National College of Ireland

Hybrid Machine Learning Approach Towards Anomaly Detection and Data Quality Assessment Of IoT Weather Data

MSc Research Project
Data Analytics

Saket Abhaykumar Kulkarni
Student ID: 23102381

School of Computing
National College of Ireland

Supervisor: Jaswinder Singh

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Saket Abhaykumar Kulkarni
Student ID:	23102381
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Jaswinder Singh
Submission Due Date:	12/08/2024
Project Title:	Hybrid Machine Learning Approach Towards Anomaly Detection and Data Quality Assessment Of IoT Weather Data
Word Count:	6471
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	13th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid Machine Learning Approach Towards Anomaly Detection and Data Quality Assessment Of IoT Weather Data

Saket Abhaykumar Kulkarni
23102381

Abstract

The growing rate of climate change requires high standards of analysis of weather data. This research focuses on climate science and its data quality assessment though clustering and anomaly detection using traditional machine learning with a combination of PCA for dimensional reduction. By using IoT weather data from almost 1200 cities around the world, the hybrid model is built to detect anomalies and classify data according to its quality. The final percentage of anomalies found in the data are between 14-15 % of the entire data with an ROC of 0.70 and the anomalies are visualised which shows clear separation between the values. The hybrid model performed better compared to individual models with evaluations metrics of Homogeneity: 0.702, Completeness: 0.527, V-Measure: 0.602, Adjusted Rand Index: 0.713. The following model is a strong base to prove that there is space for traditional machine learning in this field rather that heavy deep learning models which require much higher computational power. The pre-trained models perform well and are a the base for future development in clustering analysis of weather data.

Keywords: Machine Learning, Climate Science, Clustering, Anomalies, Data Quality

1 Introduction

Climate change and its impact on weather patterns have become increasingly significant concerns in recent years. The rise of Internet of Things has improvised the ability to collect, analyze, and interpret vast amounts of weather data Gubbi et al. (2013). This wealth of information presents both opportunities and challenges in identifying anomalies and patterns that may indicate climate shifts or extreme weather events. This research focuses on developing a hybrid machine learning model for detecting anomalies in IoT and maintaining the quality weather data. Due to the high dimensions in the weather data, the combination of Principal Component Analysis is chosen along with the Isolation forest for initial anomaly labelling and a combination of DBSCAN and Gaussian Mixture Model for final clustering and anomalies. The final results give an idea about the quality of the data for either forecasting or analysis.

The dataset used for this research is The Weather Dataset [1] from kaggle and is available as open source. The data contains readings from approximately 1200 cities around the world which shows the diversity in range of the readings. The parameters present in the data provide a collective view of weather conditions across various locations and seasons. The objective is to analyse the data and observe climate trends. Using the model mentioned above, anomalies are detected that may have significant implications for weather forecasting, climate change, and and over all environmental attack. This hybrid machine learning approach uses the strengths of each individual algorithm. , PCA helps in dimensionality reduction and identifying the most important features in the dataset while retaining the variance of 92%. Isolation Forest performs very well at detecting anomalies initially. DBSCAN is particularly useful for identifying clusters of anomalies, and GMM provides a probabilistic approach to modeling the underlying distribution of the data Chandola et al. (2009). The integration of these techniques allows for a more innovative and comprehensive analysis of weather anomalies than any single model would provide. The final results of the model provide the percentage of anomalies which can indicate how the data is spread out and if the readings are accurate to some level. The dataset and its variables are as seen in the table 1 below.

Variable Name	**Description**
`avg_temp_c`	Average temperature in degrees Celsius.
`min_temp_c`	Minimum temperature recorded in degrees Celsius.
`max_temp_c`	Maximum temperature recorded in degrees Celsius.
`precipitation_mm`	Total precipitation in millimeters.
`snow_depth_mm`	Snow depth in millimeters.
`avg_wind_dir_deg`	Average wind direction in degrees.
`avg_wind_speed_kmh`	Average wind speed in kilometers per hour.
`peak_wind_gust_kmh`	Peak wind gust speed in kilometers per hour.
`avg_sea_level_pres_hpa`	Average sea level pressure in hectopascals.
`sunshine_total_min`	Total sunshine duration in minutes.
`year`	Year of the observation.
`month`	Month of the observation.
`weekday`	Day of the week of the observation (0 = Monday, 6 = Sunday).

Table 1: Description of climate data variables.

While working on this research, the aim is to contribute to the field of data driven climate science. The insights gained from this research could potentially play a role in early warning systems for extreme weather events, and inform adaptive strategies for various sectors including agriculture, urban planning, and disaster management Rolnick et al. (2022). Climate change is a reality that is being dealt with now and ideally the biggest sector that needs constant observation and development. The affect of climate change is seen all over the world in various forms which is unpleasant to see and go through. Therefore it is extremely important to use the modern technology in hands with high quality sensors and highly performing machine learning models to provide better data for forecasting and analysis of the weather around the world. This research also shows

[1] The Weather Dataset

that the sensors are highly likely to detect a good number of anomalies and to be able to read those in real time is a challenge. Traditional methods of weather analysis and anomaly detection as valuable and are often limited in their ability to handle the volume, velocity, and variety of data generated by modern IoT weather sensors. These limitations show the need for more sophisticated, data-driven approaches that can leverage the full potential of big data in climate science.

The clustering approach provides a good view of the data and how it is distributed. The type of clusters identified by the models also help in what type of points are present in the data. Clustering machine learning models are implemented on the data and are evaluated using Homogeneity, Completeness, V, Measure, Adjusted Rand Index which provide a good understanding of the performance and the hybrid approach. The classification of anomalies provide a smoother process for analysis. The weather data collected contains a vast number of readings which shows the variance in weather all around the world. Climate science is a very import field and there is constant research being done in order to improvise forecasting. With the increase in global warming and the warmest year being 2023 Lindsey and Dahlman (2024), action must be taken as soon as possible as there is change in the climate that is not been able to keep up with.

The most important objective for this research is a develop this model to maintain and aim for the best quality of data from the IoT sensors as climate change is a huge factor affecting the entire world and the climate scientists are working on this every day. Although this requires more domain knowledge, it is an attempt to contribute in any way possible but ensuring that whatever data goes into analysis and forecasting, must go in the best version possible. A traditional machine learning approach might not be better than the big deep learning models being built over strong GPU's. Because not everyone has access to these resources and when that is the case, these smaller but efficient models come in hand and can always be developed more.

2 Related Work

2.1 IoT Anomaly Detection : A survey

The rapid expansion of IoT applications requires robust anomaly detection methods to ensure data quality and system reliability. Chatterjee and Ahmed's comprehensive survey Chatterjee and Ahmed (2022) on IoT anomaly detection methods provides insights into various techniques and their applications. The paper categorizes anomaly detection approaches into geometrical, statistical, and machine learning methods which is the motivation behind this research as well. They emphasize on the growing of machine learning and deep learning models for their flexibility and adaptability with big data and IoT data is always growing therefore there have to be ways to always tackle the size of the data that is gathered by these devices

Naive Bayes has a very strong probabilistic foundation therefore it can effectively classify normal and anomalous data points while Isolation Forest performs better in isolating outliers through random partitioning and it provides as a base model for this research. DBSCAN and GMM clustering capabilities. This brings the concentration to the current research on the hybrid model where the implementation of both these algorithms is at-

tempted Best et al. (2022).

The survey also identifies key challenges such as the need for unsupervised and semi-supervised approaches, drift adaptation, and handling multi-representation data traffic, which are pertinent to developing a comprehensive hybrid model Chatterjee and Ahmed (2022). Using the strengths of these strong machine learning models, The hybrid model aims to enhance data quality assessment and anomaly detection in IoT networks, addressing the identified gaps and advancing the field.

2.2 K-means and Naive Bayes for IoT

The research on hybrid approaches combining K-means clustering and Naive Bayes for IoT anomaly detection which is a pure clustering approach. The use of IoT devices has increased the need for robust anomaly detection algorithms that can secure data across various devices from any unnecessary use cases. The primary advantage of using a hybrid method is the combination of the unsupervised learning capability of K-means to group similar data points and the supervised learning accuracy of Naive Bayes for classification. This hybrid approach leverages the speed and scalability of K-means clustering and the precision of Naive Bayes, resulting in improved detection rates of up to 100% for different types of anomalies such as DDOS, backdoor, ransomware, and more Best et al. (2022).

Studies highlight the importance of flexibility and scalability in anomaly detection algorithms, particularly for IoT systems that generate large volumes of data. While the model proposed by Best et al. (2022) is well built and performs up to a good mark, there is still space for more as there are better performing clustering algorithms that can are are being used for this research. DBSCAN and EM clustering are a much more developed clustering algorithm compared to k-means even though k means handles the data well, it is still has some limitations. It is believed that the hybrid combination of this research is a more stronger approach towards anomaly detection.

2.3 PCA & IoT

The paper "Self-Adaptive Incremental PCA-Based DBSCAN of Acoustic Features for Anomalous Sound Detection" by Tan and Yiu (2024) presents a novel approach for detecting anomalous sounds in industrial settings, which is critical for predictive maintenance and operational efficiency. The proposed method integrates incremental principal component analysis (IPCA) with density-based spatial clustering of applications with noise (DBSCAN), enhanced by an automatic EPS calculation algorithm using a genetic algorithm. This hybrid technique effectively reduces the dimensionality of acoustic features and optimizes clustering performance, achieving a high area under the curve (AUC) of 0.84. The study demonstrates superior performance compared to existing methods like K-means++, one-class SVM, and deep learning models, highlighting its potential for real-time industrial applications in detecting machine anomalies and preventing failures Tan and Yiu (2024).

2.4 DSVDD-CAE

The paper "Robust Anomaly Detection in IoT Networks using Deep SVDD and Contractive Autoencoder" by Aktar and Nur (2024) addresses the need for securing Internet of Things (IoT) devices against increasingly sophisticated cyber threats. Traditional Intrusion Detection Systems struggle with detecting novel attacks due to their dependence on pre-defined rules and this limits their performance in diverse IoT environments. The authors propose a novel model, DSVDD-CAE, which combines Deep Support Vector Data Description with a Contractive Autoencoder to enhance anomaly detection in IoT networks Aktar and Nur (2024). This model operates in a semi-supervised learning framework, using only normal data to detect anomalies by evaluating reconstruction errors and the distance of data points from the center of a hypersphere in the latent space.

The model was evaluated on two datasets—ToN-IoT and IoTID20—demonstrating superior performance over traditional methods like KMeans, OCSVM, and Isolation Forest. The DSVDD-CAE achieved impressive precision, recall, F1-score, and accuracy metrics, significantly advancing the field of IoT security. By focusing on the compact representation of data and robust anomaly detection, the DSVDD-CAE model provides a resilient solution for protecting IoT networks from both known and novel cyber threats, ensuring a more secure and stable ecosystem. Aktar and Nur (2024). This shows the dominance of deep learning models and how well they perform with high dimensional data. This gives way to the growth of machine learning models to perform equally well with more limited resources.

3 Methodology

3.1 Overview

This section provides an overview of the methodology that is used in developing the anomaly detection system based on a hybrid approach combining Principal Component Analysis, Isolation Forest, Density Based Spatial Clustering of Applications with Noise, and Gaussian Mixture Model. The system aims to identify anomalies in climate data collected from IoT devices. The processes involved are as seen in the figure 1 and the following sections discuss the work flow and the model development in detail.

3.2 Data Acquisition

The initial stage involves retrieving the data from kaggle using the API provided for the dataset. The data is in Parquet form which is processed and converted to a dataframe which then is cleaned and thoroughly checked for unnecessary data points. EDA is performed on the data to observe the patterns and any other information that can be obtained from the data and its visaulizations. Finally, the data is pushed into a MongoDB cluster. The data is clean and can be pulled for modelling and analysis anytime. The clean data is obtained using the `pymongo` library to make the flow of data more simple and easier to access. The data is fetched into a Pandas DataFrame for further processing. The dataset is read for unnecessary columns and any other data that might not be required. The date column is converted into three different variables of year, month and week to make the numeric as well as keeping the important columns.

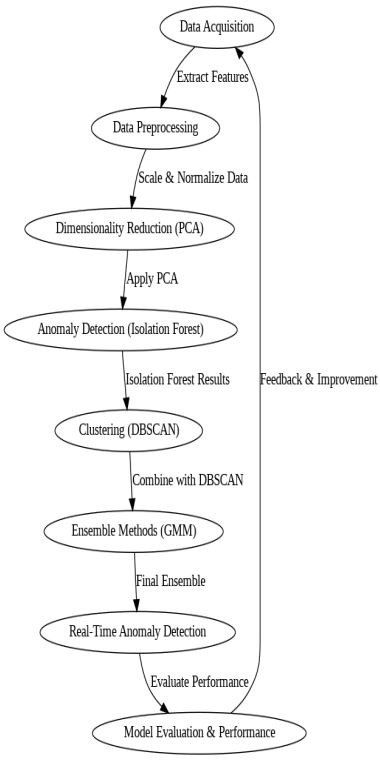

Figure 1: Methodology for the Hybrid Model

3.3 Data Preprocessing

Preprocessing on the data is done to prepare it for modelling focusing on the numeric data and also the weather data is thoroughly checked for any kind of unnecessary values and cleared for modelling. This stage involves the formation of new columns as well where the date column is converted to year, month and week to get more variance and this helps in visualisations as well. This data will be used for the modeling as the full data is getting pushed into analysis. The numerical features are all scaled up to provide a standard deviation, this is done by using the `StandardScaler` in python, which is the most important procedure for algorithms like PCA and Isolation Forest to function correctly (Lim et al.; 2023). The data is prepared for pre-processing and also scaling which sets up the stage for performing PCA.

3.4 Dimensionality Reduction using PCA

As seen in table 1 the number of variables present in the data are pretty high and covers a good range of information that is required. PCA is applied as a dimensionality reduction technique to transform the original features into a smaller set of components while retaining 90% of the variance in the data which will allow the modelling process to use the data as a whole but without each variable going individually. This reduction not only helps in minimizing computational stress but also prevents the side affects of dimensional data, which can lead to over fitting in machine learning models. The 6 transformed features which reatin the required variance are the principal components and are the input for the anomaly detection algorithms (Kim and Vasarhelyi; 2024). The performance and the effectiveness of PCA is evaluated by examining the explained variance ratio, ensuring that the most significant features are retained.

3.5 Anomaly Detection using Isolation Forest

The Isolation Forest algorithm is used as the primary method for detecting anomalies within the dataset and the labels provided by the algorithm will hold as the true lables for comparisons and evaluations. The algorithm isolates anomalies by randomly selecting a feature and then splitting the data along that feature. The model is trained on the principal component dataset derived from PCA in the previous subsection 3.4, and it identifies anomalies based on their relative isolation and how far they are form the mean. Data points that are isolated quickly are more likely to be anomalies. The contamination parameter is set to 0.09 as the initial objective with anomaly detection is to keep the number of anomalies within 10% of the entire data. The results are stored in a new column labeled `Anomaly` and this will serve at the true labels going further. After running the isoaltion forest model, the initial numeric dataset is updated with the anomaly labels from the results.

3.6 Clustering and Ensemble Methods - DBSCAN & GMM

To build upon the anomaly detection process, the Density Based Spatial Clustering of Applications with Noise algorithm is applied. DBSCAN is a clustering method that groups points in dense regions and labels points in sparse regions as noise, which are potential anomalies. This method is particularly useful for identifying clusters with irregular shapes, which might not be detected by other clustering methods Ester et al. (1996). The DBSCAN results are combined with those from the Isolation Forest to create a common anomaly label column which is an overlap between the two results. After this, a Gaussian Mixture Model is used to assign data points to clusters based on probability of being an anomaly or not. Data points with a low probability of belonging to any cluster are flagged as anomalies. This ensemble approach leverages the strengths of each algorithm, providing a robust solution for anomaly detection (Petrov et al.; 2023). The anomalies labelled are finally combined and updated in the dataset as the final anomalies and is tested for the performance.

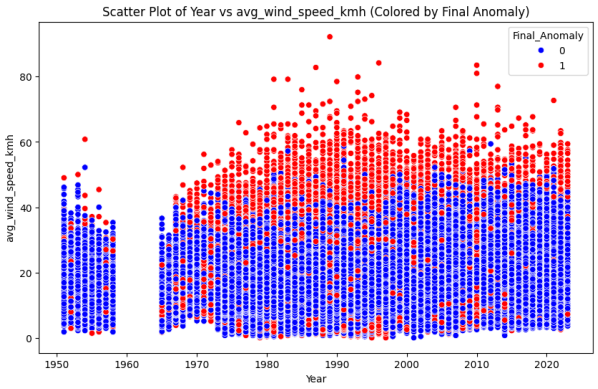

Figure 2: Final Anomalies From the Hybrid Model (Variable - Average Wind Speed)

The figure above shows the anomalies detected from the hybrid model and with the combination of PCA on the data. There is good divide between the points as seen in the plot and the outer layer of points are all detected as anomalies. There is room for improvement and that can only be done with more parameter tuning. The current model is set to achieve anomaly rate between 10 and 15% of the entire data for further analysis.

3.7 Real-Time Anomaly Detection

The final stage of the methodology involves implementing a real-time anomaly detection system using the pre-trained models. The trained models are applied to data pulled from the MongoDB database. Each batch of incoming data undergoes the same preprocessing steps and PCA transformation, before being passed through the anomaly detection algorithms. The system calculates the anomaly scores and categorizes the data points accordingly. A warning function is integrated to notify users when the percentage of anomalies in a batch exceeds a predefined threshold, enabling timely responses to potential issues in the data. This real-time capability is important for applications requiring continuous monitoring and immediate action, such as climate monitoring systems.

```
BEGIN

    // Step 1: Data Preparation
    CONNECT to MongoDB
    FETCH and CLEAN climate data
    EXTRACT time-based features (year, month, weekday)
    STANDARDIZE the dataset

    // Step 2: Dimensionality Reduction
    APPLY PCA (retain 90% variance)
    TRANSFORM data into principal components

    // Step 3: Anomaly Detection
    USE Isolation Forest to label anomalies
    APPLY DBSCAN to identify clusters and noise

    // Step 4: Ensemble Method
    COMBINE results from Isolation Forest, DBSCAN, and GMM
    IDENTIFY final anomalies via majority voting

    // Step 5: Real-Time Monitoring
    WHILE (new data available):
        FETCH and PREPROCESS data
        APPLY PCA and detect anomalies
        IF anomalies exceed threshold:
            TRIGGER alert
    END WHILE

    // Step 6: Evaluation and Feedback
    EVALUATE performance with ROC/AUC
    RETRAIN models if performance drops

END
```

Figure 3: Pseudo Code for the Final Pre Trained Model

3.8 Model Evaluation and Performance

The performance of the anomaly detection system is evaluated using various metrics like the percentage of identified anomalies, the ROC curve, and the area under the curve. The ROC curve is useful for visualizing between the true positive rate and false positive rate at different threshold settings. The combined use of Isolation Forest, DBSCAN, and GMM ensures that the model is capable of identifying a wide range of anomalies. The combination of the clustering models are evaluated using Homogeneity, Completeness, V, Measure, Adjusted Rand Index which provide a good understanding and the hybrid approach. The effectiveness of the model is also demonstrated through visualizations, such as scatter plots which highlight the distribution of anomaly scores and the separation of anomalous points from normal ones (Kim and Vasarhelyi; 2024).

4 Design Specification

4.1 Data Management and Preparation

The design of this anomaly detection system is structured to ensure robust data preprocessing, efficient dimensionality reduction, and accurate anomaly identification within climate data. Initially, climate data is acquired from a MongoDB database and undergoes

a cleansing process to remove irrelevant columns. Essential time-based features such as the year, month, and weekday are extracted to standardize the dataset. This preprocessing phase is critical in ensuring the data's consistency and reliability, forming a solid foundation for subsequent analysis.

4.2 Assumptions and Constraints

Assumptions: The data used is from sensors and it is assumed that the IoT devices collecting the climate data are reliable. This ensure some accuracy in the data being analyzed. Additionally, the data sources in this case are not in real time but are updated only once in a while, there is no continuous flow of data.

Constraints: The system operates within the computational limits of the available infrastructure at hand. It is also important to consider how fast the IoT sensors for weather are growing and will always require more processing power for real time analysis.

4.3 Dimensionality Reduction and Modelling

To handle the high dimensionality of the dataset, the design incorporates Principal Component Analysis. By reducing the dataset to principal components that retain 90% of the variance, PCA enhances computational efficiency and avoids overfitting, allowing the system to focus on the most significant features. The anomaly detection mechanism is a key component, beginning with an Isolation Forest model to identify and label potential anomalies. This is followed by the application of DBSCAN.

4.4 Ensemble Approach - Pre Trained Models

The system employs an ensemble method that combines the outputs of the Isolation Forest, DBSCAN, and a Gaussian Mixture Model (GMM), leveraging the strengths of each to provide a comprehensive anomaly detection approach. Designed for real-time detection, the system continuously fetches new data, applying the models to trigger alerts when anomalies exceed a predefined threshold. The design emphasizes continuous improvement through regular evaluation using metrics such as ROC/AUC. If performance declines, the system is capable of retraining models and adjusting parameters, ensuring long-term effectiveness and adaptability to changing data patterns.

4.5 Future Work and Scalability

1. **Scalability:** The system is designed with minimal scalability, ensuring it can handle the data as more IoT sensors are deployed and more data comes in. The current model is fit for the data but needs to be refined for new data with more dimensions to come through the model. Future work will focus on optimizing the model for large-scale industry deployments and even incorporating cloud based processing to manage computational demands.

2. **Further Model Refinement:** Future iterations of the model may include additional machine learning techniques, such as an additional layer of an auto-encoder for anomaly detection, to further enhance the system's accuracy and robustness.

4.6 Ethical Considerations

The dataset use is sourced ethically from the Kaggle dataset repository as it is open source and is also updated every Sunday with more readings from around 1200 cities. This dataset is considerably large in size and is varied from 1950 to 2024. It can be accessed by anyone and as mentioned, it is publicly available on Kaggle [2]

Data Bias and Fairness: The system is evaluated for potential biases, particularly in how anomalies are detected across different geographic regions or sensor types. It is important to ensure fairness as biased models could lead to misinformed decisions that affect certain areas or populations Ferrara (2023).

Representation: The representation of the training data is well spread out as it ensures the coverage of data from almost all around the world and in this case. Around 1200 cities are covered in the dataset. Ensuring that it covers a diverse range of climates, locations, and seasonal variations. This will help prevent the model from being biased towards the conditions most frequently encountered Ferrara (2023).

Transparency: Transparency in the decision-making process is essential. The system should provide results that allow users to understand why certain data points are classified as anomalies, facilitating trust and accountability in the system's outcomes Ferrara (2023).

Societal Impacts: The potential societal impacts of the model's decisions are an important as they might influence climate related issues with respect to the public. Ethical guidelines are very important as the people are the ones who will be affected by any change in climate so it is important in making sure that the people are informed right Ferrara (2023).

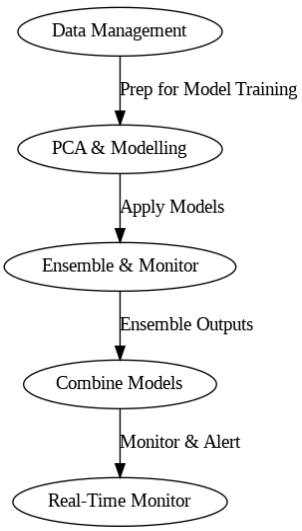

Figure 4: Design Specifications Flowchart

[2]The Weather Dataset

5 Implementation

The Research is implemented all in one system and the system used for the research is a Macbook Air with the M2 chip, 8 GB of unified memory and 256 GB of storage in the device. The programming language used throughout is python. The code runs on a google colab notebook and the hardware accelerator used is the CPU of the device. MongoDB NoSQL database is used to store the data that is fetched from the API for the modelling process to be easier as the data in MongoDB is the clean version.

5.1 Data gathering and Preprocessing

The implementation of this hybrid anomaly detection system begins with data acquisition from a MongoDB database where the clean data is stored in order to make the modelling easy. Using the `pymongo` library, the system pulls the data from a MongoDB collection that stores climate data retrieved from IoT devices. After fetching the data, the unnecessary columns are removed, and the date column processed and it includes extracting key temporal features such as the year, month, and weekday from the date column. The dataset is then split into training and test data, mainly selecting the numerical data types for the process going forward.

Figure 5: Data Stored in the MongoDB Cluster

5.2 Principal Component Analysis

The data contains various numeric variables which provide different kind of readings, most of them are very relevant and important when it comes to climate science therefore to reduce the dimensionality of the dataset, PCA is applied. The goal of PCA is to identify the most significant features while retaining at least 90% of the variance in the data. This step not only reduces computational complexity but also helps understanding of the dimensionality. Not handling multidimensional data can degrade the performance of machine learning models. The numerical data is scaled to maintain equal distribution and is transformed into principal components. After which the data is used in the anomaly detection models. The PCA model's effectiveness is validated by analyzing the explained variance ratio as seen in Figure 6, confirming that the selected components capture most of the data's variability (Lim et al.; 2023).

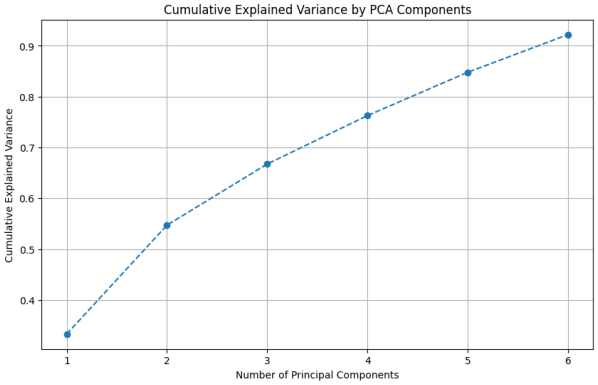

Figure 6: PCA Variance Scree Plot

As also seen in the table 1 it is very clear how high the dimensions of the data is and to run this in a model every time, it will crash the device. PCA has played the most important role in this research as it is very important consider all the values and still maintain the variance. All the models are run on the PCA data and then updated in the original dataset.

5.3 Isolation Forest for Anomaly Detection

The Isolation Forest algorithm is one of the primary methods used for detecting anomalies in the dataset. This unsupervised learning technique is particularly effective in identifying outliers by isolating data points that appear to be anomalies with respect to the rest of the dataset. The model is trained on the principal components derived from the PCA. The labels provided by the isolation forest model will be the true labels used for the evaluation and the hybrid approach. The anomalies detected by the Isolation Forest are visualized using scatter plots and it helps is observing the clear separation between the points (Kim and Vasarhelyi; 2024).

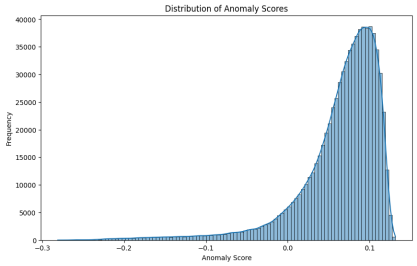

Figure 7: Anomaly Scores - Isolation Forest Model

5.4 Clustering and Ensemble Methods

In combination with the Isolation Forest, DBSCAN algorithm is used to identify clusters and detect unnecessary data points, which may signify anomalies. The DBSCAN results are combined with those of the Isolation Forest to create a hybrid model that improves the accuracy of anomaly detection. After which a Gaussian Mixture Model is used to assign each data point to a cluster based on probability and expectation maximisation. The

clusters are evaluated using Homogeneity, Completeness, V, Measure, Adjusted Rand Index which provide a good understanding of the performance and the cluster identification. The ensemble approach combines the strength of each model and also combines the anomaly readings to get a more accurate reading of the total anomalies culminating in a robust model.(Petrov et al.; 2023).

5.5 Real-Time Anomaly Detection

The final implementation involves real time anomaly detection, where the pre trained models are applied to incoming weather data from MongoDB. The system is designed to process new data at regular intervals, applying the preprocessing steps, PCA transformation, and anomaly detection algorithms in sequence. The results are analyzed to determine the percentage of anomalous data points, and even alerts are triggered if the anomaly rate exceeds a predefined threshold which is defined. This real time detection capability is critical for applications that require immediate response to potential issues detected in the data especially with weather data being large in size with various different types of readings.

6 Evaluation

This section provides an overview of some results achieved by the hybrid model along with what improvements are required for better performance. The table 6.1 provides a performance detail of the models individually and after the combination.

Model	**Parameters**	**Anomalies**	**Percentage**
Isolation Forest	contamination = 0.09	34,768	9%
DBSCAN	eps = 0.6, min_samples = 20	33,908	8.78%
Combined Anomalies	none	44,753	11.58%
GMM	n_components = 2	30,905	8%
Final Anomalies	none	53,665	13.89%

Table 2: An overview of models, their parameters, and anomaly detection results. Random State = 2381.

6.1 PCA - Principal Component Analysis

Due to the high dimensions in the data with various variables playing a factors in the quality of data and readings obtained by the sensors. It was important to reduce the dimensions of the data to be able to run models on the data. The pca is run on the entire data after scaling the data to standardize the values. After running PCA on the numeric data, 6 principal components are picked which maintain 92% variance in the data. This data is used for all modelling cases and then combined with the other models output values to create the hybridization between the machine learning models. The PCA plays the most important role as it sets the stage for model development and the role of dimensional reduction is most suited for this particular research. The data is decreased in dimensions while retaining a high variance level.

6.2 Isolation Forest - Base Model

The initial model is based on an Isolation Forest algorithm, which has proven beneficial for unsupervised anomaly detection. Isolation Forest operates by isolating observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. This process is repeated until the observation is isolated. The results indicate that the Isolation Forest model effectively identified anomalies in the weather dataset, providing a solid baseline for further model enhancements. The Model is used to run on the principal components from the PCA model, The contamination level is set to 0.09 to keep the anomaly level below 10% and the model performs well labelling the anomlaies according to the data point and after plotting the anomalies for certain variable it is very clearly seen how the anomalies are differentiated from the normal points. The model detects a total of 34768 anomaly points for this parameter and as seen in the plot below. The labels provided by the isolation forest model will be the base labels for the research. The anomalies detected are evaluated by visualisations the anomalies are clearly defined and differentiated as seen in the plot for anomlaies in the average temperate throughout the years. 8.

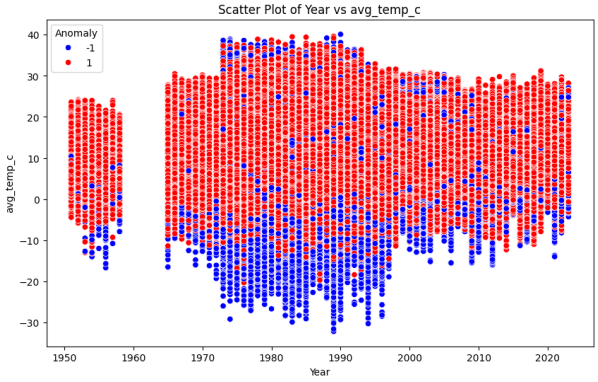

Figure 8: Anomalies Detected For the Average Temeprate Variable

6.3 Isolation Forest / DBSCAN

The Isolation Forest outputs are saved into the main dataframe and the DBSCAN is used on the data to cluster the data according to normal and anomaly points. The DBSCAN is used with the eps value of 0.7 considering the large dataset and a minimum sample of 20 which is at least 20 data points in a clusters because of the closeness of the data which is seen via visuals. The model performs well in clustering the data into different sections especially for the anomalies and the normal point and some additional small clusters which are identified by the model according to the parameters. The labels generated for the anomalies from the DBSCAN model are then updated on the main dataset as well. This allows a combination feature between the previous readings and the current reading which gives a combination value of anomalies through out the data. This is then evaluated to check the values of anomalies over all and the true values after the combination. This gives a more robust value where two values from different models are overlapped to get the final value.

6.4 PCA / Isolation Forest / DBSCAN / GMM

The final hybrid model combined Isolation Forest, DBSCAN, and Gaussian Mixture Models to use the strengths of each algorithm and combine the final set of anomalies detected. The model runs with 2 components as the requirement for this is just to spot the anomalies and the threshold for the anomaly population is set to 9% which is the same as the isolation forest model, this allows the model to maintain the overall threshold of anomalies within between 10 - 15% of the over all data after the final combination of the anomaly scores. The readings are taken into consideration and combined with the previous anomaly scores to get the overlapped common scores which will be the final anomly labels for the data.

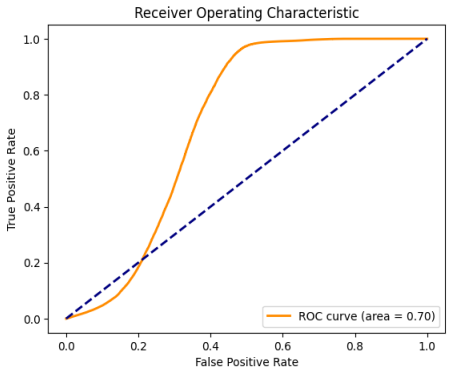

Figure 9: AUC - ROC for the Final Anomalies

The plot for the AUC-ROC shows the results of 0.70 which means that at least 70% of the time the model is able to differentiate between the points and label them as anomalies. This shows decent performance based on the final labels of the hybrid model. As seen in figure 9, there is room for improvement and the value can be increased based on the readings and with more precise classification, it can be developed to detect anomalies in a stronger manner. This final model gives a base for higher development in the field as this is not a deep learning model which takes a huge load of resources to run instead, its taken the least of resources which shows that there is space for a traditional machine learning approach with these strong algorithms.

6.4.1 Cluster Evaluation

The evaluation metrics used for the clustering methods are the Homogeneity, Completeness, V-Measure and Adjusted Rand Index. These are key metrics for the evaluation of the clustering algorithms and its performance after creating the clusters. Completeness checks the assignment of all class members to the same clusters. The homogeneity checks for the purity of each clusters formed, ideally assessing the clusters for the class of members. ARI measures similarity between the predicted clusters and the true labels which are the initial labelling done by the isolation forest and that is compared to the final anomaly values. The values closer to 1 indicate better clustering. V-Measure is the harmonic mean of homogeneity and completeness, providing a evaluation of both metrics. These metrics are important in understanding the effectiveness of clustering algorithms in

separating groups within data Amigó et al. (2009) Hubert and Arabie (1985) Rosenberg and Hirschberg (2007).

Metric	DBSCAN	GMM Model	Hybrid Anomalies
Homogeneity	0.414	0.244	0.702
Completeness	0.410	0.265	0.527
V-Measure	0.412	0.254	0.602
Adjusted Rand Index	0.625	0.463	0.713

Table 3: Evaluation metrics comparison across DBSCAN, GMM Model, and Hybrid Anomalies.

The table above 3 is the results after evaluating the clusters according to the metrics discussed previously. The individual models both performed below average not showing the best performance in terms of clustering the data points. The GMM model shows very low values which means the clustering was not at it best and the points have just been assigned to two clusters based on some level of probability. The DBSCAN model also performs along the similar line but slightly better in terms of forming the clusters and identifying the anomaly points. As seen clearly from the table, The final hybrid model which is a combined approach of all models, performs much better in identifying the clusters and the metrics show as well. The purity of the data is at the good rate as well as the assignment of different classes to each cluster is also performed at a good rate. The over all adjusted rand index shows good similarity between the true labels and the final labels. This proves that the hybrid model is a good approach towards the combination of clustering for the data quality and anomaly checks.

6.5 Hybrid Pre-trained Model

A prototype hybrid real time model is also developed using the pre-trained model and its features. This model performs well and it accepts a threshold values as well a batch size for the data. Therefore, depending on the size of the data the user can set a threshold to detect the anomalies. It also includes a warning system which helps in letting the user know when the anomaly rate has crossed the threshold.

6.6 Discussion

The findings from the various experiments point out the importance of using a hybrid approach for anomaly detection in weather data. This is a crucial part of climate science with the increase in temperatures and other climate factors all over the world becoming hard to predict and issue warning to areas that might be under danger. The Isolation Forest provided a solid baseline, while the integration with DBSCAN, PCA, and GMM progressively enhanced the model's performance.The final hybrid model performce and its evaluation metric values shows the strength of the combination over the individual models and this proves the space for growth in this model. The pre-trained model based on this also performs well based on this flagging around 10-15% of the data as anomalies.

The overall performance of the model still has space for improvisation as this is limited to the current dataset and its features, the growth of this model requires more time ,

resources as well as domain knowledge which will help in understanding the data better. The final model runs on the dataset that is pulled from MongoDB and this makes it very limited as there is no test case with unseen or unused data during training. This makes it hard to evaluate the pre-trained model. The model that is run is a prototype and only a base for a hybrid and that it leverages the strengths of each algorithms and combines the results. The model also abides by the threshold value as well the size of the data that is taken in. The batch limit for each iteration can be altered. The model also gives an idea about a warming system where if too many anomalies are found. Other considerations for the model would be to introduce quantization methods and optimisation methods like used in Tan and Yiu (2024) for a self adaptive PCA but in this case, a self adaptive feature for parameter tuning would make the process much smoother.

6.6.1 Final Analysis

In conclusion, this research adds to the field of Anomaly detection in IoT environments and data driven climate science by demonstrating the effectiveness of a hybrid machine learning model in detecting anomalies. More development for real time detection and analysis will prove to hold value for maintaining the data quality and eliminate the unnecessary data points. As mentioned previously, this is a very sensitive topic with the high increase in global warming that we all see happening around the world. This data has also shows the fluctuation and temperature and other factors influencing the weather over the years. The data is finally loaded back and the anomalies are removed, the plot below 10 is an example plot of temperature values after the removal of the anomalies and also a plot of the Evaluation metrics performed for cluster quality 11. It is very clearly seen how the combination of the models performs much better in comparison to the individual contributions.

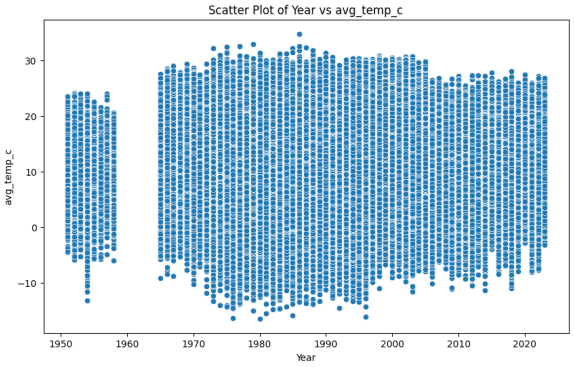

Figure 10: Average Temperature after Removing the Anomalies Detected by the Model

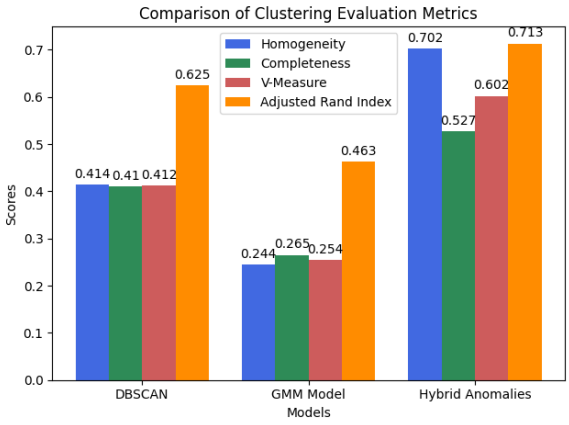

Figure 11: Clustering evaluation plot

7 Conclusion and Future Work

This study has developed and implemented a hybrid machine learning model for detecting anomalies in IoT weather data, integrating Isolation Forest, DBSCAN, and GMM

model along with PCA to reduce the dimensions of the data. The approach has demonstrated some potential in identifying unusual weather patterns across diverse temporal data. The model's ability to detect both individual points and anomalous clusters is proven by the evaluation metrics and the higher scores that is achieved by the hybrid model over the individual models. However, this work also highlights areas for future research and improvement. One promising direction is the incorporation of bigger models such as autoencoders and the introduction of Quantization methods, to capture more complex temporal patterns in large weather data. Additionally, exploring the integration of external data sources, such as satellite imagery could provide contextual information to enhance anomaly detection accuracy but this does involve more ethical considerations as it is a much bigger area of research to step into. Extending this approach to real-time anomaly detection in streaming IoT weather data presents a chance prepare early warning systems for extreme weather events. Scalability as well as the ability to use this on any kind of weather data must also be a part of the future developmet.

References

Aktar, S. and Nur, A. Y. (2024). Robust anomaly detection in iot networks using deep svdd and contractive autoencoder, *IEEE Transactions on Big Data* **8**: 1–15.

Amigó, E., Gonzalo, J., Artiles, J. and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints, *Information Retrieval* **12**(4): 461–486.

Best, L., Foo, E. and Tian, H. (2022). A hybrid approach: Utilising kmeans clustering and naive bayes for iot anomaly detection, *arXiv preprint arXiv:2205.04005* .

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey, *ACM Computing Surveys* **41**(3): 1–58.

Chatterjee, A. and Ahmed, B. S. (2022). Iot anomaly detection methods and applications: A survey, *Internet of Things* **19**: 100568.
URL: *http://dx.doi.org/10.1016/j.iot.2022.100568*

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226–231.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, *arXiv preprint arXiv:2304.07683* .

Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M. (2013). Internet of things (iot): A vision, architectural elements, and future directions, *Future Generation Computer Systems* **29**(7): 1645–1660.

Hubert, L. and Arabie, P. (1985). Comparing partitions, *Journal of Classification* **2**(1): 193–218.

Kim, Y. and Vasarhelyi, M. (2024). Anomaly detection with the density based spatial clustering of applications with noise (dbscan) to detect potentially fraudulent wire transfers, *The International Journal of Digital Accounting Research* .

Lim, J., Han, E.-S., Kim, D. H. and Lee, B. K. (2023). An optimal clustering algorithm for second use of retired ev batteries using dbscan and pca schemes considering performance deviation.

Lindsey, R. and Dahlman, L. (2024). Climate change: Global temperature. Accessed: 2024-08-12.
URL: *https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature*

Petrov, P., Kacprzyk, J. and Bureva, V. (2023). Generalized net model of density-based spatial clustering of applications with noise (dbscan) algorithm and its application on diabetes dataset, *Book Chapter*.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A. and Luccioni, A. (2022). Tackling climate change with machine learning, *ACM Computing Surveys* **55**(2): 1–96.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420.

Tan, X. and Yiu, S. M. (2024). Self-adaptive incremental pca-based dbscan of acoustic features for anomalous sound detection, *SN Computer Science* **5**(542): 1–10.