School of Engineering

Diploma in AI & Data Engineering

EGT209

Data Engineering Project

T2G2

Final Report

AY2024S1

Date Submitted: Aug 15, 2024

Table of Contents

1. Project Team Members

S/N	Module Group	Name	Admin Number	Role and Responsibilities
1	EGT209 – T2	Min Phyo Thura	233523A	Team Leader, Software++
2		Lim Jin Bin	221128Z	Member, Dashboard, Report
3		Mohammad Habib	231880L	Member, Hardware, Report
4		Alexander Chan	230648A	Member, Hardware, Report

2. Executive Summary

The Envi-Optimizer project is aimed to enhance the efficiency and safety of advanced manufacturing facilities by optimising environmental factors such as temperature, humidity, and air quality. These three sensor values are collected over two weeks via Arduino and are compared alongside weather conditions during the same period. This led us to confirm that some weather conditions such as extreme heat or heavy rain does affect the indoor environment. Furthermore, by comparing our collected data against other groups, we ensured that our data is accurate and reliable. Although no actual feedback from workers was involved, we were able to determine the optimal ranges of indoor parameters through KNN clustering, and determined what is normal and what is not through Recurrent Neural Networks (RNN). Our findings pointed out the need for a proper dehumidifier as well as a better insulation for the shutter door among others. Finally, a self-explorable dashboard with interactive slicers is attached together with this report to allow the reader to freely explore, potentially leading to more insights.

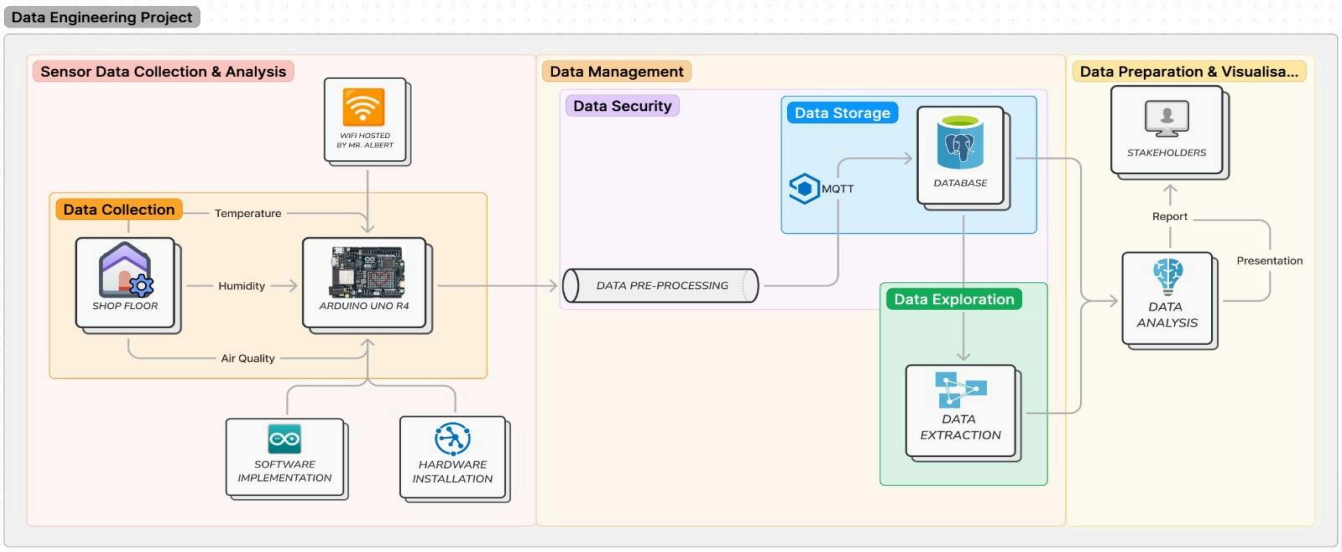
3. Introduction

The introduction should provide an overview of the project and set the context for the rest of the report. It should include a brief description of the project objectives, scope, and methodology. It should also provide some background information on the topic being studied and explain the significance of the project.

In the present scenario, there are modern manufacturing units with an array of machineries programmed to operate continuously, often in conjunction with human workers. This continuous operation leads to indoor pollution, which affects the health of workers due to heat and dust and also impairs the proper functioning of the machines due to the changing environment. Using an Arduino, a temperature-humidity sensor, and an air quality sensor, our goal is to tackle these challenges by developing a seamless data pipeline that supports data-driven decision-making in controlling these indoor parameters. The stakeholders will be able to optimise the environmental factors on the shop floor for human workers and their equipment use, ensuring energy efficiency and resource use. Below are three main pillars for the success of Envi-Optimizer:

- **Data**: Accurate collection and analysis of data will be carried out in five phases: collection, pre-processing, storage, exploration, and analysis.
- **Security**: Given the substantial resources invested in Envi-Optimizer, prioritising data security is crucial to protect the collected sensor data and end results against malicious theft.
- **Enhancements**: To exceed stakeholder expectations, further enhancements to the project are suggested, such as monitoring of the intelligent control systems and cloud storage infrastructure.

The *Data Pillar* will be discussed in detail in this report through methodology used, results and findings as well as recommendations.



Design Architecture

4. Methodology

Data collection

The Arduino is programmed in **C++** to collect sensor values and publish the collected data to the **MQTT** broker every 10 minutes. The server, which is subscribed to the MQTT broker and is configured in **Node.js**, listens for the relevant data topics (t2g2). Once the server identifies the correct table name, the broker transfers the data to the server whose script translates it into SQL statements for uploading and storing on the SQL server. The data is then stored in a table (t2g2) in a local **MySQL** database, mydb1. In addition to sensor data collection, **Beautiful Soup** is used as well to scrap other related and required data from the internet such as the weather data of Ang Mo Kio New Town from timeanddate.com.

- **Tools**:
  - *C++(Arduino IDE)* - suitable for Arduino development
  - *MQTT* - lightweight and used to connect with devices with limited connectivity
  - *Node.js* - a common scripting language suitable for json format
  - *MySQL* - Relational database management system used for data storage
  - *Python(Beautiful Soup)* - a powerful python library for web-scraping with an extensive community

- **Limitation/Challenges:**
    - The Arduino collected complete data only until 16 July, with subsequent days having incomplete data.
    - The air quality sensor is not quite accurate as it is deployed without fine-tuning. It is quite sensitive to human traffic as well.
- **Mitigation:**
    - To address the former, we reached out to t1g2, t2g3, and t2g5 for additional data to use for comparison and imputation.

Data Retrieval

Mosquitto: Small and lightweight MQTT broker
MySQL Connector: enables python programs to access MYSQL database
ODBC: Used to connect the MYSQL database to our Power BI dashboard

Data Preparation

- **Data Evaluation:**
    - After exploring the datasets, we determined that only t1g2's data was suitable for imputing the missing data from 16 July onwards.
    - t2g3's dataset was incomplete, covering only weeks 13-15.
    - t2g5's data was unsuitable due to its location on the other side of the room, resulting in slightly different values.
- **Data cleaning:**
    - **Splitting the Payload:** Dividing the data into three columns: temp, hum, and air.
    - **Handling Null and Duplicate Values:** Remove any duplicate or null entries to ensure data quality.
    - **Resampling the Data:** Adjusting the data frequency to match the desired time intervals for analysis.

Data Analysis

Data analysis techniques: Using Google Colab, we created:

- line graphs and scatter plots: To visualise the trends of temperature, humidity, and air quality over time
- box and whiskers plot: To assess the variance in temperature, humidity, and air quality for each day

Data Visualisation: An interactive and intuitive dashboard has been created to provide the necessary information of the factory conditions.

5. Results and Findings

Variance

- **Temperature (Figure 1):**
    - On Monday and Tuesday, the sensors exhibited the highest variance in temperature, suggesting significant fluctuations. These fluctuations may be the result of more people entering and leaving the room on these days.
    - Saturday and Sunday showed lower variance with relatively stable temperatures. This stability is likely due to the consistent environmental control and reduced number of people over the weekend.
- **Humidity (Figure 2):**
    - The variance in humidity appears to be higher at the beginning and end of the week (Monday and Sunday) while the mid-week days (Wednesday to Friday) display the most stability. On Monday, the variance is high as internal conditions could be stabilising after a potentially less controlled environment over the weekend. On Sunday, environmental controls may be less strict, leading to significant humidity variations.
- **Air quality (Figure 3):**
    - It can be seen that Monday showed the largest variance in air quality. Like temperature, this could be caused by the influx of people going in and out of the room. It could also be caused by the usage of the machinery within the class.

Weekly trends

- **Temperature: (Figure 4)**
    - The temperature was observed to have peaked towards the end of the week, specifically on the 7th and 14th of July.
    - The weekend's lack of activity may have an impact on these peaks, resulting in the air conditioner being set to a lower level or turned off.
- **Humidity: (Figure 5)**
    - The humidity was observed to have peaked on the morning of 13th July.
    - There are multiple peaks and valleys where humidity spikes and drops.
    - The weather is cloudy before the peaks of humidity. However, the cloudy
    - weather is not the cause of the high humidity, as it is on the contrary as the high humidity caused the formation of clouds.
- **Air quality: (Figure 6)**
    - Air quality was observed to have peaked on 2nd and 8th July.
    - This could be due to the operations of the machinery within the classroom producing more dust particles.

Comparison

Comparison of T2G5 & T2G2

After comparison with T2G5, the trends observed in the data collected by T2G2 are corroborated by the data collected by T2G5. Although, due to the differing locations in the shop floor, it can be concluded that the variance in the data readings can be attributed to there being a gradient of values for all three measurements across the entire shop floor thus resulting in similar patterns but differing values.

Comparison between Device 01 & T2G2

- **Temperature (Figure 7 & 8)**
    - There is no clear correlation between the Ambient air and shutter door surface temperature recorded by Device01 and the temperature recorded by our Arduino (T2G2). The temperatures recorded by Device01 are generally higher than that of ours . The difference in placement may also have contributed to this as the shutter door comes into direct contact with external temperatures.
    - While there may be no correlation between the 2 devices. There is a correlation between the brightness level and temperatures recorded by device01
    - Both brightness and temperatures increase together, suggesting that higher brightness levels, likely due to direct sunlight, contribute to the increase in temperatures recorded by Device01.

Comparison of weather data & T2G2

- **Temperature: (Figure 9)**
    - There are peaks of temperature on the weekends, which is most likely caused by the lack of indoor climate control. However, the sunny weather may play a part in the high temperatures.
    - On weekdays, we can see a trend of temperature rises and drops, likely caused by the switching off of climate control for the day every evening, thus increasing the temperature even though the weather is cloudy. And then turning it on in the morning of the indoor next day causing temperature drops even though it is sunny.
- **Humidity: (Figure 10)**
    - Like temperature, weather does seem to play a part in the room's Humidity
    - Cloudy days tend to have higher humidity than sunny and rainy days.

- **Air Quality: (Figure 11)**
    - Multiple peaks of air quality indicates, there are some days where air quality is worse than others. However, there is no correlation between the external weather as compared to the internal air quality.
    - One off instance on 13th July where it rains and the air quality improves. Rain may have caused the air quality improvements as the rain captures dust particulates and other pollutants causing the internal air to improve in air quality as the dust and pollutants inside the factory floor then diffuses externally improving air quality.

Optimisation

Using the KNN clustering machine learning algorithm, we split the data into 2 clusters as seen in (**Figure 12**). These classes represent 2 distinct environmental conditions based on the 3 parameters, temperature, humidity and air quality.

- **Yellow cluster:** Represents the optimal conditions, where the environment is cooler and the humidity is lower. This means that we should maintain conditions within this cluster to create comfortable and healthy factory conditions.
- **Blue cluster:** Represents the less desirable conditions that we want to avoid having in the shop floor. These conditions are categorised by higher temperatures and humidity levels.

Parameter distribution (Figure 13):

- Temperature:
    - Optimal conditions: The temperature values are relatively lower with a tighter Interquartile range
    - Hot & Humid conditions: The values are slightly higher with a similar interquartile range.
    - Recommendations: We should target temperatures of about 24 degree celsius (mean optimal temperature).
- Humidity:
    - Optimal conditions: While the interquartile range is lower, there are many outliers observed showing that the humidity can sometimes exceed the desired range.
    - Hot & Humid conditions: The humidity values are generally higher with a larger interquartile range.
    - Recommendations: As per the collected data, humidity of around 64% should be maintained, but this is still relatively high according to healthy humidity standards (~30 - 50%).

- Air quality:
    - Optimal conditions: There are many positive and negative outliers observed suggesting that air quality levels fluctuate under optimal conditions.
    - Hot & Humid conditions: The Air quality levels are generally better with a tighter interquartile range but show many outliers.
    - Recommendations: We should target air quality levels of around 42; the lower the better.

Another deep learning algorithm, Recurrent Neural Networks (RNN), has been used to detect anomalies for temperature and humidity. It is a two way process, where we first trained an RNN to predict the temperatures and then compared the predicted trend with the original trend. If there is a huge difference between the two over some period, the parameter within that period is said to deviate from what is normal.

Temperature prediction (Figure 14):

- The predicted and actual trends of the temperature readings over the two weeks are closely matched with one lag behind, meaning that the temperature, i.e., the air conditioning, is behaving as expected.

Humidity prediction (Figure 15):

- On the other hand, the predicted humidity trend is way lower than the actual trend, indicating the anomaly and that the humidity controller (if any) is not working as expected; a dehumidifier is required.

Recommendations

- **Dehumidifiers for Humidity Control:**
    - install dehumidifiers to manage the higher humidity levels observed at the beginning and end of the week. By doing so, we can reduce humidity levels, especially on Mondays when internal conditions are stabilising after the weekend, and on Sundays when humidity tends to vary.
- **Air conditioners:**
    - Currently, the air conditioner is used on weekdays until around 7- 8pm and turned off on the weekends. So, instead of this we can keep the air con on and run it at a lower capacity. This will help maintain our baseline temperature preventing large fluctuations when the system is turned back on during weekdays
- **Usage of air filtration units**
    - There are moments in the day, likely caused by usage of the machines, where air quality is highly polluted. As such, to mitigate the air pollution

generated by the usage of the machinery, usage of air filtration units where the air near the machines when they are in use are filtered to ensure dust particles are filtered out.

- **Coat the shutter door with thermo-resistant polymer**
    - The shutter door is metallic and has shown to affect the data we have collected. As such, by coating the inside and outside with a UV and heat resistant plastic layer will reduce the conductivity of the shutter door further insulating the factory floor, helping to reduce the spikes in temperature we have recorded on our device t2g2.

6. Conclusion

From this project, we have deduced that there are clear patterns in the indoor climate conditions of the factory floor which can be traced to either external weather conditions, factory floor foot traffic or the switching on of machinery and or the climate control system. As such, we have created a dashboard to help users to view the current climate conditions as well as historical climate data of the factory floor. With this information alongside with the findings of our KNN clustering algorithm, which has found the optimal factory floor conditions, users will be able to in real-time be able to set the climate to the optimal conditions via the various recommendations we have made. This will allow for better safety as air quality and temperature are controlled, efficiency as climate control is turned on as required saving energy and lastly, ensuring the machinery lasts longer as they are not exposed to extreme conditions which may damage them. One key limitation to note is that through the analysis between t2g5 and t2g2 we have concluded that conditions around the factory floor are susceptible to large variances during the same period of time. As such, it is also key that in future research, we have multiple Arduinos set up in various locations on the factory floor so that we may collect data points for all areas of the factory floor and then map the conditions to each section of the factory floor. This will help in more fine control over the climate in the factory floor, further improving upon our current system.

7. Security

Security plays a big role in any data-pipeline, and Envi-Optimizer is no exception. For secured transmission and storage of data in this project,

- The server identifies the correct topic (t2g2) before storing the data.
- A username and password is required to retrieve data from MySQL database.
- The database is locally hosted and is accessible only within a small area, thereby protecting against miscellaneous attacks from the internet.

- The MQTT message is published without guidance, i.e., an onlooker could not know at first glance which numbers represent which parameters in “t2g2:114,12.34;56.78;90;Min;”.

A higher security feature that could have been easily implemented would be to encrypt the reading through some mathematical operations instead of publishing the direct readings.

8. Enhancement

ODBC MySQL Driver

The use of the ODBC driver allowed us to update our dashboard with the most up-to-date data at the click of a button, provided that the computer is connected to the server in the area. This helped the stakeholders on-site to directly view the updated reading in a visually pleasing Power BI dashboard. It also eliminated the need to run a python script to download and import the data manually as well as splitting the message and cleaning it each time, thereby streamlining the data retrieval process.

3D Printed Enclosure

A 3d printed enclosure was designed to help keep the arduino and the shield safe from knocks and scratches. Furthermore, for further expansion, there are rails on the top of the enclosure which can allow for clip-on brackets which can allow for the secure mounting of the sensors. In the future, this will allow for the ease of mounting the Arduino in all sorts of locations in the factory floor allowing the collection of data in all areas of the factory. In addition, due to the modular nature and the ease of access a person has to the sensors, if and when they fail, they can be easily swapped out whilst ensuring the sensors are connected to the correct ports on the shield. All in all, the enclosure makes the Arduino safe to deploy anywhere, allows for easy maintenance and lastly ensures the Arduino is easy to identify in the midst of all the machinery in the shop floor.

9. References

Weather Data - *Weather in July 2024 in Ang Mo Kio New Town, Singapore. (July 2024).* *https://www.timeanddate.com/weather/@1884365/historic?month=7&year=2024*

Enclosure Inspiration - *Arduino Uno Enclosure | Autodesk Community Gallery. (n.d.). Autodesk Community Gallery.* *https://www.autodesk.com/community/gallery/project/21486/arduino-uno-enclosure*

10. Appendices

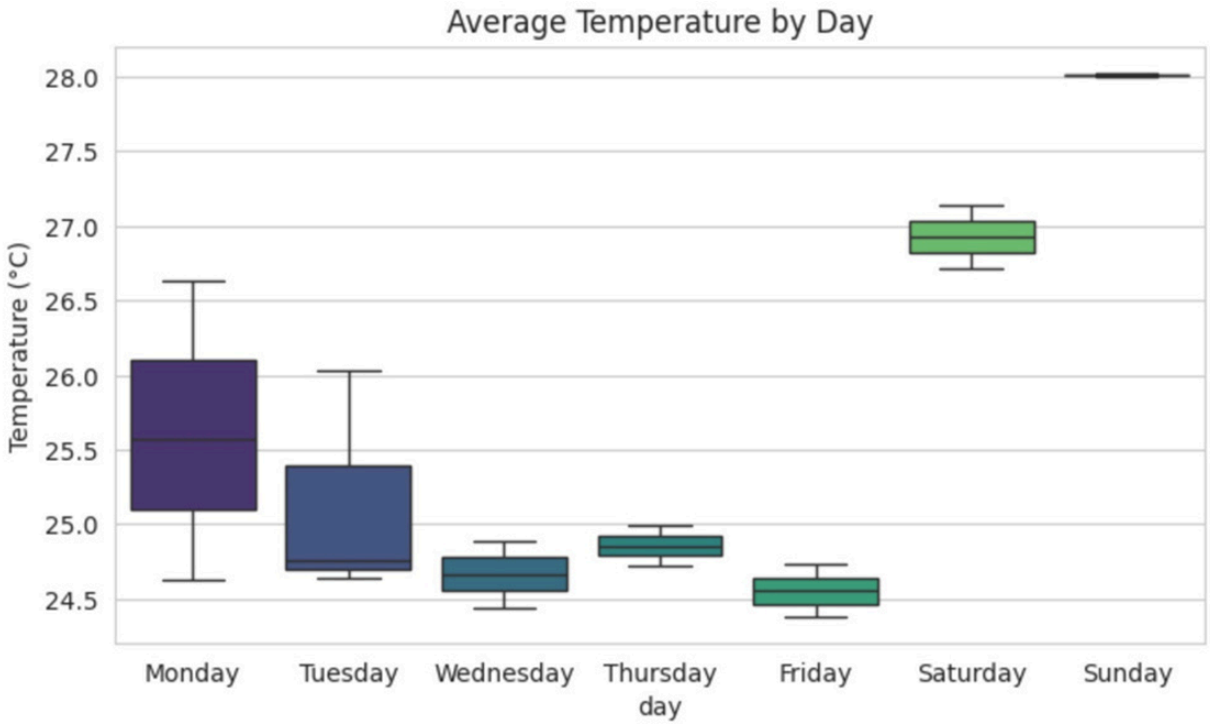


Figure 1 - Temperature variance for each day

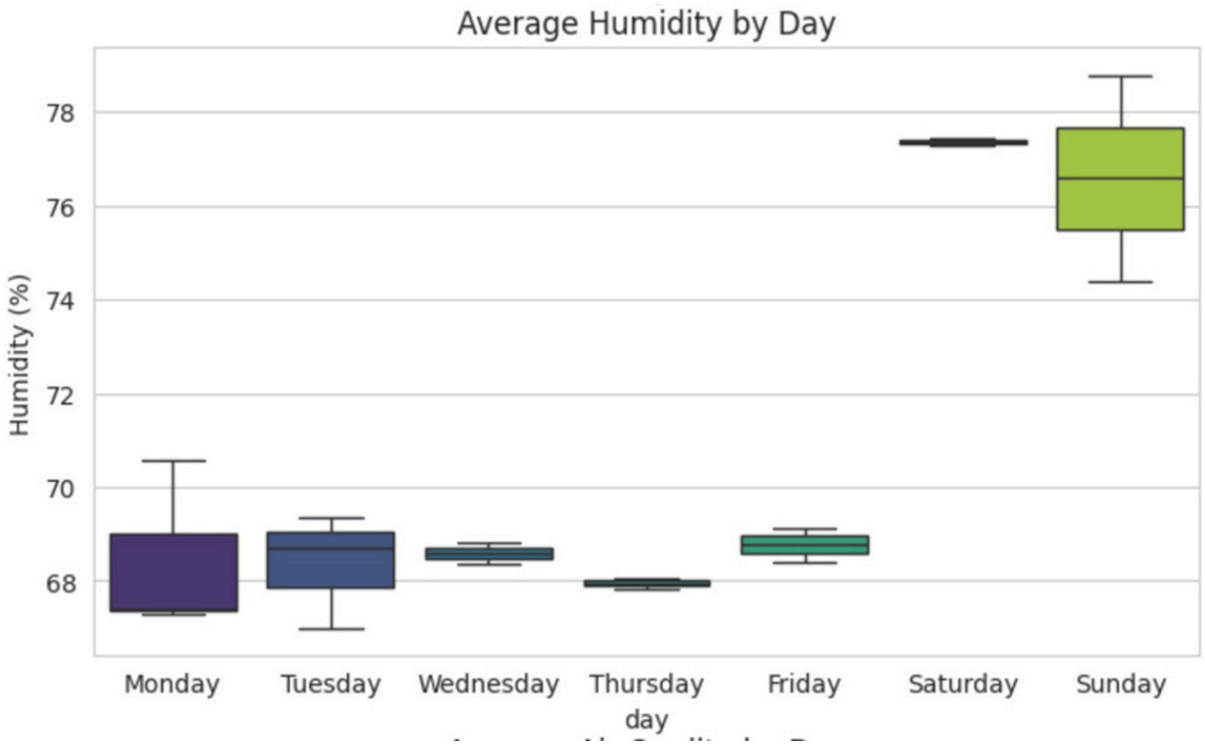


Figure 2 - Humidity variance for each day

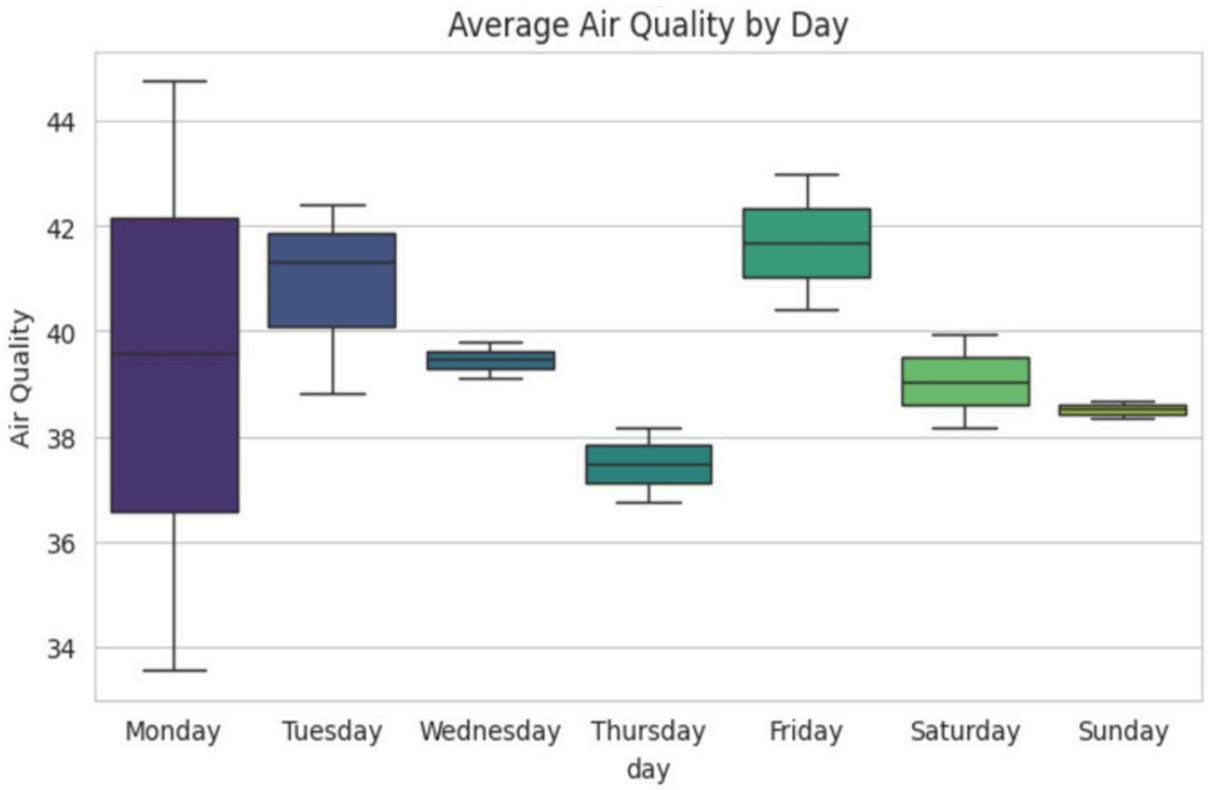


Figure 3 - Air quality variance for each day

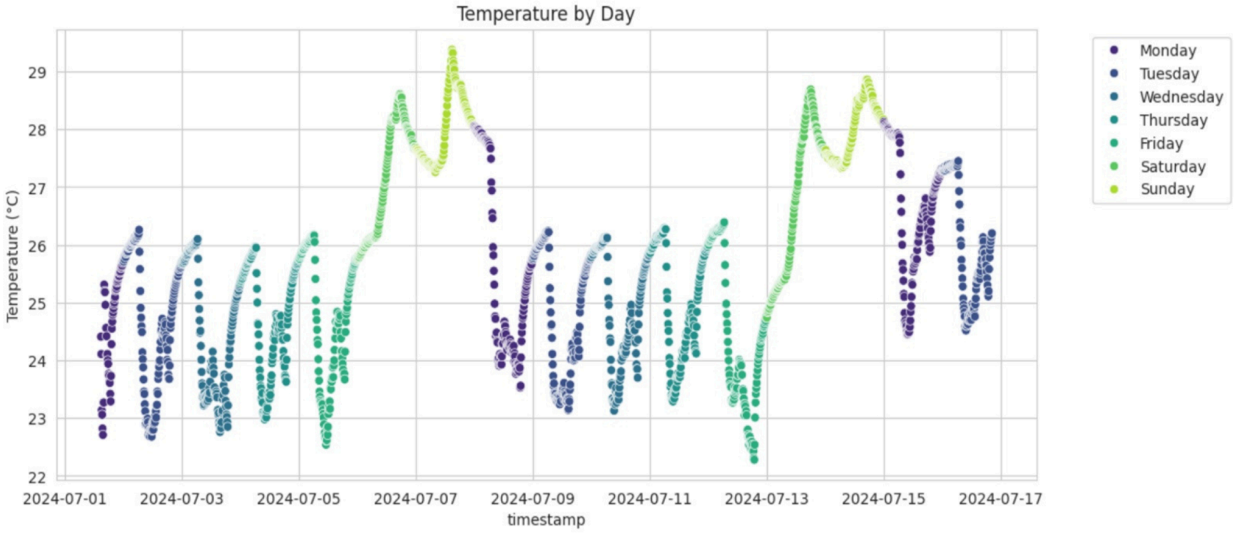


Figure 4 - Temperature trend by date

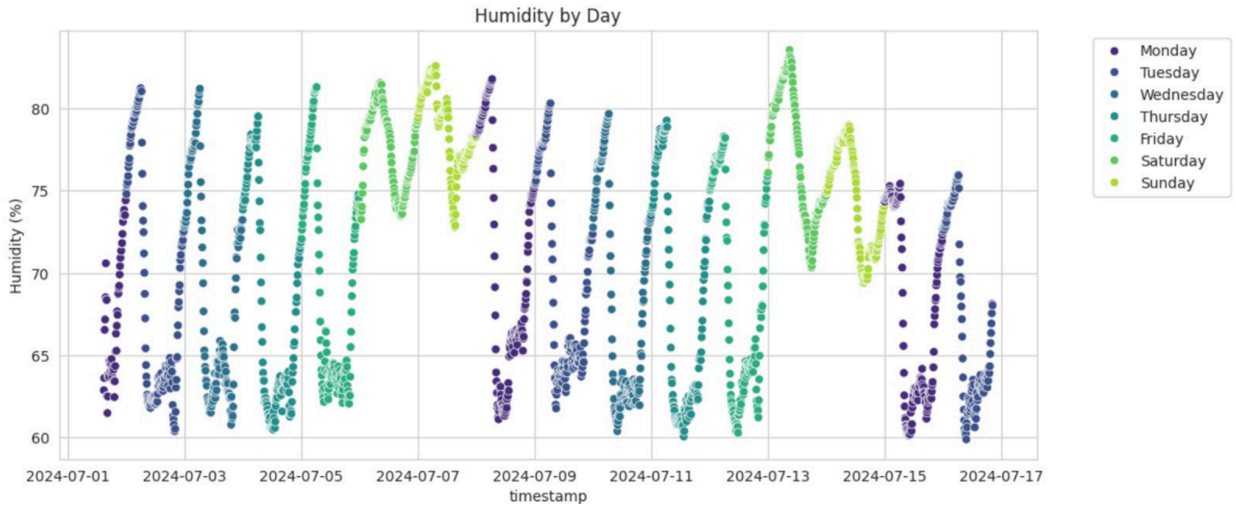


Figure 5 - Humidity trend by date

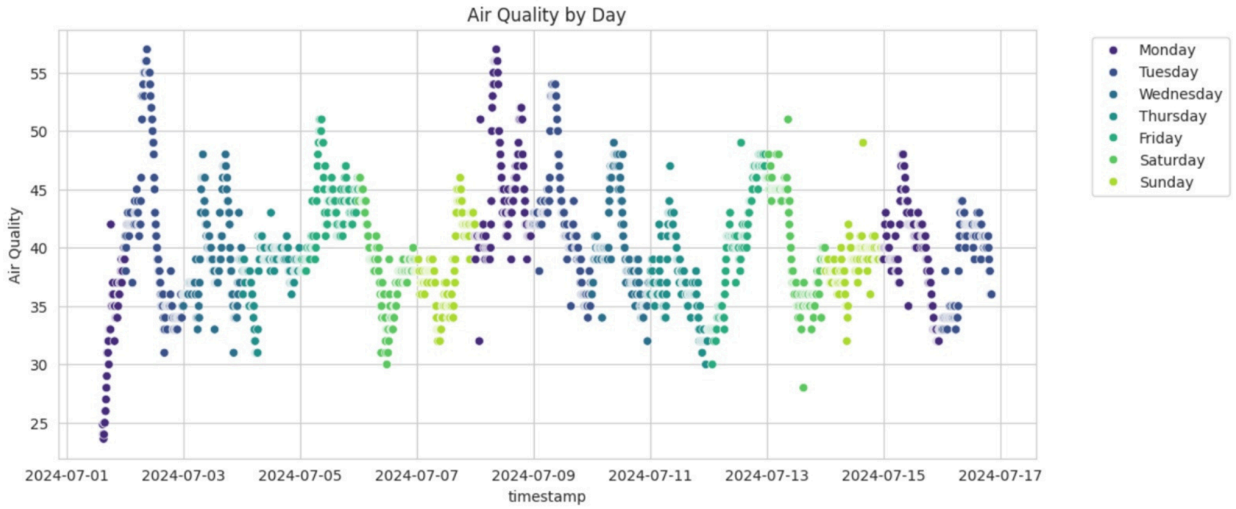


Figure 6 - Air quality trend by date

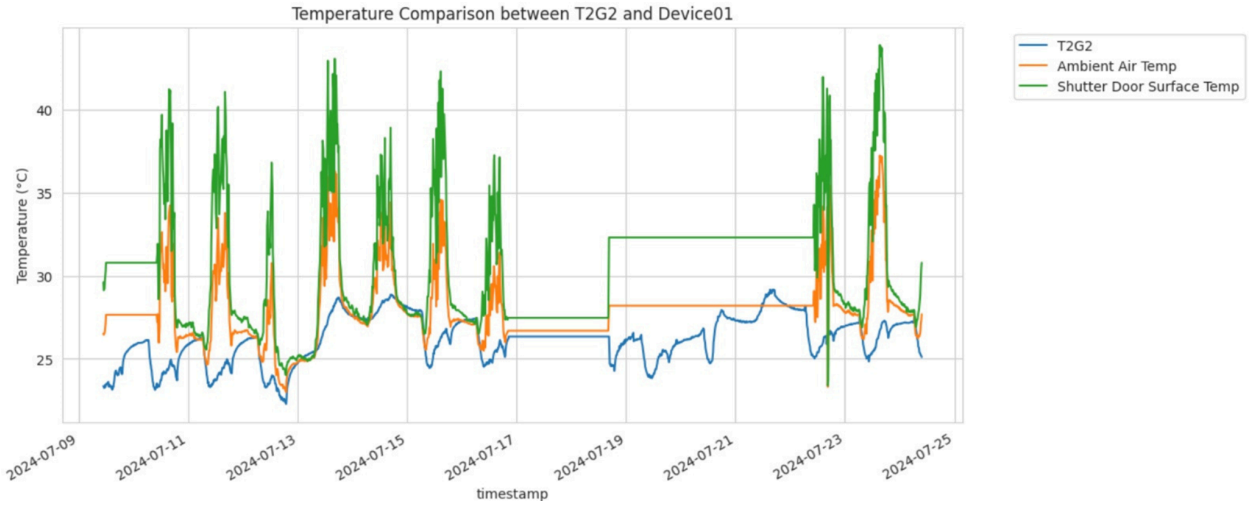


Figure 7 - T2g2 and Device01 Temperature

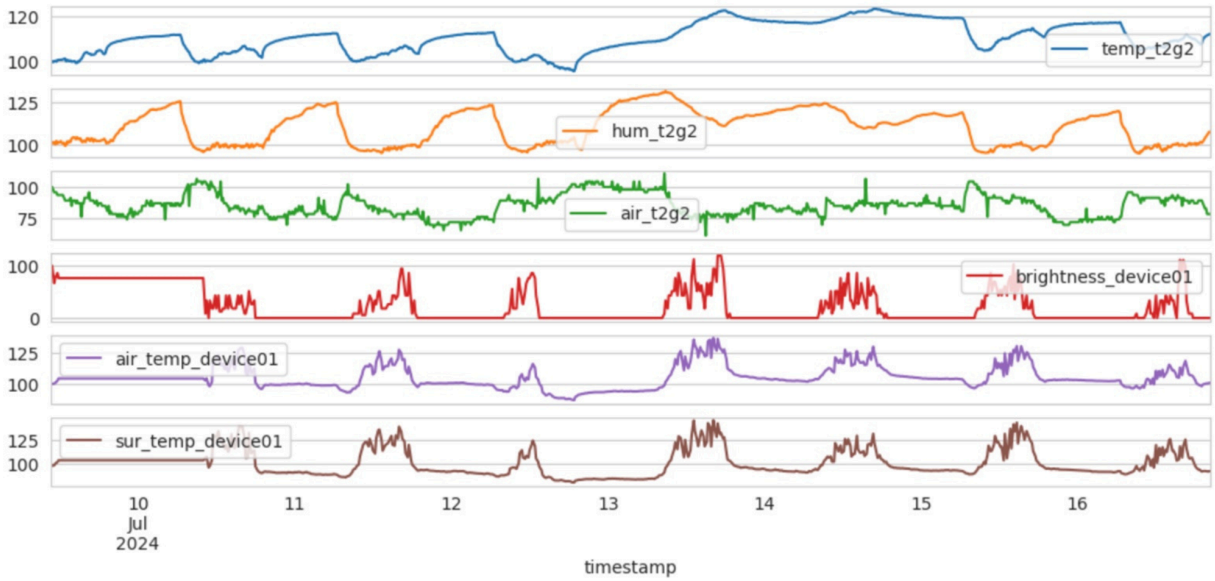


Figure 8 - Individual graphs for T2G2 and Device01

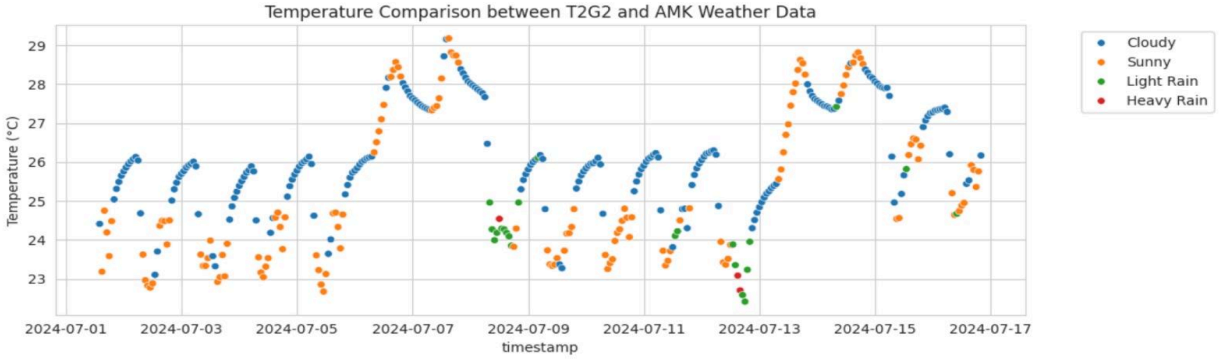


Figure 9 - Weather conditions & Temperature

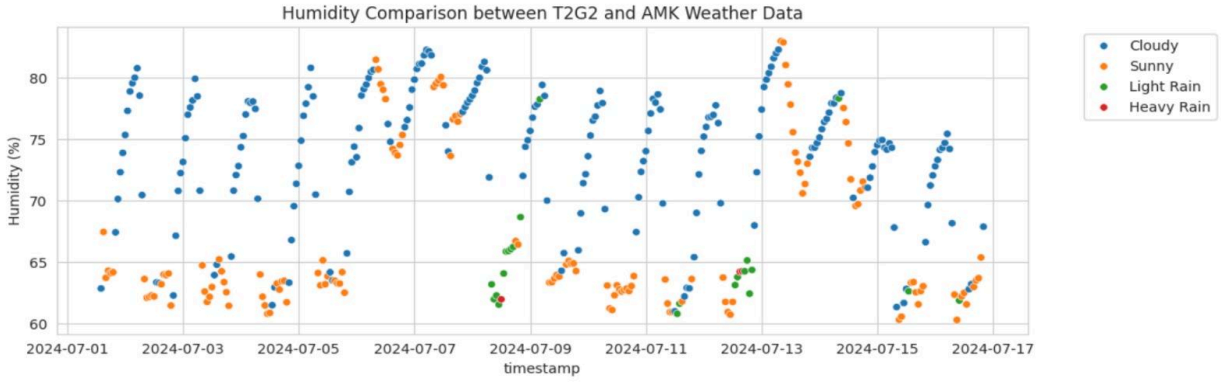


Figure 10 - Weather conditions & Humidity

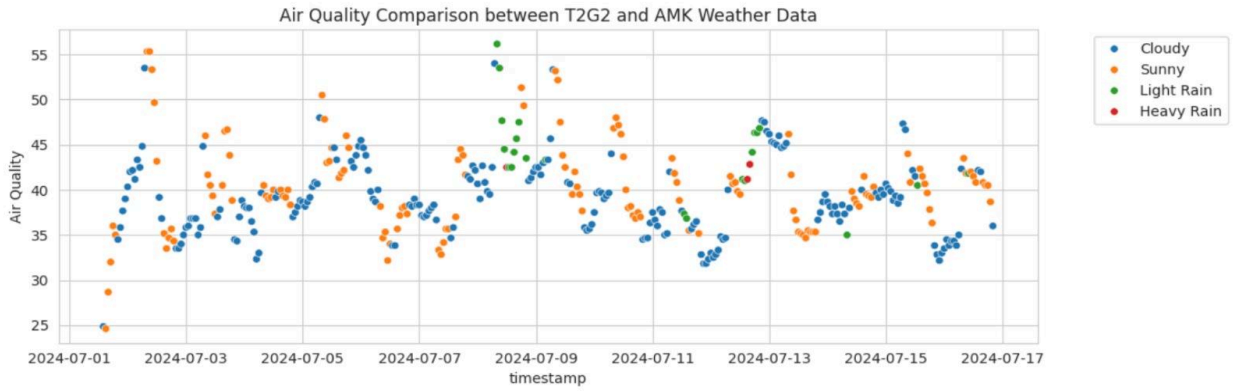


Figure 11 - Weather conditions & Air quality

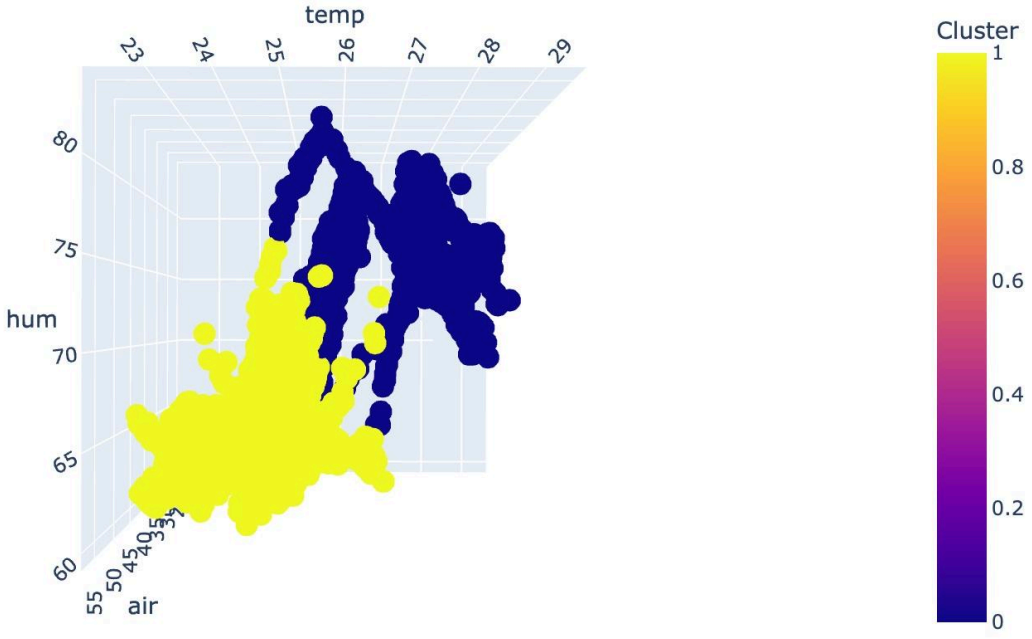


Figure 12 - KNN Clustering

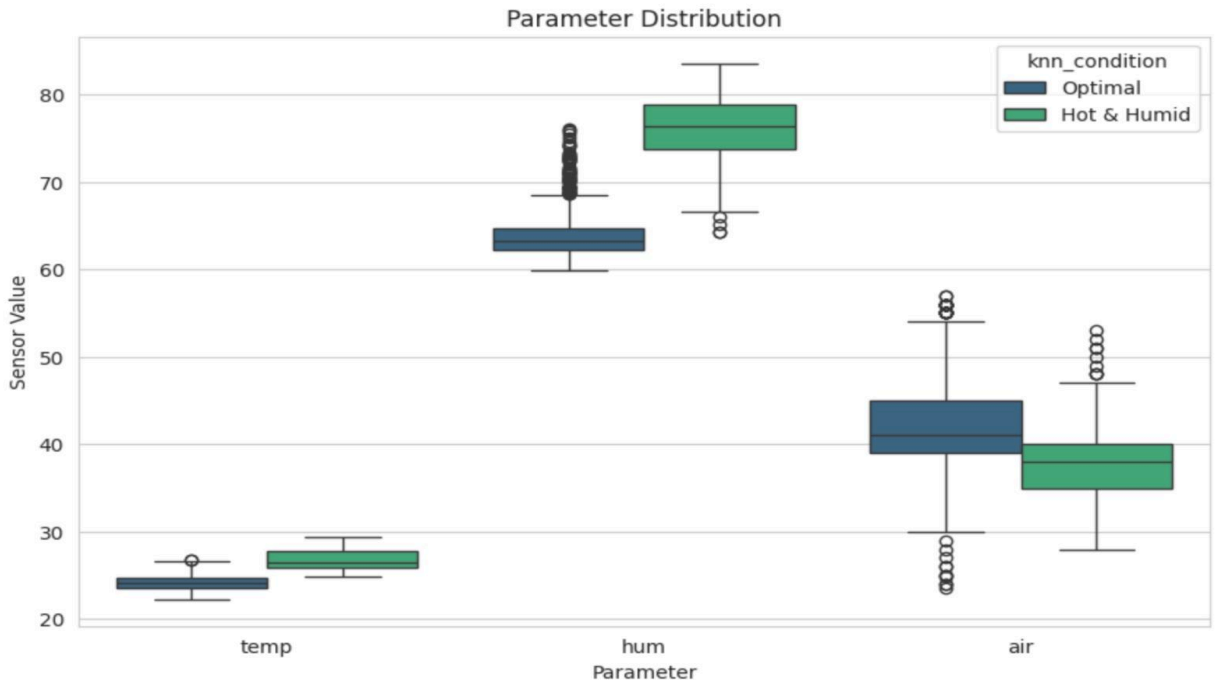


Figure 13 - Parameter Distribution

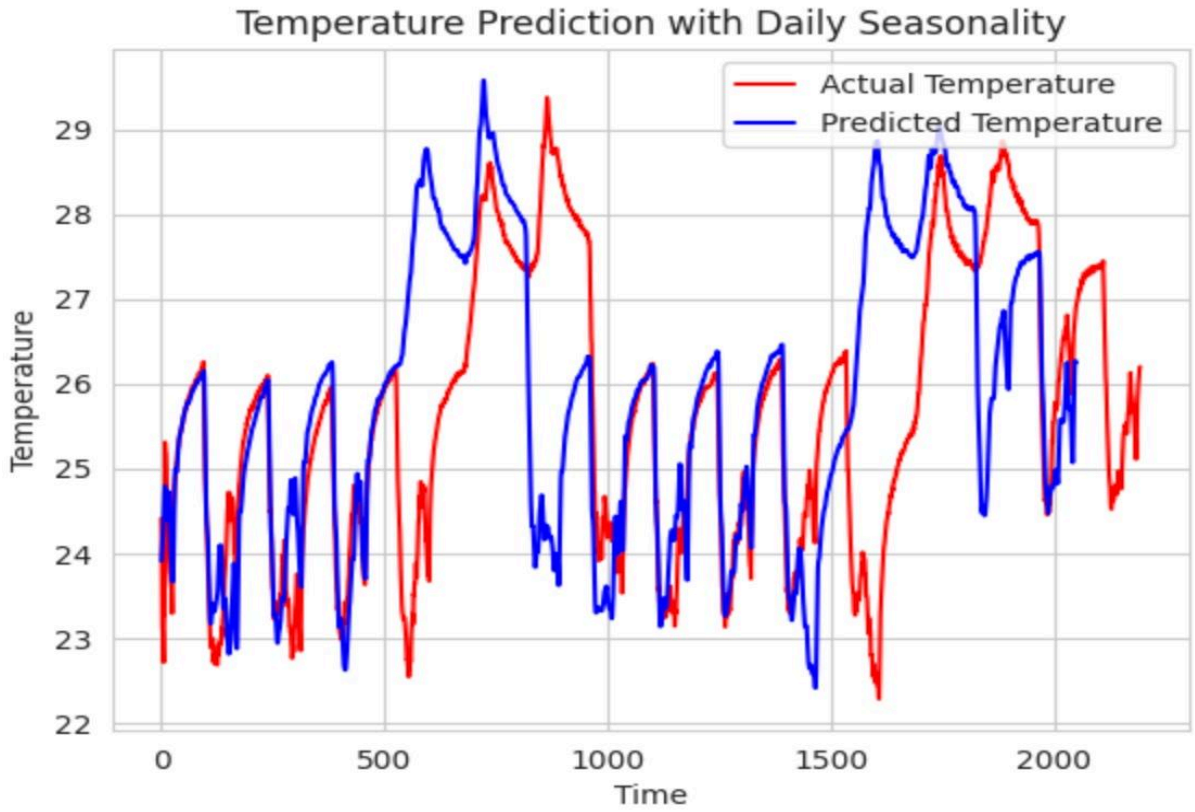


Figure 14 - Temperature prediction

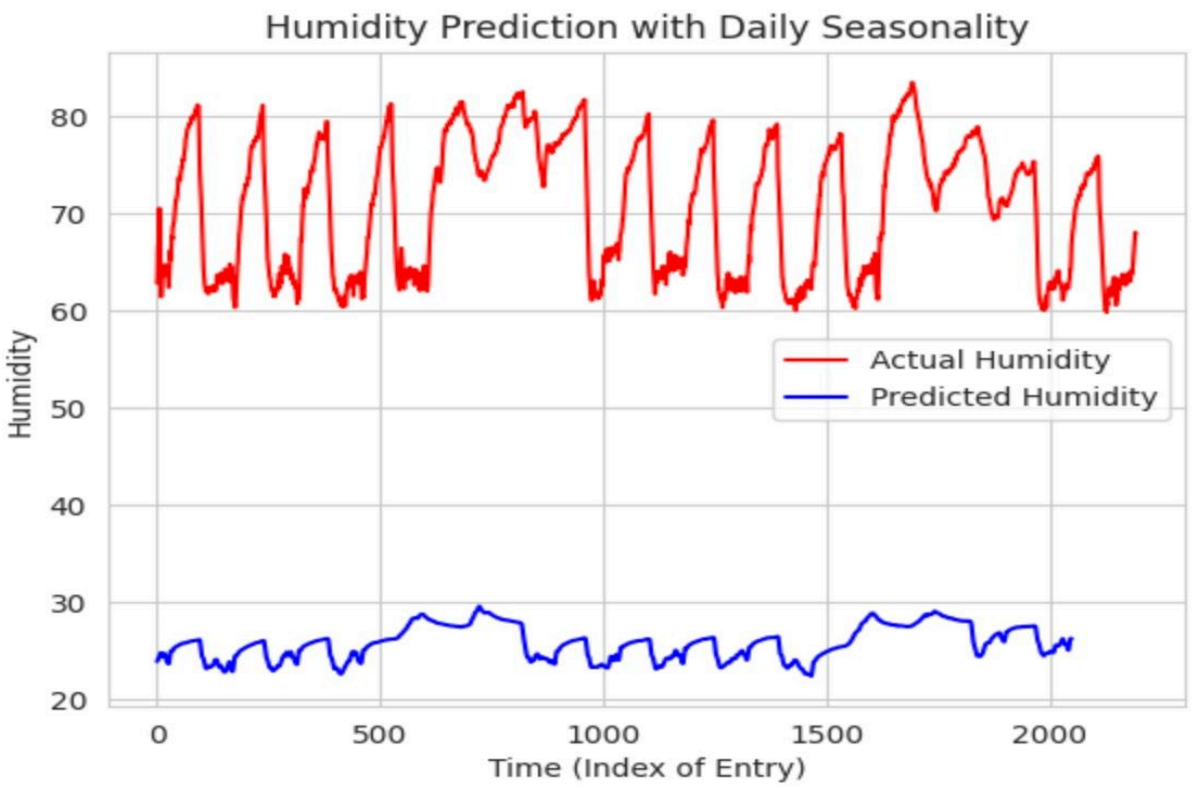


Figure 15 - Humidity Prediction

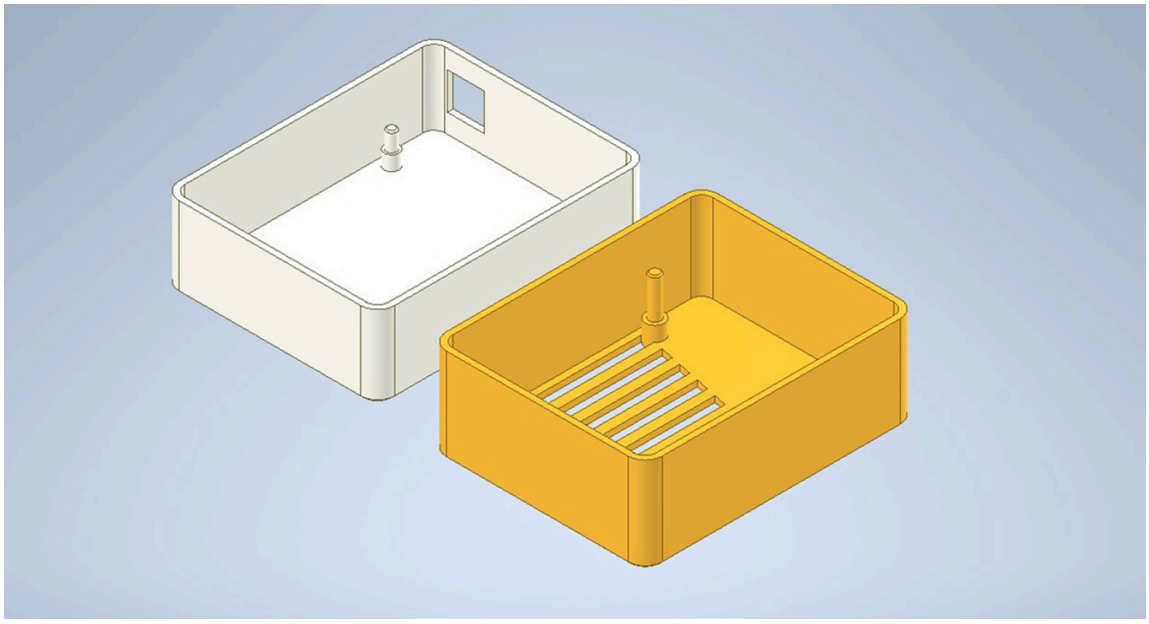


Figure 16 - 3D modelling of prototype (internal)

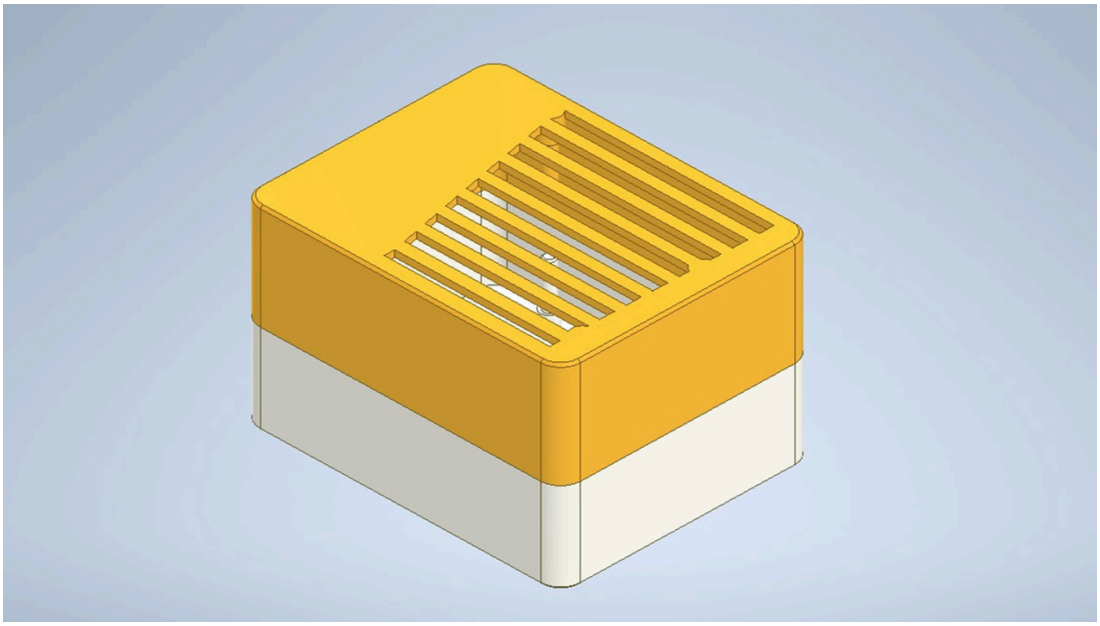


Figure 17 - 3D modelling of prototype (external)

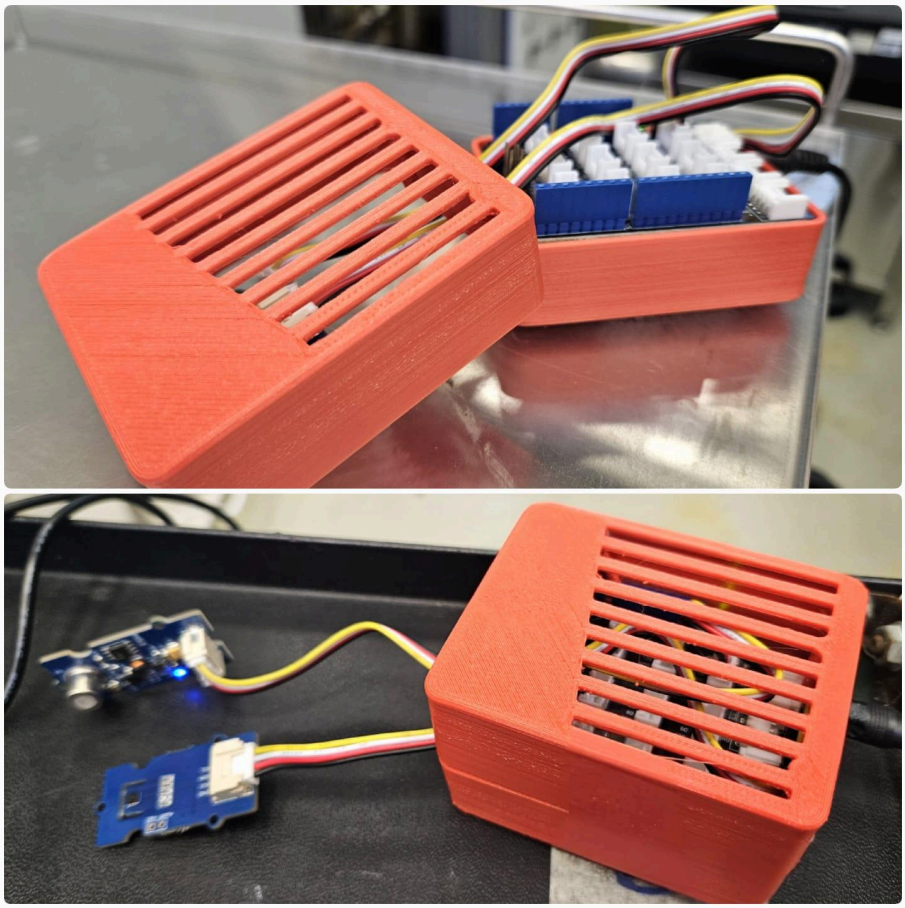


Figure 18 - Image of deployed Enclosure