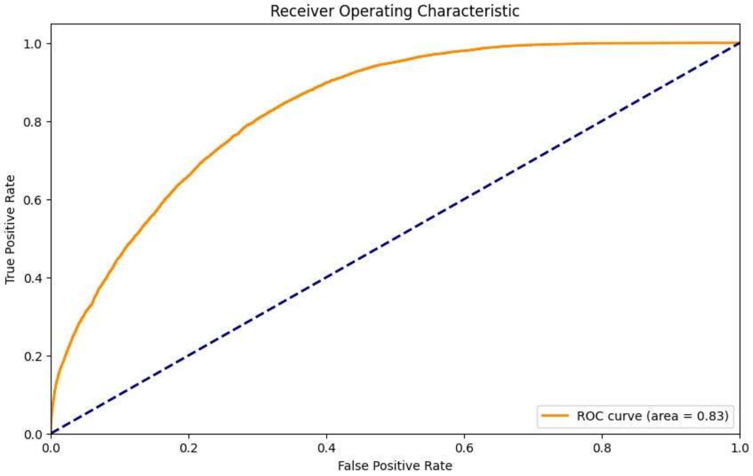

Figure 5. ROC Curve Analysis

4. Conclusion

The science of anticipating atmospheric conditions through data analysis is known as weather forecasting, and it is essential for a number of sectors, including disaster relief, aviation, and agriculture. The increasing demand for more precise weather forecasts to lessen the effects of extreme weather events is addressed in this study. Using cutting-edge machine learning models to improve forecast accuracy is one of the goals. The assessment models employed in this study are PCA, RFE, LSTM, and XGBoost. The proposed methodology integrates these models, leveraging an ensemble approach to combine their strengths. The model achieved an accuracy of 95.9% with an AUC of 0.83, indicating robust performance. The proposed model contributes significantly to the field of weather forecasting by offering a more reliable and precise prediction framework, which is vital for timely and informed decision-making.

Acknowledgements:

To my supervisor, Dr. Sifatullah Siddiqi, I would like to thank you for your guidance and support. Integral University's Research & Development department in Lucknow has assigned the paper communication number IU/R&D/2024-MCN0003239

References

1. Merz, Bruno, Christian Kuhlicke, Michael Kunz, Massimiliano Pittore, Andrey Babeyko, David N. Bresch, Daniela IV Domeisen et al. "Impact forecasting to support emergency management of natural hazards." Reviews of Geophysics 58, no. 4 (2020): e2020RG000704.
2. Merz, Bruno, Christian Kuhlicke, Michael Kunz, Massimiliano Pittore, Andrey Babeyko, David N. Bresch, Daniela IV Domeisen et al. "Impact forecasting to support emergency management of natural hazards." Reviews of Geophysics 58, no. 4 (2020): e2020RG000704.Kishore MT, Satyanarayana V, Ananthanpillai ST, Desai G, Bhaskarapillai B, Thippeswamy H, et al. Life events and depressive symptoms among pregnant women in India: Moderating role of resilience and social support. *International Journal of Social Psychiatry.* 2018; **64**:570-577
3. Hochman, Assaf, Francesco Marra, Gabriele Messori, Joaquim G. Pinto, Shira Raveh-Rubin, Yizhak Yosef, and Georgios Zittis. "Extreme weather and societal impacts in the eastern Mediterranean." Earth System Dynamics 13, no. 2 (2022): 749-777.
4. Edenhofer, Ottmar, Ramon Pichs-Madruga, Youba Sokona, S. Agrawala, I. A. Bashmakov, G. Blanco, J. Broome et al. "Summary for policymakers." (2014)Bauer, Peter, Alan Thorpe, and Gilbert Brunet. "The quiet revolution of numerical weather prediction." Nature 525, no. 7567 (2015): 47-55.
5. Fekadu Dadi A, Miller ER, Mwanri L. Antenatal depression and its association with adverse birth outcomes in low and middle-income countries: A systematic review and meta-analysis. *PLoS One*. 2020; **15**:e0227323.

6. Duan, W. S., Rong Feng, L. C. Yang, and Lin Jiang. "A new approach to data assimilation for numerical weather forecasting and climate prediction." Journal of Applied Analysis and Computation 12, no. 3 (2022): 1007-1021.
7. Lean, P., E. V. Hólm, M. Bonavita, N. Bormann, A. P. McNally, and Heikki Järvinen. "Continuous data assimilation for global numerical weather prediction." Quarterly Journal of the Royal Meteorological Society 147, no. 734 (2021): 273-288.
8. Clark, Peter, Nigel Roberts, Humphrey Lean, Susan P. Ballard, and Cristina Charlton-Perez. "Convection-permitting models: a step-change in rainfall forecasting." Meteorological Applications 23, no. 2 (2016): 165-181.
9. Capecchi, Valerio, Andrea Antonini, Riccardo Benedetti, Luca Fibbi, Samantha Melani, Luca Rovai, Antonio Ricchi, and Diego Cerrai. "Assimilating X-and S-band Radar Data for a Heavy Precipitation Event in Italy." Water 13, no. 13 (2021): 1727.
10. Maiello, Ida, Sabrina Gentile, Rossella Ferretti, Luca Baldini, Nicoletta Roberto, Errico Picciotti, Pier Paolo Alberoni, and Frank Silvio Marzano. "Impact of multiple radar reflectivity data assimilation on the numerical simulation of a flash flood event during the HyMeX campaign." Hydrology and Earth System Sciences 21, no. 11 (2017): 5459-5476.
11. Ricchi, Antonio, Davide Bonaldo, Guido Cioni, Sandro Carniel, and Mario Marcello Miglietta. "Simulation of a flash-flood event over the Adriatic Sea with a high-resolution atmosphere–ocean–wave coupled system." Scientific Reports 11, no. 1 (2021): 9388.
12. Zheng, Yue, Kiran Alapaty, Jerold A. Herwehe, Anthony D. Del Genio, and Dev Niyogi. "Improving high-resolution weather forecasts using the Weather Research and Forecasting (WRF) Model with an updated Kain–Fritsch scheme." Monthly Weather Review 144, no. 3 (2016): 833-860.
13. Bauer, Peter, Peter D. Dueben, Torsten Hoefler, Tiago Quintino, Thomas C. Schulthess, and Nils P. Wedi. "The digital revolution of Earth-system science." Nature Computational Science 1, no. 2 (2021): 104-113.
14. Fente, Dires Negash, and Dheeraj Kumar Singh. "Weather forecasting using artificial neural network." In 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1757-1761. IEEE, 2018.
15. Mishra, Shivam, Aakash Shukla, Sandeep Arora, Himandhu Kathuria, and Mandeep Singh. "Controlling Weather Dependent Tasks Using Random Forest Algorithm." In 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1-8. IEEE, 2020.
16. Suksri, Saktaya, and Warangkhana Kimpan. "Neural network training model for weather forecasting using fireworks algorithm." In 2016 International Computer Science and Engineering Conference (ICSEC), pp. 1-7. IEEE, 2016.
17. Higgins JP and Green S: Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons, Ltd, Chichester, UK, 2008.
18. Higgins JP and Green S: Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons, Ltd, Chichester, UK, 2008.
19. Anaka, Rukevwe Emmanuel. "Review of AI-Enhanced Weather Forecasting Application for Communication Networks."
20. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521, no. 7553 (2015): 436-444.
21. Shrestha, Ajay, and Ausif Mahmood. "Review of deep learning algorithms and architectures." IEEE access 7 (2019): 53040-53065.
22. Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and fnm Prabhat. "Deep learning and process understanding for data-driven Earth system science." Nature 566, no. 7743 (2019): 195-204.
23. Guo, Huadong, Wenxue Fu, Guang Liu, Huadong Guo, Wenxue Fu, and Guang Liu. "European Earth observation satellites." Scientific Satellite and Moon-Based Earth Observation for Global Change (2019): 97-135.
24. Zheng, Feifei, Ruoling Tao, Holger R. Maier, Linda See, Dragan Savic, Tuqiao Zhang, Qiuwen Chen et al. "Crowdsourcing methods for data collection in geophysics: State of the art, issues, and future directions." Reviews of Geophysics 56, no. 4 (2018): 698-740.
25. Boukabara, Sid-Ahmed, Vladimir Krasnopolsky, Jebb Q. Stewart, Eric S. Maddy, Narges Shahroudi, and Ross N. Hoffman. "Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges." Bulletin of the American Meteorological Society 100, no. 12 (2019): ES473-ES491.

26. Dewitte, Steven, Jan P. Cornelis, Richard Müller, and Adrian Munteanu. "Artificial intelligence revolutionizes weather forecast, climate monitoring and decadal prediction." Remote Sensing 13, no. 16 (2021): 3209.
27. S. Sen, S. K. Saha, S. Chaki, P. Saha, and P. Dutta, "Analysis of pca based adaboost machine learning model for predict midterm weather forecasting," Computational Intelligence and Machine Learning, 2021
28. El Mhouti, Abderrahim, Mohamed Fahim, Asmae Bahbah, Yassine El Borji, Adil Souf, Ayoub Aoulalay, and Chaimae Ouazri. "A Machine Learning-Based Approach for Meteorological Big Data Analysis to Improve Weather Forecast." International Journal of Computing and Digital Systems 14, no. 1 (2023): 1-xx.
29. Ponniah, Krishna Kumar, and Bharathi Retnaswamy. "A novel deep learning-based intrusion detection system for the IoT-Cloud platform with blockchain and data encryption mechanisms." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-18.
30. Salman, Afan Galih, Yaya Heryadi, Edi Abdurahman, and Wayan Suparta. "Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting." Procedia Computer Science 135 (2018): 89-98.
31. Singh, Deep Karan, and Nisha Rawat. "Machine learning for weather forecasting: XGBoost vs SVM vs random forest in predicting temperature for visakhapatnam." Int. J. Intell. Syst. Appl 15 (2023): 1-12.
32. https://www.kaggle.com/datasets/manidevesh/weather-data-set-australia