IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Distributed Decomposed Data Analytics with Weather Insights

S.Sankareswari
IT Department
Finolex Academy of Management and Technology, Ratnagiri, India

Yash Shankar Gosavi
IT Department
Finolex Academy of Management and Technology, Ratnagiri, India

Baseema Irfan Kazi
IT Department
Finolex Academy of Management and Technology, Ratnagiri, India

Suyash Sachin Kanase
IT Department
Finolex Academy of Management and Technology, Ratnagiri, India

Sanika Ashok Tambe
IT Department
Finolex Academy of Management and Technology, Ratnagiri, India

***Abstract*— Weather condition prediction is a crucial requirement for various fields, such as agriculture, which requires accurate weather conditions for crop growth. The purpose of our project is to enhance weather prediction and provide real-time data on weather parameters. It includes integrating data from nodes, such as sensors like ESP280, and fetching it to the cloud, followed by applying ML operations on the cloud. In this project, we have used AWS Cloud to analyze and process data. This project consists of the KNN algorithm to predict the outcomes. Additionally, this project successfully delivers the aspect of a distributed decomposed data analytics technique.**

I. Introduction

Weather prediction is a crucial aspect in various fields like the energy sector, transportation, agriculture, etc. Technologies like IoT, Machine Learning, and Cloud Computing are very useful for integrating real-time weather data and analyzing it.

This project consists of a Distributed Decomposed Data Analytics (D3A) mechanism where data is in a decentralized form. The data collection, processing, analyzing, etc., are all done on different nodes in the system. The main aim of the project is to efficiently implement the D3A mechanism and apply cloud computing methods to process data.

In this project, AWS Cloud is used for its diverse tools, such as Lambda functions and S3 buckets. This project contributes to fields like agriculture, where weather data is highly needed. By using various sensors, it achieves accurate readings, and processing on the cloud ensures operations do not rely on local machines. Users can access data or make changes from anywhere.This project can provide accurate weather conditions to users on dashboards on their phones. It can also increase the number of sensors used to improve accuracy. Additionally, it can be applied in various other fields like disaster management. This system is in its early stage; therefore, there are not enough sensors and computational power. However, it already has an accuracy rate of more than previous testing.

The major contributions of this paper include:

(1) The utilization of machine learning in prediction of weather conditions in short periods of time, which can run on less resource-intensive machines.

(2) Implementation of automated systems to collect real-time data from a dedicated sensors.

(3) Thorough evaluation of the proposed technique ,comparison of several machine learning models and cloud computing mechanism in the prediction of future weather conditions.

II. Literature Survey

According to AHM Jakaria, Md. Mosharaf Hossain, and Mohammad Ashiqur Rehman, traditional weather forecasting models typically operate on high-performance computing (HPC) environments, leveraging hundreds of nodes. However, such models demand substantial energy consumption. In response, the authors proposed a lightweight machine learning (ML) approach that utilizes historical weather data from multiple weather stations. Their system trains simplified ML models, incorporating data from neighboring cities along with Nashville’s meteorological records to enhance predictive accuracy.

The study primarily employs Random Forest Regression (RFR) as its predictive model. Based on Root Mean Squared Error (RMSE) calculations, the system achieves an estimated accuracy of 95%, with an RMSE value of approximately 3.0. However, the research is geographically constrained, as it exclusively focuses on Nashville, Tennessee, and its surrounding regions. Furthermore, the model is limited to predicting only temperature, neglecting other critical meteorological parameters such as humidity, atmospheric pressure, and wind speed. This narrow scope reduces its overall applicability in broader climate studies.

Another significant limitation is that the model solely relies on historical data, making it ineffective in capturing real-time atmospheric variations. Additionally, the study does not explicitly define the prediction horizon (e.g., hourly, daily, or weekly forecasts), thereby hindering its adaptability across diverse application domains.

Moreover, the paper lacks comprehensive details on data preprocessing and model training procedures, which are crucial for ensuring model reliability and accuracy. The absence of these specifics raises concerns regarding the reproducibility and robustness of the proposed methodology.

This project successfully implemented a distributed and decomposed data analytics system. We have created a circuit from sensors that collect data and store it in an Excel file. This process works in real-time, unlike previous research papers where historical data from the city and neighboring cities was used. In our current system, we are creating our own weather data by recording weather parameters like pressure, humidity, and temperature, and uploading them to the cloud.so we have created a decentralized system. The system includes a machine learning algorithm, KNN, which processes the data and helps predict outcomes with an accuracy of over 90%. The system operates on AWS Cloud, providing various tools like Lambda functions, S3 buckets, and IAM users, which make operation handling easier

III. Objective

To create a real-time weather prediction system that improves forecasting accuracy through distributed data processing.
A variety of IoT sensors (BME280) and multiple edge computing devices (Arduino, ESP8266) make efficient data collection and preprocessing easier before transmission.
Parallel data processing across multiple nodes considerably improves scalability and accelerates computations.
The application of multiple machine learning techniques, including K-Nearest Neighbors (KNN), serves to detect many weather anomalies along with predicting particular conditions.
An important optimization of cloud resource usage involves employing edge computing to effectively reduce latency and network congestion.
Improving decision-making by providing accurate and timely weather understandings to support agriculture, improve disaster management and increase ecological monitoring..

IV. Methodology

Data from the satellite or own networks and data collected on cloud on these Weather forecasting systems rely on. Their prediction can be further enhanced and made in real time by adding the real time data of IoT, using multiple sensors.

Climate change has worsened the situation by disrupting weather patterns across the globe. According to a World Resources Institute study, global agricultural productivity might be reduced by 17% by 2050 as fertile farmlands face ruin due to extreme climate events caused by climate change. Flooding, inundation, drought, cyclones are depleting the soil of its nurturing properties for agriculture.

In view of such uncertainties, the weather forecast and its dissemination to farmers assumes more importance. Weather forecasts in farming are the answer to these problems .It is like having intelligence that helps farmers make better agricultural decisions by assessing the weather's impact on crops. This intelligence helps make smarter decisions, such as optimizing irrigation, timing fertilization, and choosing the best days for fieldwork to enhance productivity.

The proposed system will be taking input from following resources:

1. IoT Sensors deployed at a location (Air temperature, pressure and humidity)
    - BME280
    - ESP8266
2. Historical Data

The proposed framework is shown in Fig.1. The analytics

on the gathered Large volumes of data will be done in three phases:

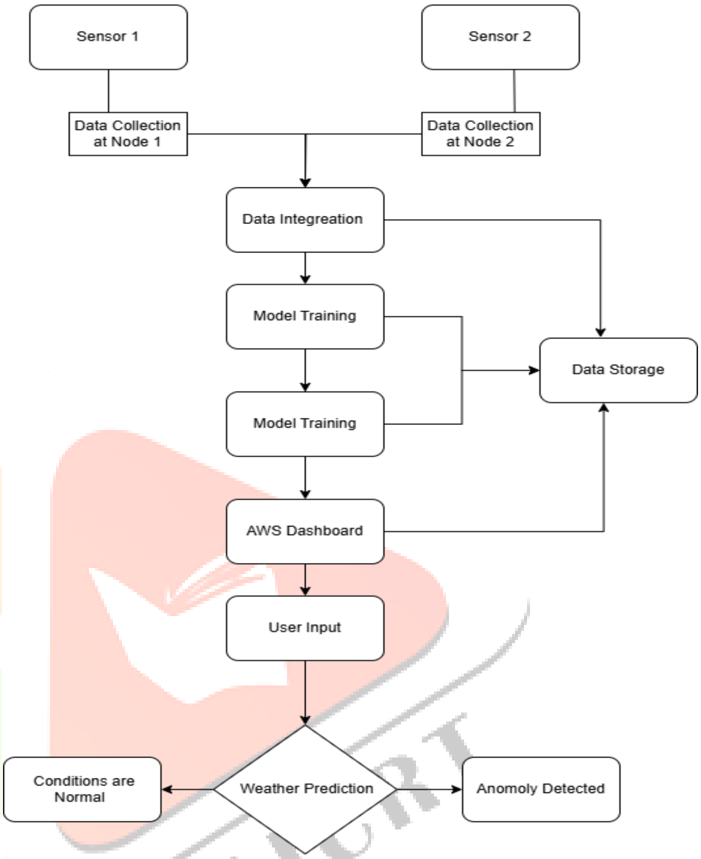


Fig 1. Working of Distributed Decomposed Data Analytics for predicting weather.

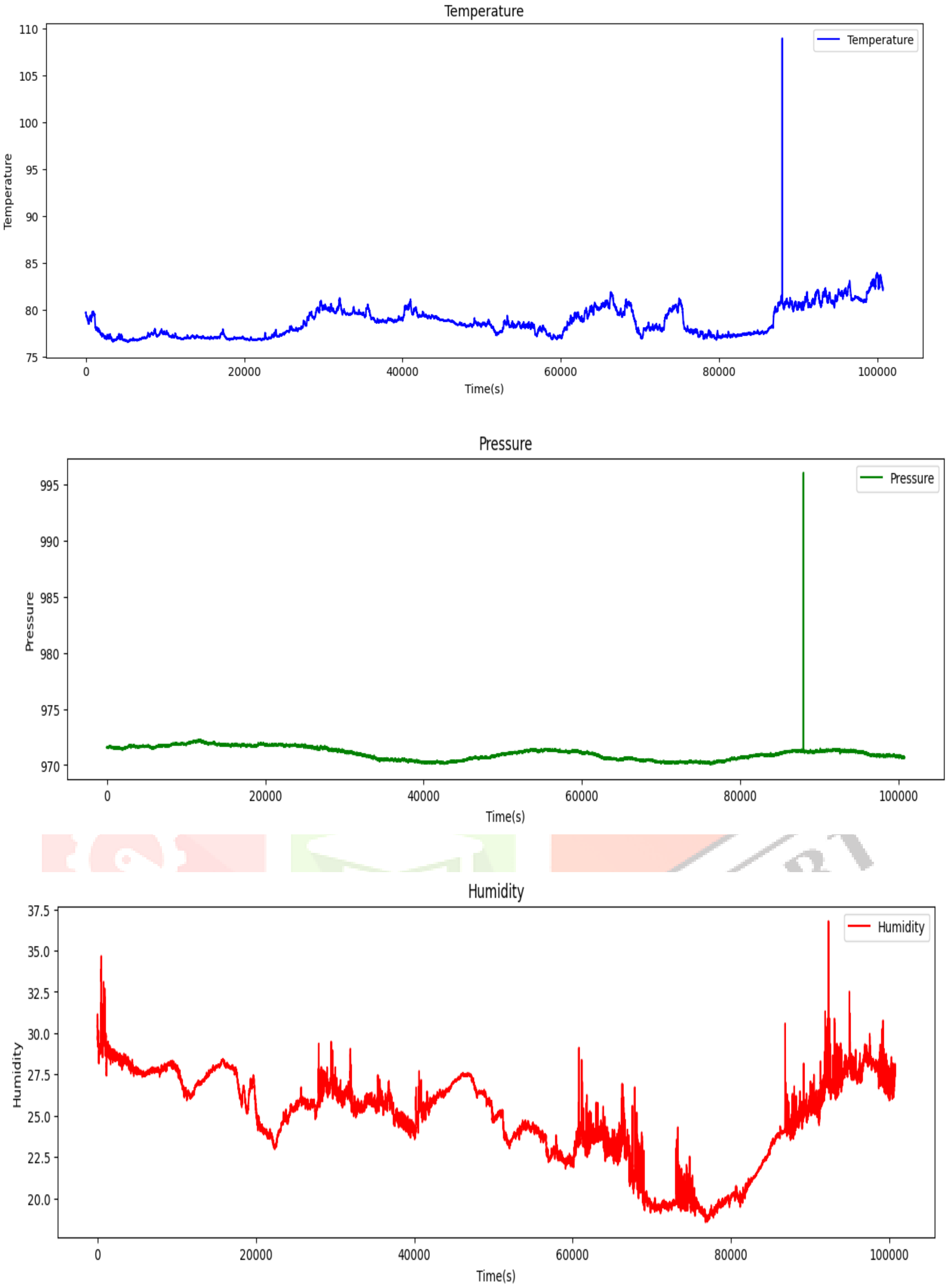


Fig 2. Temperature, humidity and Pressure Data Plots for BME280 Data used for the Prediction Analysis

Data Collection Phase

There are two BME280 sensors — one connected to an Arduino board and another to an ESP8266 NodeMCU. Each sensor collects environmental data such as temperature, humidity, and pressure. The collected data is sent to its respective local machine for initial processing.

Data Transmission:

The local machines continuously transmit the recorded data to an EC2 instance at a rate of one record every two seconds. This ensures that the system receives real-time data for further analysis.

Model Prediction:

On the EC2 instance, a Python script named practice.py processes the incoming data. This script uses three pre-trained machine learning models:

knnmodel.pkl for weather prediction,

scaler.pkl for data normalization, and

labelencoder.pkl for encoding categorical data.

Weather Prediction:

The practice.py script makes real-time predictions by classifying the weather into categories such as rainy, sunny, foggy, stormy, or cloudy based on the input data. This process runs continuously, providing ongoing and accurate weather insights.

A. *Data Procecesing on Cloud*

In our cloud-based system, sensor data is collected and initially processed at the edge before being uploaded to a distributed storage system. Data is gathered from multiple IoT sensors and temporarily processed using microcontrollers such as Arduino and ESP8266 to filter noise and reduce redundancy before transmission.

the raw data files from sensors are uploaded to an S3 bucket, where AWS Lambda triggers preprocessing, followed by data cleaning in Google Colab. The collected data is then stored in an AWS S3 bucket, serving as a distributed storage solution. Instead of utilizing AWS Lambda for integration and processing, the system now employs Google Cloud for preprocessing, which includes data transformation, cleaning, and feature engineering. This distributed approach ensures efficient data handling and scalable processing of large datasets across multiple computing nodes.

Model training is conducted using machine learning algorithms, where a pre-trained model (model.pkl) and historical data are utilized to enhance prediction accuracy. The trained model is deployed on an AWS EC2 instance, allowing real-time predictions based on new sensor inputs. This approach effectively distributes workload across cloud platforms, enhancing computational efficiency and reducing latency in inference.

B. *Learning phase*

The learning algorithm is implemented in Python using libraries such as Scikit-learn, Pandas, NumPy, and Dask. Scikit-learn is used to implement the KNN algorithm, while Pandas and NumPy handle data processing and numerical computations. The Dask library enables parallel computing across multiple CPUs or GPUs, ensuring efficient processing of large datasets. By leveraging multi-GPU environments, the system improves performance and scalability in data-intensive applications.

The proposed weather prediction system integrates IoT sensors, edge computing, and machine learning to provide accurate and real-time forecasts. Environmental parameters, including temperature, humidity, and pressure, are captured using BME280 sensors deployed across different locations. Initial data processing is performed at the edge using Arduino boards and ESP8266 microcontrollers to reduce redundant data before transmission.

The processed data is decomposed into smaller units and distributed across multiple computing nodes for parallel analytics. Each node applies the K-Nearest Neighbors (KNN) algorithm, trained on historical weather data, to perform localized predictions. The partial results from different nodes are then aggregated to generate a comprehensive weather forecast. This output is made accessible through a cloud-based dashboard or user interface.

This distributed decomposed analytics approach minimizes cloud dependency, reduces latency, and enhances scalability. It is particularly suitable for real-time applications in agriculture, disaster management, and climate monitoring. The system’s ability to detect anomalies and predict extreme weather conditions ensures timely alerts, enabling stakeholders to take proactive measures to mitigate potential risks.

This methodology successfully achieves Distributed Decomposed Data Analytics by:

- Distributed Processing: Data collection, preprocessing, model training, and inference are distributed across multiple platforms, including AWS S3, Google Cloud, and EC2.
- Decomposed Workflow: The process is divided into independent steps (sensor data collection, preprocessing, model training, and inference) while ensuring seamless integration.
- Advanced Analytics: The application of machine learning techniques for weather prediction classifies and forecasts environmental conditions effectively.

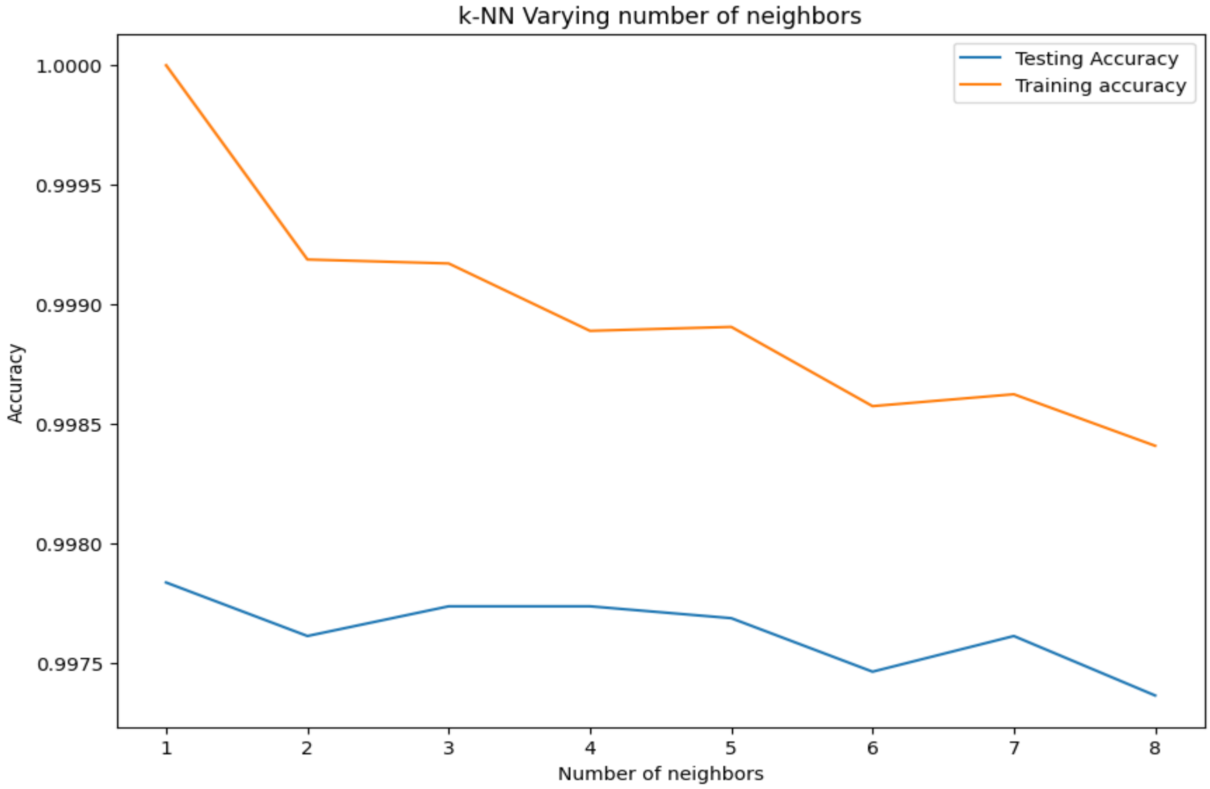


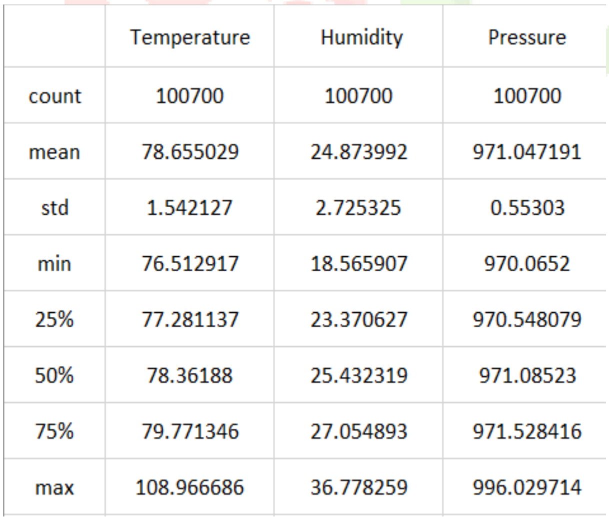
Fig 3. Performance of K-NN Classifier

V. Results

The initial implementation is focused on the IoT sensor data analytics. BME 280 is a commonly used IoT sensor for sensing temperature, humidity and air pressure A dataset having 100,700 samples is used for the analysis of temperature, air pressure and humidity data. The data description is as follows:

Table I

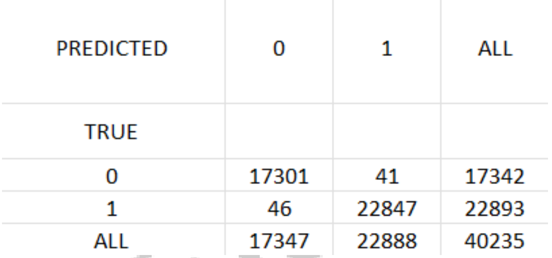
Data Description for BME280 Data



	Temperature	Humidity	Pressure
count	100700	100700	100700
mean	78.655029	24.873992	971.047191
std	1.542127	2.725325	0.55303
min	76.512917	18.565907	970.0652
25%	77.281137	23.370627	970.548079
50%	78.36188	25.432319	971.08523
75%	79.771346	27.054893	971.528416
max	108.966686	36.778259	996.029714

Training Accuracy: 0.998409

Testing Accuracy : 0.997365



PREDICTED	0	1	ALL
TRUE			
0	17301	41	17342
1	46	22847	22893
ALL	17347	22888	40235

Conclusion

This project introduces a distributed decomposed analytics framework for weather prediction, integrating IoT sensors, edge computing, and machine learning to enhance forecasting accuracy. By leveraging BME280 sensors for real-time environmental data collection and Arduino-based edge processing, the system reduces cloud dependency and improves efficiency. Parallel analytics using the K-Nearest Neighbors (KNN) model ensures rapid and scalable weather predictions. The proposed approach not only enhances anomaly detection and extreme weather forecasting but also provides valuable insights for agriculture, disaster management, and environmental monitoring, enabling informed decision-making and improved resilience against climate uncertainties.

REFERENCES

[1] Biswas, M., Dhoom, T., & Barua, S. (2018). *Weather forecast prediction: An integrated approach for analyzing and measuring weather data*. International Journal of Computer Applications, 182(34), 20-24. doi:10.5120/ijca2018918265

[2] Jakaria, A. H. M., Hossain, M. M., & Rahman, M. A. (2020). *Smart weather forecasting using machine learning: A case study in Tennessee*. arXiv preprint arXiv:2008.10789

[3] V. A. Bharadi and S. Tolye, "Distributed Decomposed Data Analytics of IoT, SAR and Social Network Data," Finolex Academy of Management and Technology, Ratnagiri, India, Year

[4] Gan, L., Man, X., Zhang, C., & Shao, J. (2024, May 9). EWMoE: An effective model for global weather forecasting with mixture-of-experts. arXiv preprint arXiv:2405.06004.

[5] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022, November 3). Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. arXiv preprint arXiv:2211.02556.

[6] Agarwal, A. B., Rajesh, R., & Arul, N. (2023, October 17). Spatially-resolved hyperlocal weather prediction and anomaly detection using IoT sensor networks and machine learning techniques. arXiv preprint arXiv:2310.11001.

[7] Gan, L., Man, X., Zhang, C., & Shao, J. (2024, May 9). EWMoE: An effective model for global weather forecasting with mixture-of-experts. arXiv preprint arXiv:2405.06004.

[8] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022, November 3). Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast. arXiv preprint arXiv:2211.02556.

[9] Agarwal, A. B., Rajesh, R., & Arul, N. (2023, October 17). Spatially-resolved hyperlocal weather prediction and anomaly detection using IoT sensor networks and machine learning techniques. arXiv preprint arXiv:2310.11001.

[10] Gan, L., Man, X., Zhang, C., & Shao, J. (2024, May 9). EWMoE: An effective model for global weather forecasting with mixture-of-experts. arXiv preprint arXiv:2405.06004.