AWAIR FINAL DESIGN REVIEW REPORT

Vyomesh Iyengar, Inesh Jacob, André Larocque, Swetaa Suresh, Rodrigo Tiscareno, Lauren Yoshida
Team 2

Abstract – *Companies in the HVAC sector lack efficient remote detection of system abnormalities, relying on customer complaints, leading to inefficiencies and unprepared technicians. Alert Labs, a smart solutions manufacturer, seeks to integrate advanced analytics with proprietary sensor technology for predictive maintenance and improved diagnostics. This project involves anomaly detection, predictive maintenance, and the development of a user-friendly interface for technicians. By leveraging IoT, machine learning, and cloud computing, the aim is to enhance HVAC service quality and operational efficiency.*

Keywords: HVAC, anomaly detection, predictive maintenance, machine learning, statistical analysis, thermodynamics

1. PROBLEM ANALYSIS

HVAC companies (heating, ventilation, and air conditioning) systems lack ways to remotely detect anomalies in their customers' systems, relying on the reactive process of customer complaints to identify malfunctions. Since homeowners typically do not have an understanding of their HVAC unit's issues, technicians are dispatched without any context. This impacts HVAC service providers as they cannot prepare technicians with knowledge or equipment to provide service. This also impacts homeowners detrimentally as they must communicate issues of which they have no understanding. These problems present an opportunity to improve the efficiency and effectiveness of HVAC services.

1.1. Industry Partner

Alert Labs is a manufacturing company that builds solutions for water monitoring and HVAC systems. Along with its line of products, they receive data from over 74 sensors corresponding to indicators in HVAC systems produced by a third-party manufacturer. They are seeking a method to identify anomalies in critical data points that reflect the overall health of an HVAC unit. Additionally, they aim to perform predictive maintenance to determine optimal servicing times, ensuring the longevity of HVAC products.

1.2. Market Analysis

Companies like 75F and Monnit adopt differing approaches to data analytics, as seen in *Figure 1*. 75F specializes in intelligent automation systems for the HVAC industry. They use external sensors which limit their ability to collect some data, but integrate seamlessly with existing systems.

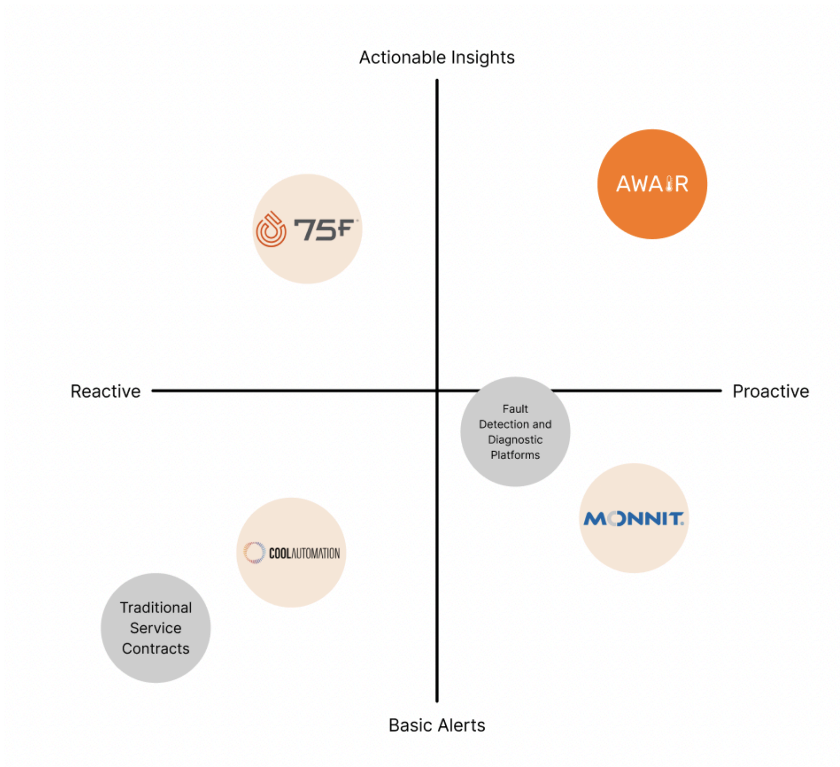


Figure 1. Competitor comparison of platforms.

Similarly, Alert Labs' current model, Sentree, has units placed alongside HVAC systems to gather data. While this method provides substantial data, there are challenges in accurately diagnosing errors. The industry has a need for solutions that collect HVAC data internally, providing in-depth analysis and precise diagnostics for technicians. Addressing this, our solution integrates advanced analytics with proprietary sensor technology embedded within HVAC units. This ensures higher accuracy and reduces reliance on external sensors. Algorithms and statistics-based techniques analyze historical data to forecast when HVAC systems will require maintenance, reducing downtime and preventing failures. Machine learning (ML) models dynamically adjust HVAC settings to optimize energy consumption while maintaining comfort levels.

1.3. Problem Statement

The team aims to solve the problem of detecting and diagnosing operational anomalies in HVAC systems, predicting their impact on energy efficiency, and improving communication among stakeholders to create a reliable system that allows for better use of resources. *Figure 2* provides a representation of how our solution differs from the status quo.

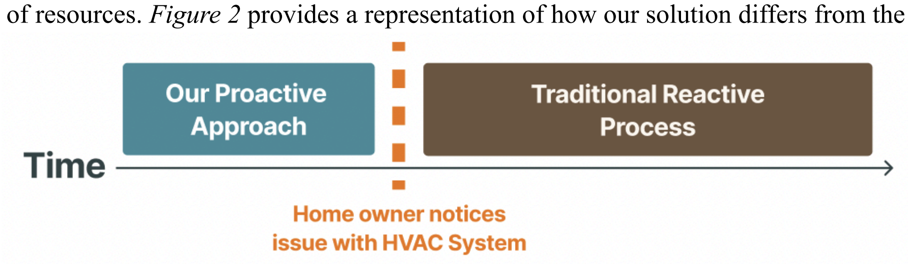


Figure 2: Time comparison of a proactive vs reactive HVAC service.

2. Requirements Analysis

The requirements-gathering process was split into two: communication with users to understand pain points, and exploring our provided dataset to discover reliable correlations and calculations.

2.1. User Requirements

One user interview was conducted with an HVAC technician. The interview's goals, script, and questions are listed in *Appendix A*. Based on this interview, an affinity map was created to plot the core insights in *Appendix B*.

The affinity map determined key areas to address: troubleshooting, monitoring, and solving issues. While installations were important to their role, they did not have dependencies on metrics or data as the process was mostly standardized. It was observed that the prioritization of problems depends on their impact on customer safety, and whether customers have a pre-existing service contract. For issue diagnosis, technicians use nameplates found on HVAC units to understand measured values expected ranges to compare against observed values.

Drawing from information provided in the interview and prior conversations with Alert Labs, three core user groups were identified: HVAC technicians, HVAC service providers, and homeowners. To learn their individual pain points, design considerations, and to contextualize their work, a user persona was created for each group (*Appendix C*).

Based on this, the core problem for each user group was determined:

- *Technicians:* There is a lack of data before entering homes, causing a lack of preparation. Solving this problem would help in diagnosing issues more efficiently.
- *Service providers:* Due to not knowing the severity of customer complaints, the seniority of HVAC technicians cannot be leveraged to solve more complex issues. Solving this problem would allow for better allocation of resources.
- *Homeowners:* Since HVAC units require knowledge of technical components, they cannot diagnose problems or identify inefficiencies. Solving this would allow for cost reduction and environmental sustainability.

To understand how this solution would integrate into the workflow of an issue diagnosis and resolution, the current process was mapped alongside the envisioned new process in *Figure 3*. This allowed for a better understanding of the core flow of an HVAC technician, the areas of interaction between user groups, and the pain points faced.

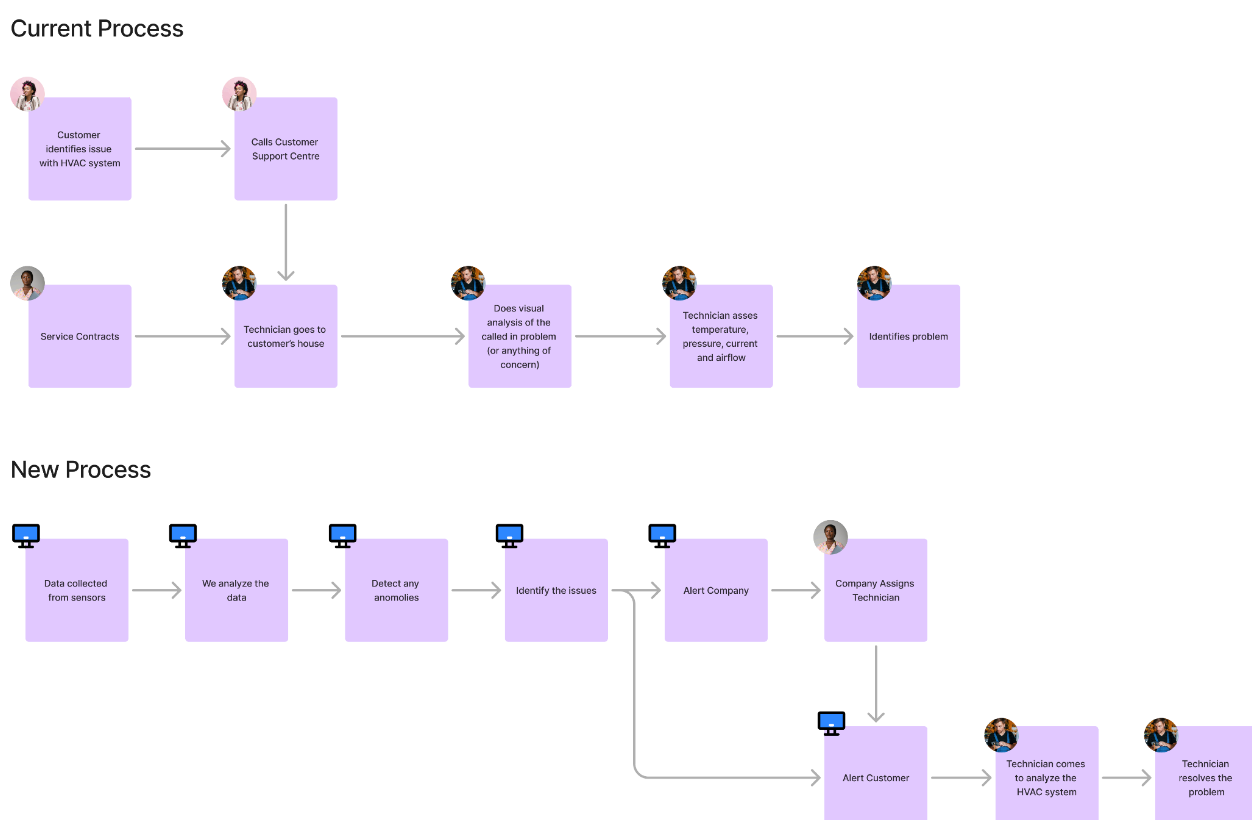


Figure 3. Current process, the stakeholders involved, and the changes anticipated with the solution.

Based on the proposed changes, customers do not need to do any identifying work as the system can do it. It also only requires the HVAC technician much later in the process, removing parts of the work highlighted as pain points. *Table 1* describes the product requirements found based on the user interview and insights organized.

Table 1: Product requirements derived from user interviews and analysis.

Requirement	**Associated Stakeholder**	**Reasoning**	**Description**
Anomaly Visualization	Technician	Visual of anomaly location to diagnose issue	Visualize anomalies in the HVAC system in a diagram.
Anomaly Analysis	Technician	Allows the technician to view the data points for investigation	Displays the data points that lead to anomalies.
Current HVAC Health	Operations Team	Displays inefficiencies that need to be checked up on and insights on issues.	Allow the operations team to view the system's behaviour. If irregularities arise, it will show the location and provide the predicted issue.

2.2. Data Requirements

Predictive analytics requires high-quality and high-volume data to establish parameters for anomaly detection. Alert Labs provided data from their partnered HVAC system manufacturer who have their own sensors. The data consists of over 600 million log records detailing changes to thermodynamic attributes and system modifications made by customers. There are 86 data attributes describing the state of the system at any point in time. These attributes define the different modes and configurations of the system, the environment and ambient values such as outside temperature and humidity, and the pressure and temperature at each stage of the refrigerant cycle.

The NoSQL design of the provided database meant that the irregular schema and structure of the source data were not optimized for large queries. An internal data dictionary, displayed in *Figure 4*, was created to classify the actualized attributes observed in the received data.

Type	Column Name	Attribute	Units	Notes	Data Type	Category	Source	Description	Questions	Type
Metadata	timestamp	timestamp	GMT(UTC+0)	Unknown what this represents in this context.	timestamp	Admin	Metadata			Captured
Trivial	CI_COMPN	company_name		Generic column for company name. Looks like redacted values.	str	Admin	Metadata			Captured
Metadata	arrivalTimestamp	arrival_timestamp	GMT(UTC+0)		timestamp	Admin	Metadata			Clarify
Trivial	CI_PHONE	phone_number		Generic column for phone numbers. Looks like redacted values.	str	Admin	Metadata			Captured
Trivial	CI__NAME	contractor_name		Generic column for contractor names. Looks like redacted values.	str	Admin	Metadata			Captured
Trivial	CI_EMAIL	contractor_email		Generic column for contractor emails. Looks like redacted values.	str	Admin	Metadata			Captured
Trivial	EQP_NAME	equipment_name		Measuring device.	str	Admin	Metadata			Captured
	DISPUNIT				float					Missing
Metadata	reportedTimestamp	reported_timestamp	GMT(UTC+0)	Reportedly the time that the data was passed to the cloud.	timestamp	Admin	Metadata			Clarify
	VACAEMON				float64					Missing
Owner Settings	VACSTATE	vacation_active		{"No"=0, "Yes"=1}	bool	Autopoll	Owner		What is vacation mode? Does it have any implications?	Clarify
	VACASTIM									Missing
	VACASMON									Missing
	VACAEYOC									Missing

Figure 4. Demonstrates the capture of some of the attributes in the data.

To enhance understanding of the data, single- and multi-variable data visualizations were created to visualize the distribution, ranges, and outliers within the data. Refer to *Appendix D* for a detailed overview of the data exploration portion of the analysis.

2.3. Physics Requirements

The physics component of our project aims to calculate the Coefficient of Performance (COP) which requires quantifying both the input power and the heat transfer of the HVAC system. The input power is directly available as a data column, but heat transfer needs to be calculated.

Heat transfer can be determined through either the air or the refrigerant. For the refrigerant, the energy differential is computed from the enthalpy and the mass flow rate. However, the provided data was insufficient to calculate mass flow rate, so we could not calculate the heat transfer via refrigerant analysis. Therefore, we derived the heat transfer of the air instead, using input/output temperatures and fan speeds.

Thus, our COP calculation requires the input power, supply and return air temperatures, and CFM of air to evaluate the overall efficiency of the HVAC unit.

3. Solution Development

3.1. System Diagram and Scope Creep

To ensure that the system processed the unlabelled data and allowed for both data cleaning and data processing, the team leveraged a variety of ensemble methods.

3.1.1. Data Processing

After ingesting the raw telemetry data from the HVAC systems, the team addressed missing column values by applying predefined thresholds for each data type. This was necessary because certain columns only recorded changes when a value was updated. If a value remained constant across rows, it was omitted from subsequent rows unless explicitly refreshed. However, this value could not be assumed indefinitely. For example, suction pressure values persist only for up to 600 seconds. If no new data is received within/after that time, the value is no longer considered valid, and the corresponding field in the row remains empty until a new datapoint is received. Once the forward filling is complete and the rows are filled, the telemetry data is separated into sensor data, metadata, and measured values. To better categorize the systems, the rows are also joined with external weather data based on the HVAC's nearest large city and the time that the telemetry recording was taken. Based on this data cleaning, it became possible to categorize HVAC systems by tonnage, SEER, mode, weather conditions, and wet-bulb temperatures, accomplished by Mode Classification described in *Section 3.2.2*.

3.1.2. Data Analysis

The data analysis was conducted in a three-part process, as seen in *Figure 5* and *Figure 6*. The first part was a statistical analysis, where a system was classified into a specific HVAC mode, and the values from each row were compared to the mean and standard deviation of that mode in order to identify outliers. This process is described in more detail in *Section 3.2*. The second part was to run the values through an isolation forest algorithm which detected potential anomalies across each component type, and is described in more detail in *Section 3.3*. The process of determining whether a datapoint is flagged as an anomaly or not based on the approach used to determine it is shown in *Figure 5*. Thirdly, the physics model of the HVAC system is used to calculate the heat transfer and amount of power used, which determines the system's efficiency as denoted by the COP. This is further described in *Section 3.4*. If the predicted anomaly negatively impacts the COP value this validates its anomalous label. Based on this categorization, the anomalies are passed onto the dashboard and labelled for technicians and service providers to view.

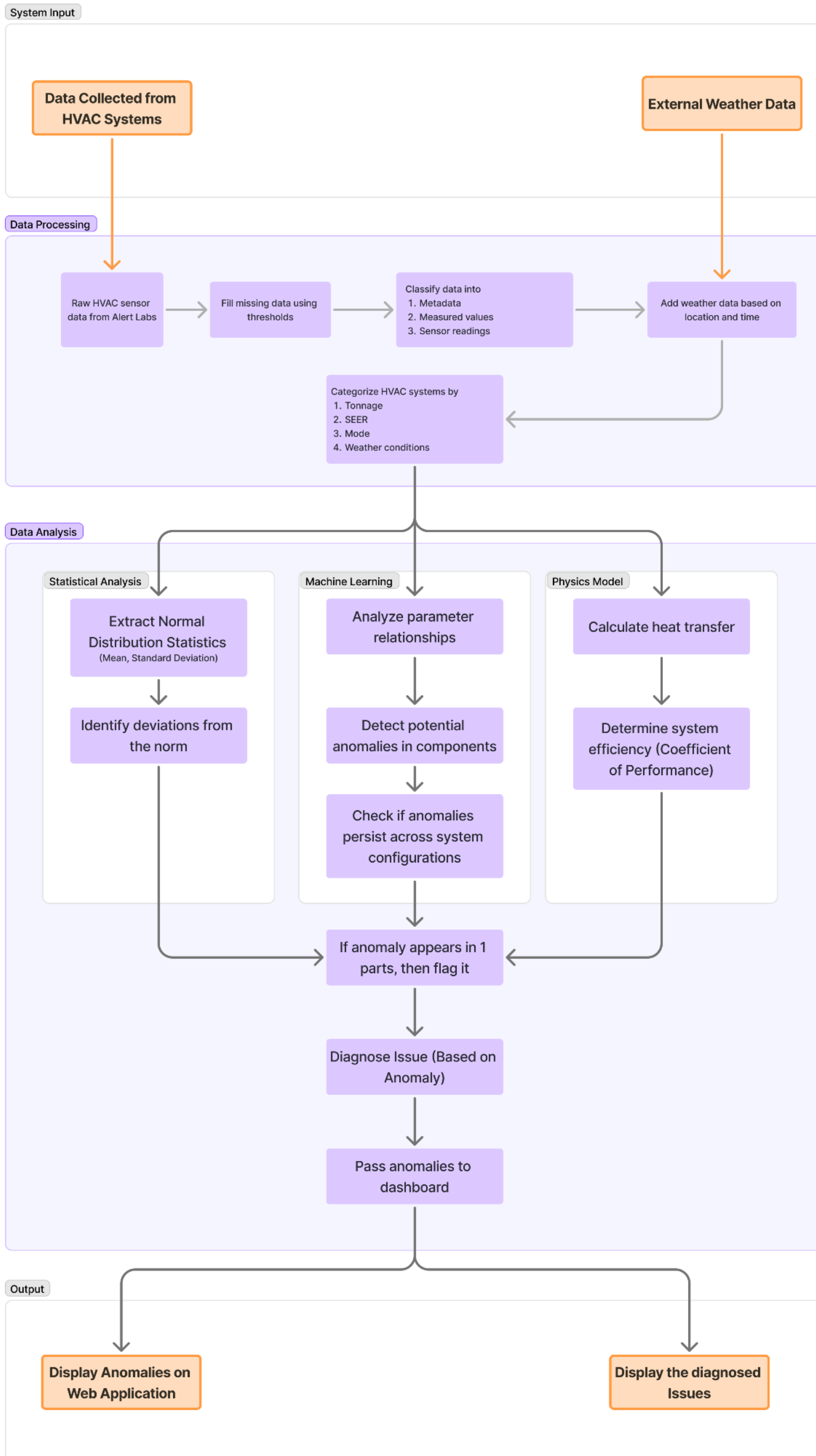


Figure 5. A process diagram outlining the input, processing, and analysis of data through the AWAIR system.

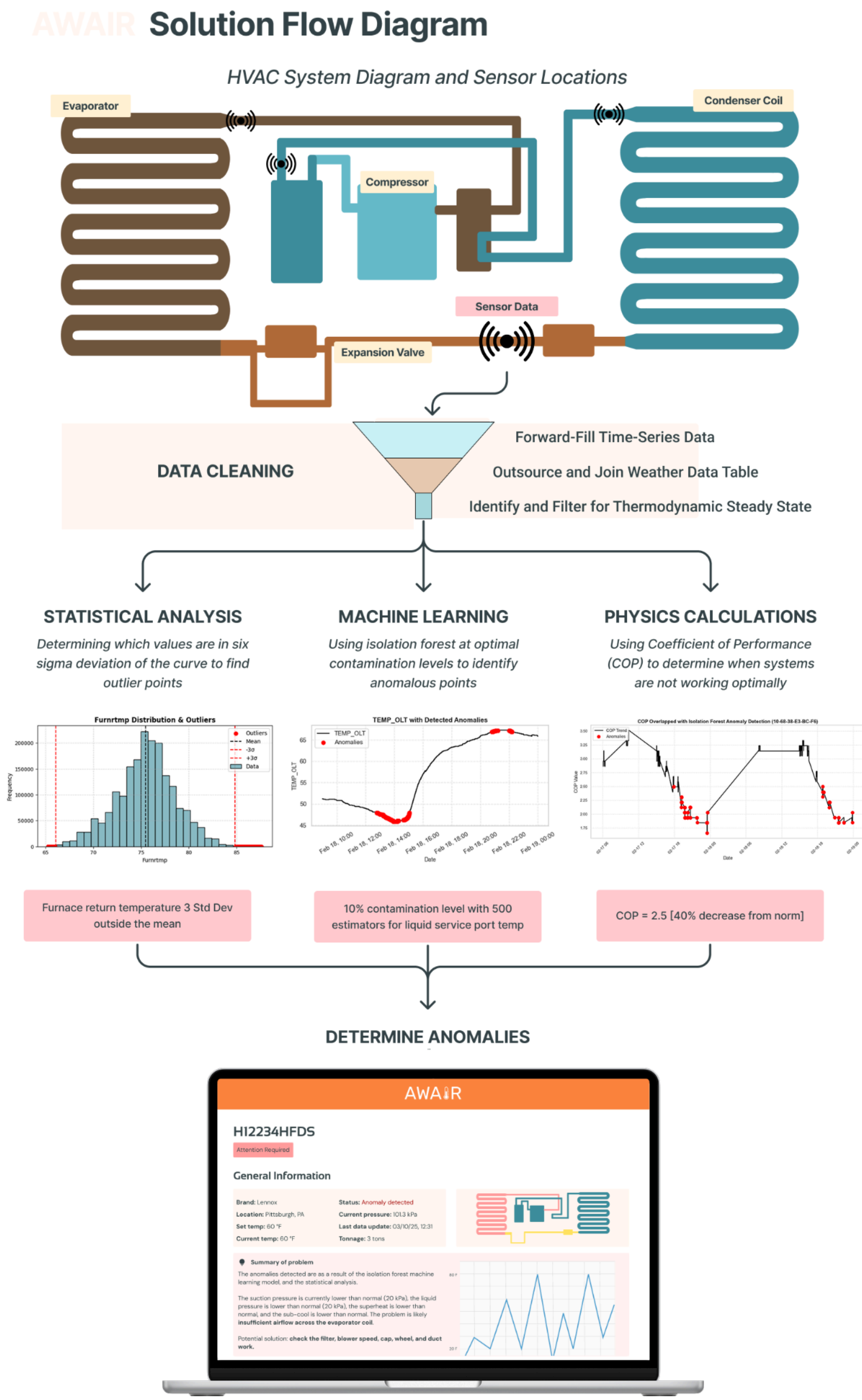


Figure 6. A general view of the data processing from sensor measurement to front-end anomaly detection display.

3.1.3. Scope Creep and Mitigation

To address key problems within the given time, we limited the project's scope. We focused on displaying essential information (system statuses, anomaly reasons, diagnostics) on the frontend dashboard, excluding features like user profiles, technician assignment, and service progress. These can be added later but are not core features. We conducted a system-level physics analysis rather than a per-component analysis since only some components had the necessary sensors. This holistic approach yielded better results. Additionally, we did not use nameplate information to determine regular value ranges, as it was not consistently available across all 279 HVAC systems, making it unsuitable for scalable analysis.

3.2. Statistical Analysis

As a primary methodology of anomaly detection, a heuristic standard deviation method was utilized. Leveraging the normal distribution of most temperature- and pressure-based metrics recorded in the HVAC telemetry data, the method involved identifying the anomalous points based on the deviation-based distance of a given observation from the mean. To execute this method, the mean and standard deviation of each feature in-scope were calculated. An HVAC system operating in heating mode in Kitchener, Ontario will output different temperature and pressure parameters than a system operating in Houston, Texas. This would mean that these two systems would be classified into different modes with a respective mean and standard deviation which would be used to determine if a given point is an outlier. As a precursor to mode classification for telemetry data, steady-state analysis and identification were essential to filter the data to a state where the HVAC system was not idle.

3.2.1. Steady State

Steady state refers to a process in time-series data when a set of telemetric variables are unchanging within a given time threshold. In the data available, steady-state was identified through the consistent nature of features that indicated the status and behaviour of the system. These features included the outdoor unit current utilization, liquid service temperature and pressures, and true suction pressure and temperatures. The superheat and subcool values, which were calculated and joined to the data, were also utilized. To determine the unchanging nature of data, a 10-observation rolling window of standard deviation was calculated and empirical thresholds were defined. If a data point did not follow the required standard deviation of its column, it would not be classified as a steady state due to the changing nature of the system.

3.2.2. Mode Classification

Mode classification is the process by which we group the time-series, telemetry data into logical and behavioural "buckets" that describe the properties of the system for a given time. *Appendix E* describes the criteria for how each data point was split into its corresponding bucket.

If all combinations were exhausted, there would be 633,600 possible buckets for each combination of attributes. Each data point in the HVAC telemetry data was classified per its bucket combination across the 279 provided systems. For each bucket identifier that the data point aligned to, there would be a running mean and standard deviation to which the new data point would aggregate to and then be stored in the MongoDB cluster.

Identifying the mean and standard deviation for each combination of property and behaviour for given telemetry data would prove effective in adding robustness to anomaly detection. By increasing the like-for-like comparison in anomaly detection, the degrees of freedom are limited to other trivial features in comparison. Mode classification was automated for 279 systems and steady-state telemetry data. The data stored in the MongoDB cluster delivered value to both Alert Labs and the statistical anomaly detection approach.

3.2.3. Executing Statistical Analysis

The statistical analysis was conducted in two stages. The mode classification for each of the 279 systems is a prerequisite for the analysis. Telemetry data from the system is filtered for steady state, enhancing the data to only include consistent, activated telemetry data. From that stage, each data point is partitioned by the six criteria defined in the mode classification.

For each set of rows that follows a unique partition, the bucket information is queried and the statistical mean and standard deviation of the in-scope features are retrieved. The values of a given data point are analyzed against the mean and 3-σ on either side of the normal distribution, as highlighted in *Figure 7* for furnace return temperature.

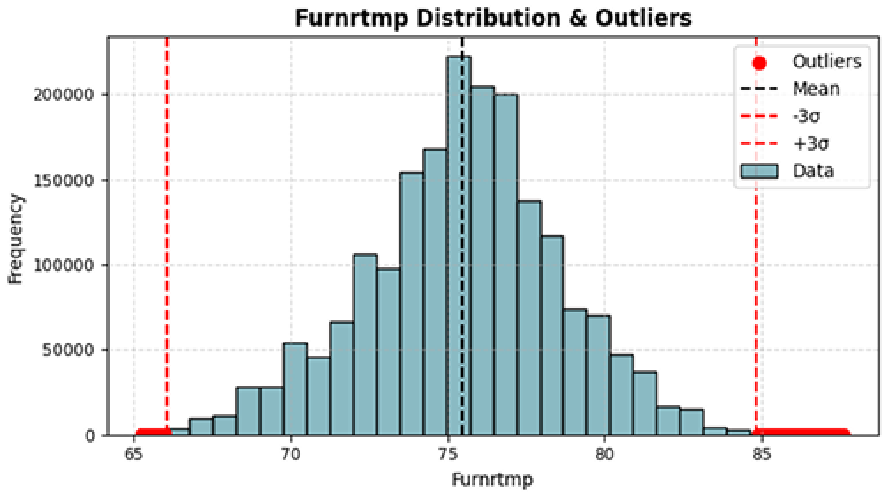


Figure 7. The 3-σ anomaly classification of the furnace return temperature feature of a system.

From *Figure 7*, the anomalies detected that fall outside the six-σ range would be classified as anomalies. Across a row, for the features in-scope, an outlier count is determined by how many features are classified as anomalous. Each anomalous feature is also recorded. From the number of anomalous features, a label is determined for the row, with the following thresholds described in *Table 2*.

Table 2. The anomaly labels and their corresponding feature count thresholds.

Anomaly Label	Anomalous Feature Count
Normal	0
Somewhat Normal	1
Abnormal	2-5
Very Abnormal	6+

These labels would be used as an indicator for the severity of an anomaly for a given datapoint. As the statistics distributions are relying on the bucket information statistics, the distribution is based on similar HVAC properties and behaviour as other seen data points, increasing the robustness of the algorithm.

3.3. Machine Learning

ML was leveraged to provide a more robust classification for anomaly detection. Isolation Forest was used as the de facto algorithm to analyze and detect anomalies. Before selecting Isolation Forest, multiple models were explored to finalize which ML model was best suited for anomaly detection. Three models that were utilized that were eventually dropped in favour of Isolation Forest were k-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Principal Component Analysis (PCA). The major requirements that the model needed to cover when working with unlabelled data was to pick up anomalies in an unbalanced and sparse dataset and be easily explainable. Refer to *Appendix F* for a detailed overview of the machine learning algorithms explored for the use-case.

3.3.1. Model Selected

Upon reviewing literature on other approaches to detecting anomalies in HVAC systems, the Isolation Forest model was commonly referenced and, based on the findings, seemed to be the most accurate at pulling the outlier points from the data. An assortment of plots created from the algorithm, represented in *Figure 8*. Applying it to the dataset, the model was particularly good at identifying anomalies, even with the high number of features present. Furthermore, due to the nature of the model working well with sparse or imbalanced datasets, it proved to work well in a dataset with these attributes. The main output of the model was an anomaly score which provides a clear metric of how to rank and categorize a particular system, on a given day, as experiencing an anomaly. Given this output, it was easy to explain to the client how the conclusions were made and, therefore, how the anomalies were detected.

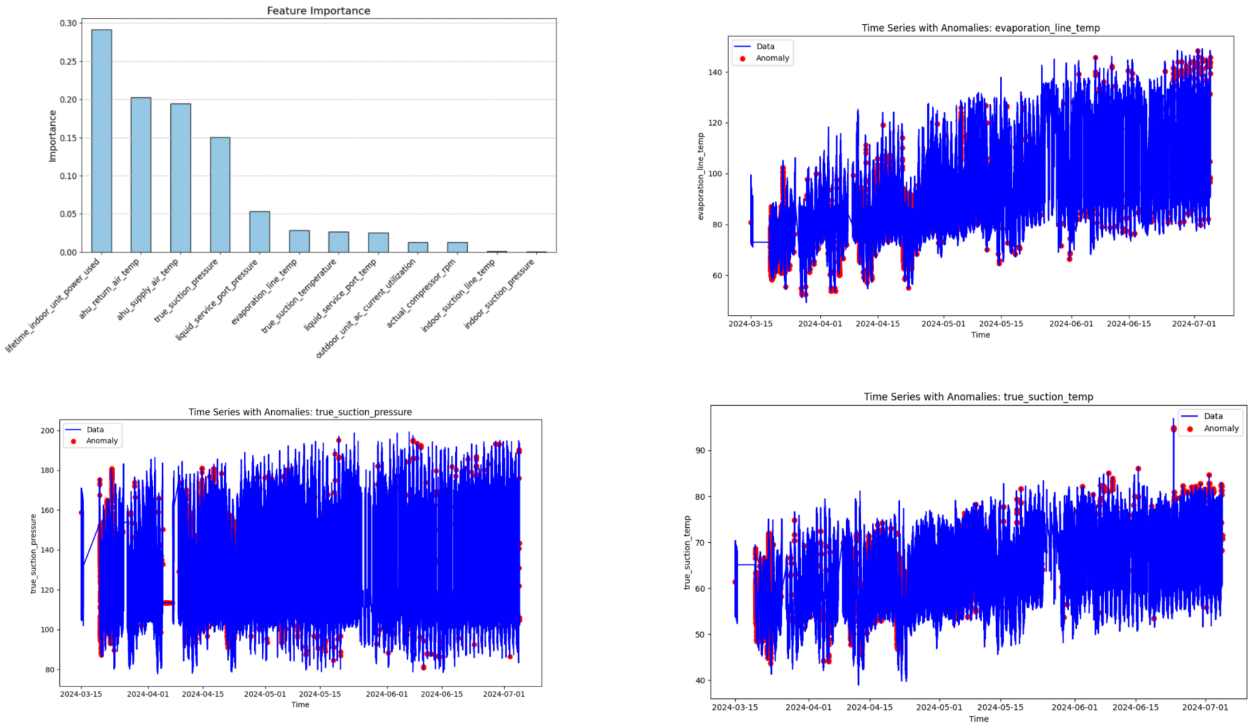


Figure 8. Feature significance of the isolation forest and anomaly erection plots for given HVAC system data.

3.3.2. Implementation

The model was tuned primarily through the contamination level. This parameter tells the model the expected proportion of anomalies believed to be in the dataset. Through in-depth testing, the contamination level settled on was five percent. A mixture of trial and error, as well as domain knowledge, both on HVAC system behaviour and the data itself, led to this conclusion. This number was used to ensure that we were only catching the most anomalous data while accounting for the fact that, due to the live nature of the data, there could be unrelated anomalies within it. If anything, the model at the set contamination level was meant to underestimate anomalies to reduce false positives as the number of anomalies should be relatively minimal due to the nature of HVAC systems.

Integrating ML with statistics and physics was a key aspect of our anomaly detection. The Isolation Forest model was not the primary way of detecting anomalies on the system due to several reasons mentioned earlier. One major limitation is that the data was unlabelled and therefore, difficult to confirm whether what the model was picking up was anomalous or just bad data or regular system behaviour that had not been seen. Instead, ML was used as a secondary validation tool to strengthen the confidence of anomalies detected through the statistics approach. At no point would an anomaly be flagged solely based on ML.

3.4. Physics Analysis

The COP is the standard performance metric for HVAC units across the industry. The formula in *Figure 9* provides a ratio between the heating or cooling effect delivered by the HVAC system and the power consumed by the compressor.

$$\text{COP} = \frac{|T_{\text{supply}} - T_{\text{return}}| \cdot \left(\frac{\Delta K}{\Delta^{\circ} F}\right) \cdot c_{p,\text{air}} \cdot \rho_{\text{air}} \cdot \text{CFM}_{\text{air}} \cdot \left(\frac{\frac{m^3}{s}}{\text{CFM}}\right)}{P_{\text{compressor}}}$$

Figure 9: Complete Coefficient of Performance equation.

In this expression, T_{supply} (the air temperature output) and T_{return} (the air temperature input) in °F capture the HVAC unit's temperature change on the air. That temperature difference is converted into a heat quantity by multiplying it by the specific heat capacity of air ($c_{\square,air}$) and the density of air (ρ_{air}). The airflow (CFM_{air}) is then converted to a volumetric flow rate in cubic meters per second (m^3/s) to maintain consistent units. Consequently, the numerator of the equation represents the total heat transferred per unit time, while the denominator ($P_{compressor}$) is the power drawn by the compressor. By dividing the absolute value of the heat transfer by the power, the COP reveals how effectively the system converts electrical energy into heating or cooling. A higher COP corresponds to a more efficient HVAC system, providing a greater thermal output for the same amount of power consumed.

This value, representing electrical efficiency, is useful to verify that a detected anomaly is impacting the performance of an HVAC unit. If anomalies are detected on a well-performing system, it could indicate noisy data rather than an issue worth investigating. This metric was useful internally as a validation technique for our ML component. Refer to Appendix G for more details on the COP equation.

3.5. Data-Driven Diagnostics

As a technician, certain attributes are investigated, and if they experience a certain combination of behaviours, it is probably a known course of action. These common cases can be found in *Table 3*. Given deviations from normal detected in suction pressure, liquid pressure, superheat, subcool and the outdoor unit power, some issues can be mapped to known cases. Based on the severity of the deviation from the normal mean, each attribute is given a rating of below normal, normal, or above normal. This threshold was set at ±5%. If this output matched any known problem, then a potential problem and solution were generated along with why this occurred.

Table 3: Common HVAC Problems with Known Solutions.

Diagnostics					Conditions/ Solutions
Suction Pressure	**Liquid Pressure**	**Superheat**	**Subcool**	**Amps**	
Lower than normal ↓	Lower than normal ↓	Lower than normal ↓	Lower than normal ↓		Insufficient airflow across the evap coil. Check filter, blower speed, cap, wheel and duct work.
Lower than normal ↓	Lower than normal ↓	Higher than normal ↑	Lower than normal ↓	↓	Undercharged, check for leaks and properly recharge system using SH and SC.
Lower than normal ↓	Lower than normal ↓	Higher than normal ↑	Higher than normal ↑	↓	Restriction, look for large temp drop. Plugged drier, stuck closed txv, restricted coil, etc.
Higher than normal ↑	Higher than normal ↑	Higher than normal ↑	Higher than normal ↑		Excessive load on evap coil, due to high airflow, leaking return duct, uninsulated return
Higher than normal ↑	Higher than normal ↑	Lower than normal ↓	Lower than normal ↓	↑	Insufficient airflow across condensing coil. Check coil, fan motor, cap, motor size.
Higher than normal ↑	Higher than normal ↑	Lower than normal ↓	Higher than normal ↑	↑	Overcharged. Recover refrigerant and adjust charge using SH and SC.
Higher than normal ↑	Higher than normal ↑	Lower than normal ↓	Maybe Higher or Lower ↑↓	↑	Noncondensibles/Air in system. Properly evacuate the system and recharge.
Higher than normal ↑	Lower than normal ↓	Lower than normal ↓	Lower than normal ↓		Over feeding metering device, txv stuck open, loose or improperly installed sensing bulb.
Higher than normal ↑	Lower than normal ↓	Maybe Higher or Lower ↑↓	Maybe Higher or Lower ↑	↓	Bad values in compressor, runs but doesn't pump, low amp draw and high surface temp possible.
Higher than normal ↑	Higher than normal ↑	Higher than normal ↑	Lower than normal ↓	↑	Low outdoor air flow (dirty condenser) OR High indoor temperature
Lower than normal ↓	Higher than normal ↑	Higher than normal ↑	Higher than normal ↑	↑	Sever restriction in refrigerant flow (with excess charge)
Normal	Higher than normal ↑	Normal	Higher than normal ↑	↓	Mild restriction in refrigerant flow (with excess charge)

3.6. Frontend Development

3.6.1. Data Retrieval

To access and retrieve telemetric data to the user interface, we leveraged the representational state transfer (REST) architecture of the backend system. Given the size of the telemetry data for a single system, only two weeks of data were called for, with only one day displayed. Empirically, two weeks was a sufficient degree of data in which the ML and diagnostics algorithms could leverage historical data, given that the algorithms require vast historical data.

3.6.2. User Interface

The front-end of our web application was designed to provide real-time insights into HVAC unit status and performance. Our conceptual design was guided by three key objectives: ensuring an intuitive dashboard layout, providing detailed diagnostics, and effectively visualizing anomaly trends. Early iterations prioritized clarity in system statuses while allowing users to drill down into specific issues as needed. We refined the interface through multiple wireframes and prototypes, incorporating user feedback to enhance usability. React was chosen as the front-end framework for its modular architecture, scalability, and efficient state management.

The final front-end system consists of a React-based web application that connects to a FastAPI backend system which retrieves data from MongoDB clusters to display real-time HVAC data. The two primary views include:

1. Dashboard Page *(Figure 10)* – Provides an overview of all HVAC units, displaying their status at a glance.

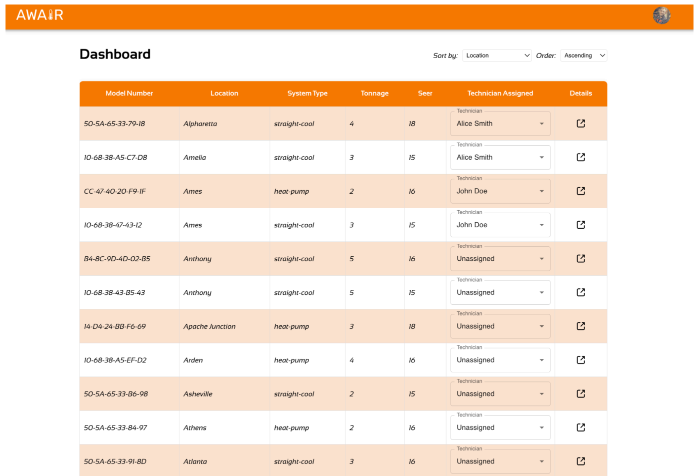


Figure 10. Dashboard Page.

2. System View Page (*Figure 11)* – Offers a detailed breakdown of a specific unit, including metadata, component statuses, diagnostics, and telemetry data.

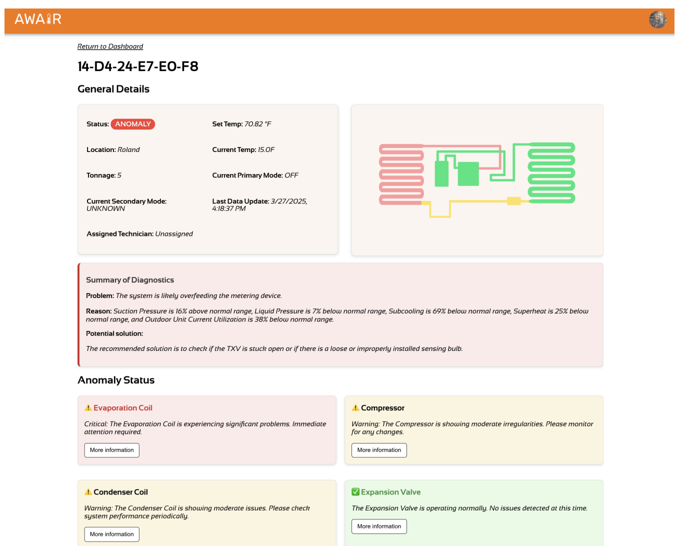


Figure 11. System View Page.

The System View Page monitors the four key components of an HVAC unit. For each component, pressure and temperature pairs are monitored and inspected for anomalies. Each component's corresponding temperature and pressure readings can be defined through *Table 4*.

Table 4. Each component has its corresponding temperature and pressure monitoring pair.

Component	**Temperature / Pressure Pairs**
Evaporation Coil	Indoor Suction Pressure, Indoor Suction Line Temperature, Furnace Supply Temperature

Compressor	True Suction Pressure, True Suction Temperature, Evaporation Line Temperature
Condenser Coil	Liquid Service Port Temperature
Expansion Valve	Liquid Service Port Pressure

Given that the condenser coil and expansion valve do not have pressure or temperature sensors, they rely on each other's missing metric to counterbalance this fact. Additionally, the evaporation coil will only analyze the indoor suction metrics, while the compressor will only monitor the true suction values.

Each component's status follows a hierarchical rule:

- Green: No anomalies detected.
- Yellow: Two sensors return warnings.
- Red: A sensor within the component returns a critical anomaly.

A crucial feature of the System View is the diagnostics modal, which dynamically updates when an anomaly is detected. It provides a structured summary of diagnostics, detailing:

- Problem: The nature of the issue
- Reason: Key sensor readings contributing to the issue
- Solution: Recommended actions

The anomaly status section determines the overall system condition through a voting mechanism driven by the statistical and ML results, as depicted in *Figure 12*.

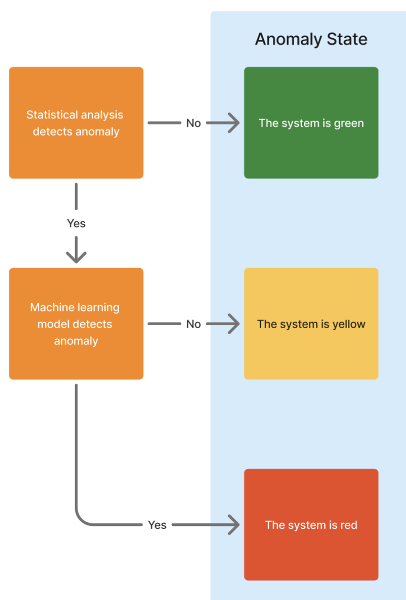


Figure 12. Voting mechanism for the anomaly state of a system.

The system features telemetry data visualization *(Figure 13)*, which displays sensor readings over the last 24 hours. These readings are plotted alongside the statistical thresholds described in Section 3.2.3. Anomalies are visually highlighted when data points exceed defined thresholds, helping users quickly identify trends and deviations.

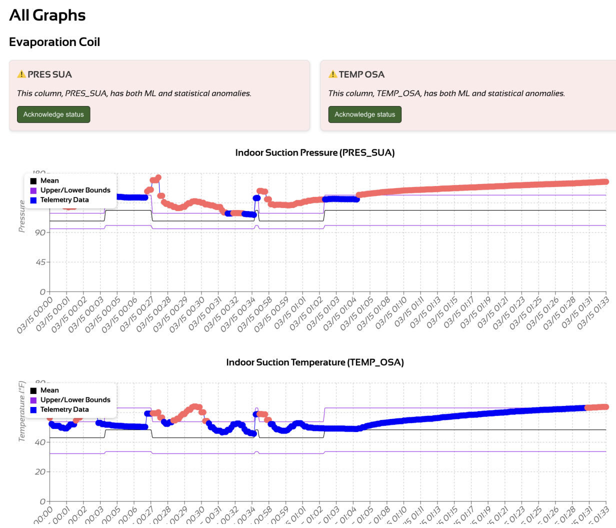


Figure 13. Telemetry Data on System View Page.

By integrating these components, the front-end delivers an intuitive, data-rich interface that enables users to monitor HVAC system health, diagnose issues, and take corrective action efficiently.

4. Solution Evaluation

4.1. Verification

Verification mainly concerned the correct implementation of anomaly detection. Our solution intentionally used a three-part approach whose components verified each other. The statistical model was verified by ensuring the right comparison between data and statistics was being performed. Data was being compared to its similar and corresponding properties and behaviour. The ML results, as the algorithms were unsupervised, were not able to be verified with labelled data. However, we purposefully underestimated the number of anomalies to detect, and in addition, only used ML results to support the evidence of anomaly presence. ML results did not deterministically conclude anomalous behaviour. The COP thermodynamic model was utilized to support the identification of anomalies from either method, and we worked with Professor Fraser, a mechanical thermodynamics professor, to ensure the thermodynamic calculations were verified.

4.2. Validation

To validate our solution, we leveraged many qualitative and quantitative methods. For ML validation, we leveraged a variety of literature including papers such as Pattern-Based Contextual Anomaly Detection in HVAC Systems (*Munir, n.d.*) and Predictive Maintenance Strategies for HVAC Systems (*Singh, 2023*), to confirm that an isolation forest was the best approach. COP values were validated using industry standards and averages. This allowed us to compare the values we calculated against a benchmark. The statistical approach was validated through the anomaly detection from the two other components. Once our system is implemented, the technicians can help validate the anomalies, which is a functionality built into our frontend solution. The primary validation method used for the final product was talking to experts at Alert Labs, which allowed us to connect with technicians as well as experts in the HVAC industry. Bi-weekly meetings with the client allowed us to get constant feedback on the graphs and insights displayed on the solution.

4.3. Impact

4.3.1. Social Impact

Our anomaly detection system positively impacts stakeholders across the HVAC ecosystem. Technicians, service providers, and homeowners gain benefits by improving communication, preparedness, and overall experience with HVAC maintenance. Technicians specifically benefit from having remote visibility into potential issues before arriving on-site, enabling them to gather the right tools and expertise. This preemptive approach reduces stress for both technicians and homeowners, as problems can be diagnosed and resolved faster. This increase in efficiency reduces costs and the logistical burden on HVAC service providers as well.

Preserving the privacy of the data provided was part of the implementation design. By conserving all confidential information, proprietary algorithms, and secret credentials, our client was shielded from data breaches. All data was kept in secure, closed databases.

4.3.2. Environmental Impact

Our solution aims to reduce energy consumption and greenhouse gas emissions by detecting HVAC anomalies early, enabling proactive maintenance and lowering electrical loads, which decreases indirect CO_2 emissions from power generation. Our system also flags refrigerant leaks, particularly of R410a, which is 2,000 times more potent as a greenhouse gas than CO_2. Early leak detection reduces wasteful repairs and minimizes harmful emissions. Additionally, our digital interface optimizes logistics, reducing unnecessary technician visits and cutting fuel consumption and travel emissions. By improving indoor air quality and reducing energy use, Awair supports the UN SDG goals of Good Health and Well-Being, Responsible Consumption and Production, and Climate Action.

4.3.3. Prosperity Impact

Our anomaly detection solution economically benefits the HVAC industry. Manufacturers gain real-world performance insights to improve products. Service providers can schedule proactive maintenance, reducing emergency repairs and increasing technician efficiency. Homeowners save on operating costs, avoid breakdowns, and build trust with providers, boosting contract renewals. Our solution promotes local procurement, connecting homeowners with specialized, nearby technicians, supporting small businesses and regional economic growth. By preventing failures and promoting efficiency, Awair demonstrates the value of preventive maintenance, reducing costs and encouraging sustainable HVAC practices.

Acknowledgements

The team thanks Professor Mehrdad Pirnia for his advocacy, Professor Roydon Fraser for his support, and Professor Carrie Mitchell for her feedback and fresh perspective. We would also like to thank the Alert Labs team, specifically Charlie Harland, Max Pfeifle, Daniel Urquhart, Kevin Wright, Cody Hines, Aditi Mukhopadhyay, Joshua Doncaster-Marsiglio and Mike Adams for providing feedback and guidance on our approach. Lastly, we thank Professor Ada Hurst for helping keep us on track with the deliverables and course requirements.

References

Hu, Y., & Yuill, D. P. (2022). Multiple simultaneous faults' impacts on air-conditioner behavior and performance of a charge diagnostic method. *Applied Thermal Engineering*, *215*, 119015. https://doi.org/10.1016/j.applthermaleng.2022.119015

Munir, M., Erkel, S., Dengel, A., & Ahmed, S. (n.d.). *Pattern-based contextual anomaly detection in HVAC ...* Pattern-Based Contextual Anomaly Detection in HVAC Systems. https://www.dfki.de/fileadmin/user_upload/import/9505_pattern-based-contextual-anomaly-detection.pdf

Singh, D., Arshad, Mohd., Tyagi, B., & Kalia, G. (2023, October 12). *Predictive maintenance strategies for HVAC systems: Leveraging MPC, Dynamic Energy Performance Analysis, and ML classification models*. IRE Journals. https://www.irejournals.com/paper-details/1705111/

Temperature - dry bulb/web bulb/dew point. (n.d.). https://www.weather.gov/source/zhu/ZHU_Training_Page/definitions/dry_wet_bulb_definition/dry_wet_bulb.html

Appendices

Appendix A:

Goal of the interview:

- Understand key roles and responsibilities as an HVAC technician.
- Where they interact with data and how it is measured.
- The insights drawn, how it is done, and its value.
- Determine the current process in evaluating various types of issues and how this may differ based on the call type.

Hello, <participant name>. My name is <facilitator name>, and I am part of the Capstone Team from the University of Waterloo working with Alert Labs on creating an analytics tool for the team. Thanks for joining me.

The way it's going to work today is that I'll start out by asking you a few questions about your work and your common roles and responsibilities. Then we'll go into more specifics of how you identify problems within HVAC systems and we'll leave time for any questions you may have for us. Some members of our team may be taking notes throughout the session, and some may hop in if they have other questions. Before we begin I'll take a moment to say that privacy is very important to us. Your name and contact information won't be shared with anyone or included in our reporting – we used those just for scheduling and arranging to meet with you today. In fact this is true for any of the data that you show us today. Feel free to talk about as much or as little as you're comfortable with. We'd also like to record the session. We use our recordings for note-taking, to make sure we have captured all your feedback accurately. The recording won't be shared with anyone except the people working on the project, and we won't tie it to your name or make it available to the public. Is that okay with you? [if yes, begin recording] Great, let's continue.

Questions

1. Roles and Responsibilities:
    - Can you tell me a little bit more about the overall roles and responsibilities of an HVAC technician?
2. Common Issues and Diagnostics:

   - What are the most common issues that are typically found in different HVAC systems, and how do you go about identifying them?
   - How do you determine if there are leaks in refrigerant lines and ensure proper refrigerant levels?
3. Issue Detection and Root Cause Analysis:
   - If there are any issues detected within a system, how is the root cause determined?
   - How do you prioritize which HVAC systems need attention?
4. System Placement and Standards:
   - What are the factors that impact where an HVAC system is placed, and is this a standardized process?
5. Inspections and Maintenance:
   - Are inspections/cleanings done on HVAC systems once they have been installed, and if so, how often? What is the process of setting up inspections and determining what areas of the system to evaluate?
6. Monitoring and Efficiency:
   - How is the efficiency of the HVAC system monitored once it is installed, and are there any changes to the unit if there is a decrease in efficiency detected?
   - What specific data points or metrics are most critical for you to monitor?
   - What types of visualizations (graphs, charts, heat maps) do you find most useful?
7. Replacement Parts and Problem-Solving:
   - If replacement parts are needed, are they typically brought with the technician? Is there a standard set of parts, and are there any issues that could require additional parts?
   - Walk us through the process of identifying a problem, and what happens if a replacement part is required.
   - When customers call with complaints, do you have a way to recommend replacement of parts based on the information they provide?
   - What information do you use to determine a product's health and lifetime?
8. Alerts and Efficiency:
   - As a technician or a customer support representative, what alerts would be beneficial in either providing better value to customers or streamlining the efficiency of your work?
9. Additional Insights:
   - Are there any other things you think we should know?

Appendix B:

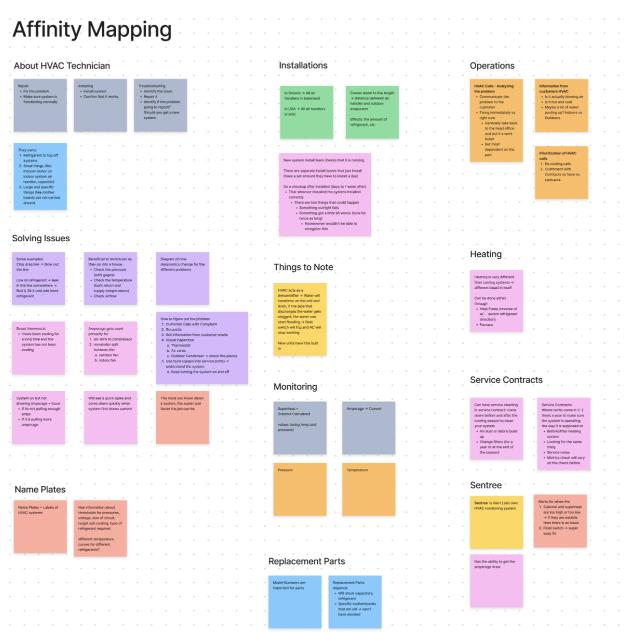
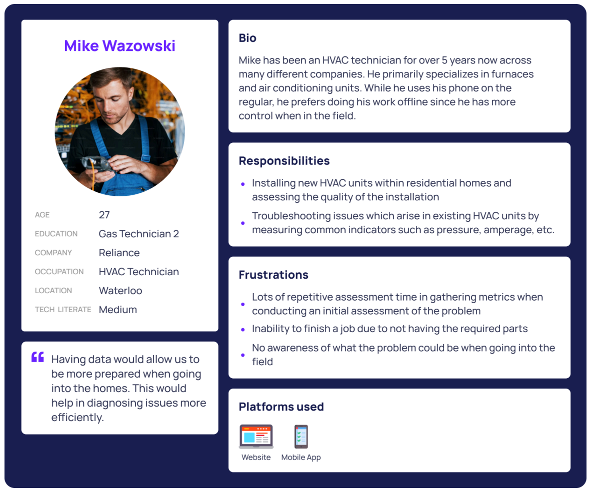
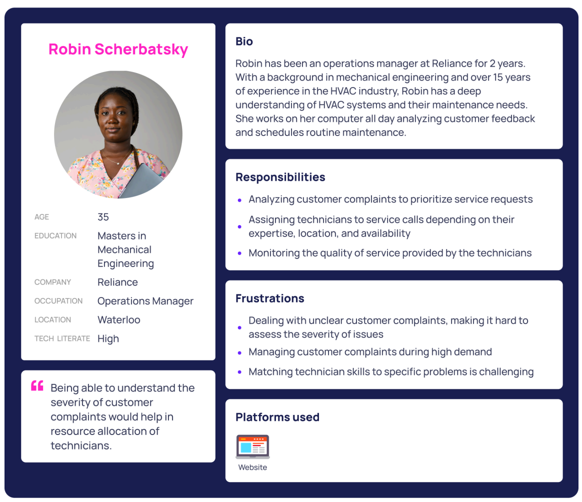


Figure B1. Affinity map grouping the common themes from the user interview (linked here).

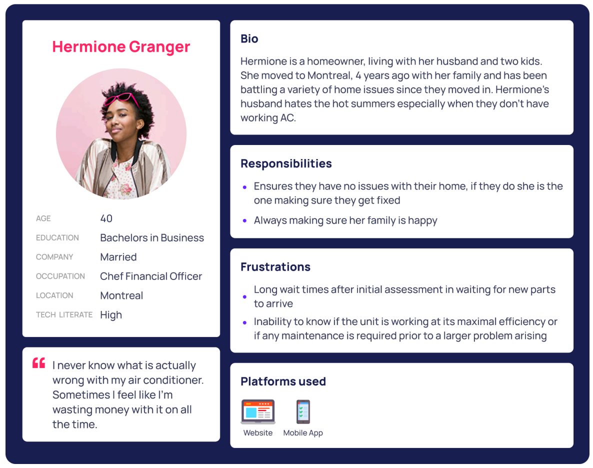
Appendix C:



User persona for technician (above)



User persona for HVAC service provider (above)



User persona for homeowner (above)

Appendix D:

Gathering an initial understanding of the bounds and acceptable ranges for each case or scenario was critical to building behavioural profiles of HVAC systems. *Figure D1* describes one of 100+ visualizations created to determine the distribution and ranges of a critical variable, the furnace return temperature.

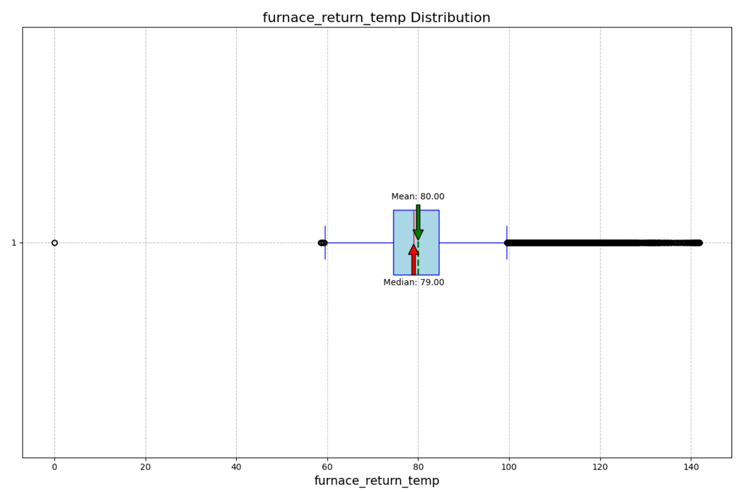


Figure D1. Distribution of furnace return temperature.

Multivariable analysis was conducted on paired attributes within the HVAC system. As depicted in *Figures D2* and *D3*, correlations were identified among collected attributes to determine how strong the relationships are between different variables. For example, *Figure D2* illustrates a strong relationship between pressure and temperature at the service port, while *Figure D3* demonstrates an inverse correlation between true suction pressure and temperature. Understanding these correlations aids in developing behavioural profiles of HVAC systems, enabling technicians to pinpoint where and when problems are present.

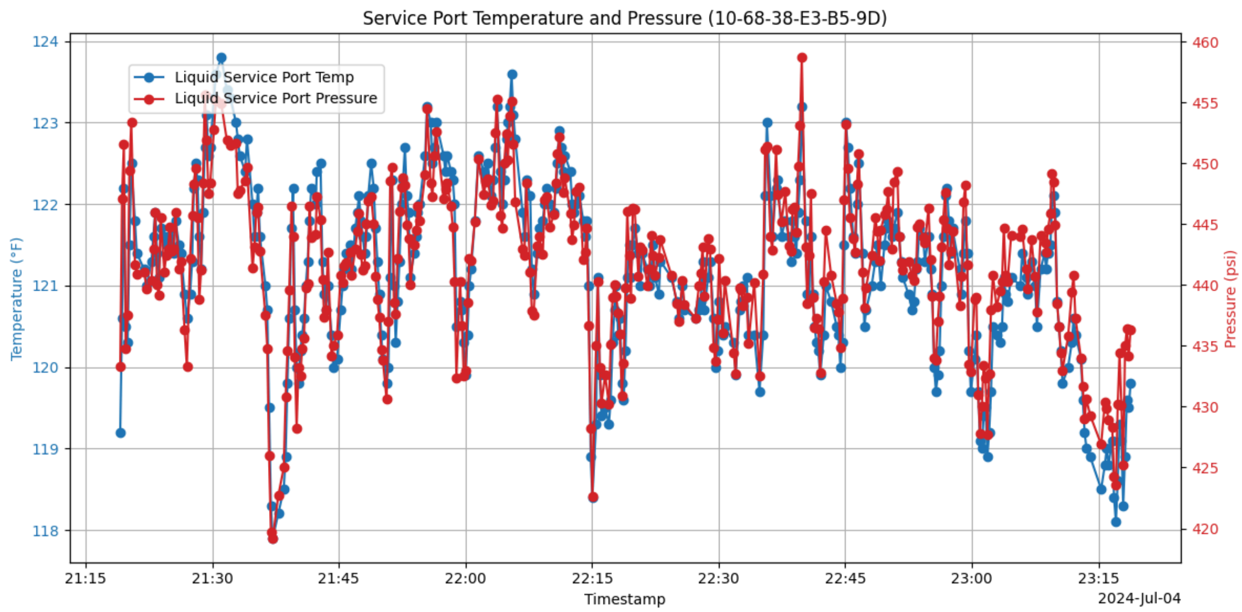


Figure D2. Relationship of service port temperature and pressure.

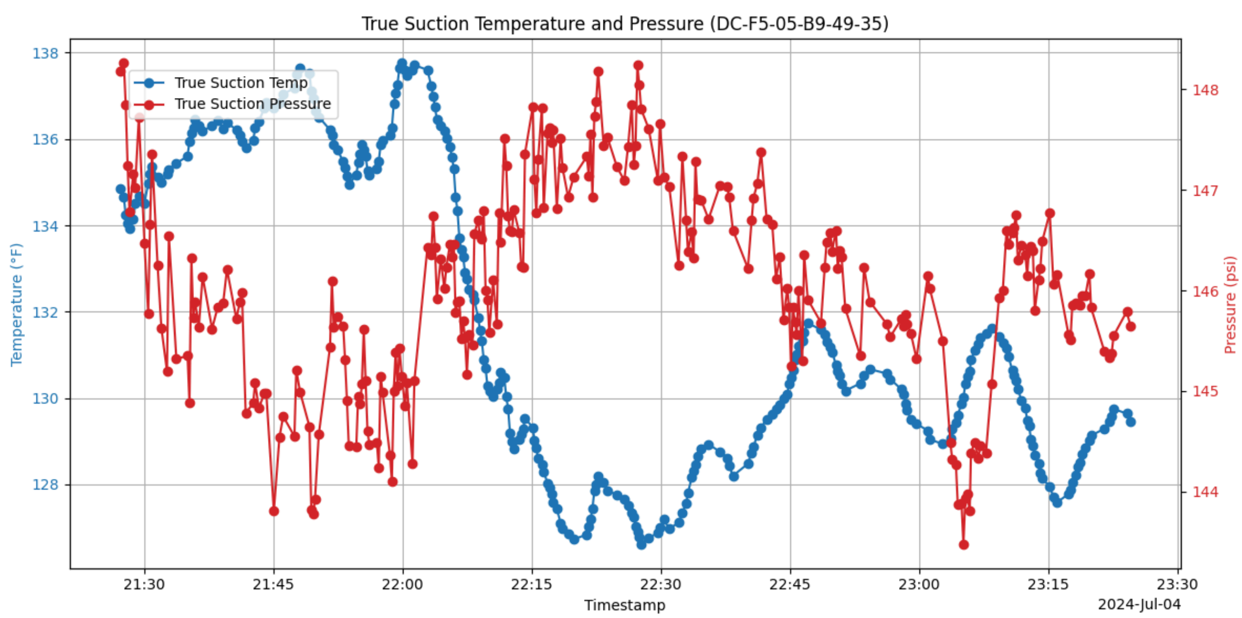


Figure D3. Relationship of true suction temperature and pressure.

Appendix E:

Table D1. The criteria for which to split HVAC telemetry data into different classes of behaviour.

Criteria	**Description**	**Number of Possible Values**
Tonnage	Refers to a unit's ability to cool the environment.	4

Seasonal Energy Efficiency Ratio (SEER)	Describes the energy consumption and efficiency of the unit's cooling ability.	3
Primary Mode	Driven by the HVAC operating mode of the system as well as the target compressor speed.	10
Secondary Mode	A granular description of the primary mode which details the intensity of the HVAC mode, driven by the compressor speed.	20
Wet Bulb Temperature	The temperature of adiabatic saturation. This figure is derived from the general wet-bulb temperature equation using values of relative indoor humidity and the set thermostat temperature. These are split into buckets from 7℃ and 31℃.	12
External Weather Temperature	Describes the weather of the city that the HVAC resides in at the time of measurement. The temperatures are captured into buckets from -21℃ and 46℃.	22

Appendix F:

Models Explored

First, KNN was selected as a candidate to explore because it was assumed that the anomalies would have only a few close neighbours in high-dimensional space. However, due to the high number of features that were present in the data, this model ran into the curse of dimensionality. This meant that normal data was labelled as anomalous, rendering the model unsuitable for this. As shown in *Figure F1*, the anomalous points have significant overlap with the normal data. The other major problem was the difficulty in explaining its output to the client and how the model arrived at the particular anomalies it selected.

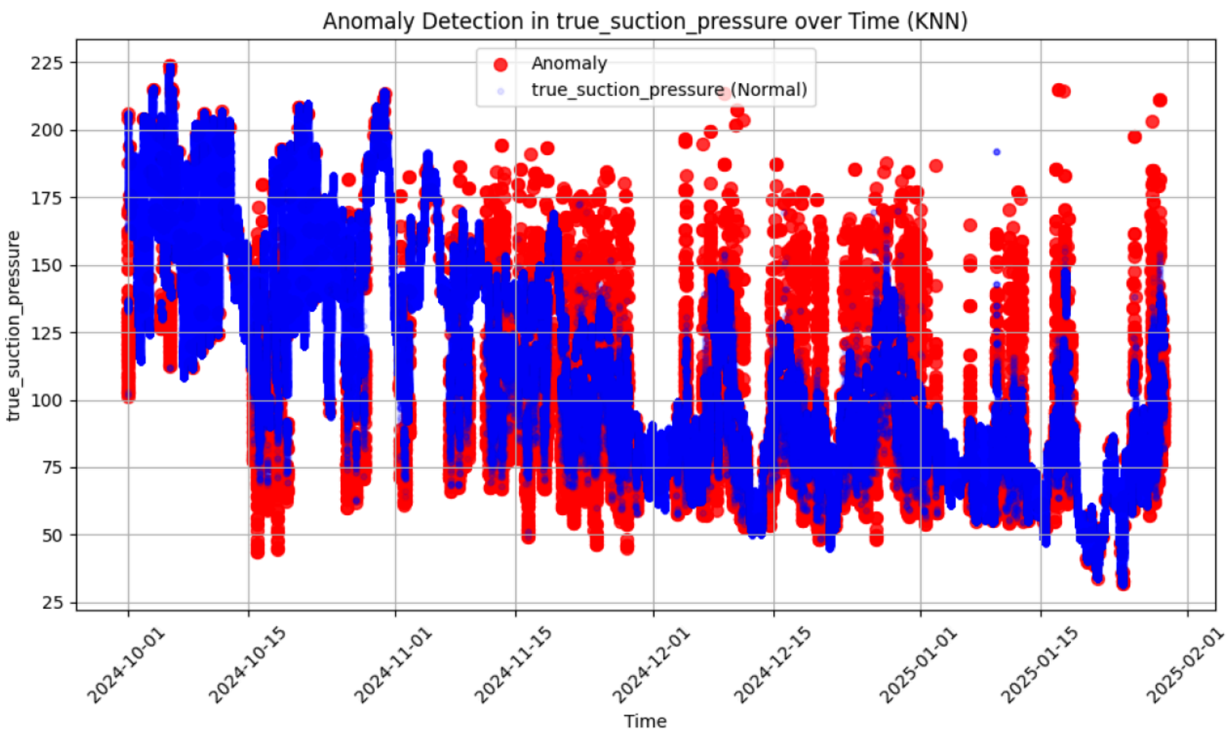


Figure F1. KNN anomaly detection on the system's true suction pressure.

Next, exploring SVM, the focus was on a particular variant called One-Class SVM. This particular model relies on learning decision boundaries around normal data points. The problem with this arose since the definition of what is 'normal' could not be confirmed due to the unlabelled nature of the data. This is shown in *Figure F2*, with the unexpected classifications of data points as anomalous in the middle of what looks like normal operating procedure. With this model relying heavily on the assumption that its classification of normal was accurate, it struggled when there were significant differences from the 'normal' which was expected.

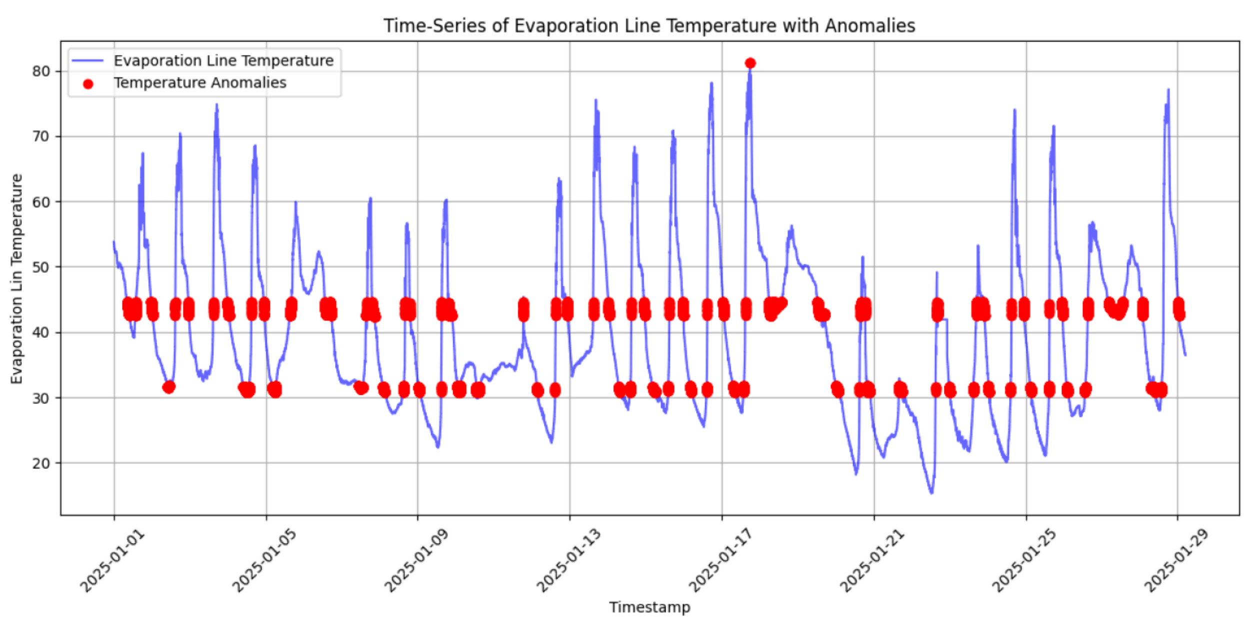


Figure F2. SVM anomaly detection on the system's evaporation line temperature.

Another model that was tested was PCA. This model relies on the belief that normal data will appear in a low-dimensional space and is captured using linear transformations. However, failures often happen in non-linear interactions which is where this model is susceptible to failures. Moreover, the model looks at outliers as it is a variance-based approach. Given that the data, especially prior to January of this year, was filled with sparse data from inaccurate sensor readings, this created issues when trying to detect true positives, as illustrated in *Figure F3.*

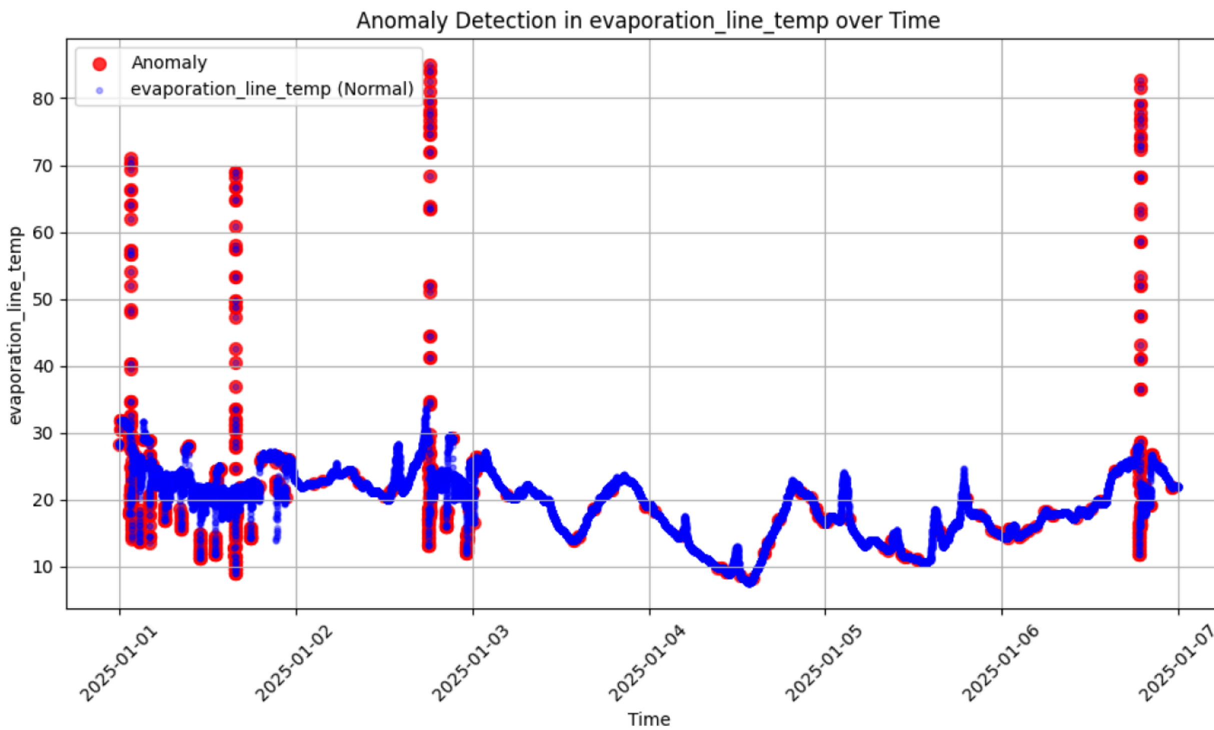


Figure F3. PCA anomaly detection on the system's evaporation line temperature.

Appendix G:

$$COP = \frac{|T_{supply} - T_{return}| \cdot \frac{\Delta^\circ K}{\Delta^\circ F} \cdot c_{p_{air}} \cdot \rho_{air} \cdot \mathrm{CFM}_{air} \cdot \frac{\mathrm{m}^3/\mathrm{s}}{\mathrm{CFM}}}{P_{compressor}}$$

Where:

T_{supply} = Air Supply Temperature in °F

- Column: FURNSTMP
- Column: TEMP_LAT

T_{return} = Air Return Temperature in °F

- Column: FURNRTMP
- Column: TEMP_RAT

Δ°K / Δ°F = Temperature conversion factor between °K and °F

- Constant: 5/9

c_p^{air} = Specific Heat Capacity of Air at constant pressure in kJ/kg°K

- Constant: 1.005

ρ_{air} = Air Density in kg/m³

- Constant: 1.293

CFM_{air} = Airflow in CFM (cubic feet per minute)

- Column: F_ACTCFM

m^3/s / CFM = Volume flow rate conversion factor between m^3/s and CFM

- Constant: 0.0004719

$P_{compressor}$ = Unit Power in kW

- Column: DR_ODUIP