A Cross-lingual Search Engine for Retrieval of Green House Gas Emission Factor

Yachai Limpiyakorn1, Nattapot Nuwong2

1, 2 Department of Computer Engineering, Chulalongkorn University, Bangkok 10330, Thailand

Corresponding Author: Email address: yachai.l@chula.ac.th

*Abstract*—*Climate change is a critical global challenge primarily driven by greenhouse gas (GHG) emissions. Managing GHG data effectively is essential for accurate carbon footprint assessments. However, existing information retrieval systems often lack cross-lingual search often lack cross-lingual search capabilities, limiting access to Green House Gas Emission Factor (EF) data across languages, This study present a cross-lingual search engine designed to retrieve EF data effectively, integrating dictionary-based synonym matching and Elasticsearch indexing, The system enables seamless bilingual search functionality, it supports synonym-based search expansion, ensuring relevant result regardless of query language. By implementing an automated pipeline for data retrieval and indexing, the system enhances retrieval accuracy and efficiency thereby aiding organizations in precise carbon footprint calculations.*

Keywords— Cross-Lingual, Information Retrieval, GHGs Emission Factor, Synonym.

# Introduction

The world is currently facing extreme weather conditions, primarily due to the continuous and substantial emissions of greenhouse gases (GHGs) into the atmosphere. These emissions have widespread impacts on ecosystems, climate, biodiversity, and the quality of life at both local and global levels. Consequently, the effective monitoring and management of GHG emissions data is of paramount importance, especially for Thailand. The principal agency responsible for assessing, collecting, compiling, and disseminating GHG-related information is the Thailand Greenhouse Gas Management Organization (TGO), which operates under the Ministry of Natural Resources and Environment. TGO has developed comprehensive databases and report covering a wide range of sectors including industry, energy, agriculture, and communities as well as information for assessing and calculating carbon footprints. Moreover, it determines GHG emissions factors (EF) based on international standards and guidelines provided by the Intergovernmental Panel on Climate Change(IPCC). However, despite the TGO’s efforts to publish GHG emission factor data in various formats on its website (such as online databases and electronic files), there remains a gap the absence of a search system that can comprehensively handle EF queries in a cross-lingual context. In the content provide on TGO’s Website, some EF keywords appear in English, some in Thai, and some in both languages. Additionally, user preferences vary, with some users favoring Thai search terms while others prefer English. These linguistic discrepancies can cause conventional single-language search approaches to yield inaccurate or incomplete result. This research proposes the development of a cross-lingual search program for retrieving of GHG emission factor. The system is designed to support in both Thai and English and to incorporate synonym matching to enhance user accessibility regardless of language. It enables rapid and comprehensive searches, with the search result displaying EF entries across different versions to allow users to select the most appropriate value. This functionality is particularly beneficial for practitioners who need to choose correct EF values when preparing an organization’s carbon footprint report, thereby reducing errors, avoiding redundant work, and ensuring efficient resource utilization.

# Background

1. *EF (Greenhouse gas emission Factor)*

The Greenhouse gas Emission Factor is calculated based on the amount of greenhouse gas emitted and absorbed per unit of activity. It is used to assess the volume of greenhouse gas emission resulting from various activities, such as energy consumption, fuel consumption, transportation, or industrial production processes. This factor is a critical variable that enables organizations, government agencies, small businesses to calculate and analyze EF. It is essential for reporting, impact monitoring, and planning strategies to reduce carbon footprint. The calculation relies on secondary data sources, ranked by reliability from highest to lowest [1-3]

1. Thailand Greenhouse Gas Management Organization (TGO)
2. Research studies conducted in Thailand
3. Publicly available databases, including Open LCA
4. Intergovernmental Panel on Climate Change (IPCC)
5. *IP (Information Retrieval)*

Information Retrieval is a fundamental conncept in information science that focuses on retrieving relevant documents from a large collection based on user queries. Various model have been developed to improve retrieval efficiency and effectiveness, including

1. Fundamental IR Models

• *Boolean Model*: Uses logical operators such as AND,OR and NOT to define search conditions. Documents that meet all conditions are retrieved based on an exact match, but this model lacks ranking mechanisms to order results by relevance.

• *Vector Space Model (VSM)*: Represents documents and queries as vectors in a high-dimensional term space. The relevance of a document to a query is computed using cosine similarity, allowing for ranked retrieval based on degree of similarity.

*• Probabilistic Models (e.g BM25*): Estimate the probability that a document is relevant to a query. These models incorporate factors such as Term Frequency (TF), Inverse Document Frequency (IDF), and document length normalization to refine retrieval rankings

2. Performance Evaluation

To assess the effectiveness of IR systems, several widely used evaluation metrics are applied

• *Precision:* Measures the proportion of retrieved documents that are relevant. It is calculated as eq where TP (True Positives) represents relevant documents retrieved, and FP (False Positives) represents non-relevant documents retrieved.

• *Recall:* Evaluates the system ability to retrieve all relevant documents from dataset. It is calculated as: eq where FN (False Negatives) refers to relevant documents that were not retrieved.

• *Mean Average Precision (MAP):* A ranking-based evaluation metric that computes the average precision at different recall levels across all queries. It is particularly useful for systems that return ranked lists of results eq where AP(q) represents the average precision for query q, and Q is the total number of queries.

These metrics are essential for assessing IR system effectiveness and optimizing retrieval algorithms to enhance user experience and search accuracy.

1. *CLIR (Cross-Lingual Information Retrieval)[4]*

Traditional information Retrieval systems assume that queries and documents are in the same language. However, in CILR users may input queries in one language (e.g., Thai) while expecting relevant documents in another language (e.g., English) or in both languages. To enable efficient CLIR, a mechanism is required to bridge linguistic differences between languages. Two primary approaches are widely employed

1. Synonym-Based (Dictionary-Based)[5]

This approach relies on a bilingual dictionary containing predefined word-pair mappings between Thai and English. During the tokenization and indexing process, query expansion is performed to include synonymous terms in the target language. For example, if a user searches for “ก๊าซเรือนกระจก” the system will retrieve documents containing “greenhouse” and “gas greenhouse” as well. The bilingual dictionary operates as part of domain-specific knowledge and includes the following components

* + *Alternative Labels:* Terms that can be used interchangeably and the same meaning such as abbreviations, acronyms, or spelling variations.

*example:* CTO – Chief Technology Officer

* + *Synonyms:* Words or phrase that convey the same or similar meaning despite minor contextual differences.

*example:* human - homo sapiens, mankind

* + *Taxonomy:* A hierarchical system that organizes terms into structured categories and defines relationships among them.
  + *Ontology:* A formal representation of complex relationships between entities in a domain, including hierarchical dependencies (e.g. employee reporting to a manager)
  + *Knowledge Graph:* A practical implementation of ontologies that explicitly defines entity relationships  
    example: Michael is an employee, Micheal report to Jim, thus Jim is Micheal manager

2. Embedding-Based

This approach leverages Multilingual Natural Language Processing (NLP) models, such as BERT, LaBSE and others, to encode both Thai and English sentences into a shared vector space (latent space). When a user inputs query in Thai, the system converts the query into a vector and matches it against document vectors, regardless of their language. If two phrases have similar meaning, their vector representations will be close to each other in the latent space.

1. *NLP (Natural Language Processing)*
   * *Tokenization* In Thai

.

* + *Stop Words*

.

* + *NER (Named Entity Recognition)*

1. *Technology stack*

• *Elasticsearch[6]*

• *Apache Airflow*

Noise is implicitly addressed through discussions on SNR, which is a critical metric used for assessing the robustness of a communication link. A higher SNR implies a cleaner (less noisy) signal, which is crucial for ensuring reliable data transmission over LoRa networks. With higher spreading factors, the concern arises, how LoRa devices achieve better immunity against multipath fading and signal degradation, which are forms of noise. The discussion ties the concept of noise directly to the reliability and effectiveness of the communication channel. Focusing on managing noise through technological strategies like choosing appropriate spreading factors can significantly enhance performance of LoRa-based systems. Regularly measuring noise levels, optimizing node placement, and using appropriate hardware, can mitigate the impact of noise on the network

# METHODOLOGY

Fig. 2 illustrates the system design of an IoT-based UVC air sterilizer that lays the groundwork for both hardware and software development. LoRaWAN is opt to manage the communication between end devices and gateways in a LoRa network. The measures including RSSI, SNR, Noise, are used for evaluating how well the system could monitor and manage the air sterilizers, particularly the communication efficiency and system reliability in a real-world setting. The proposed solution was tested by simulation of typical operational conditions encountered in environments like hospitals or laboratories, where maintaining air quality is crucial.

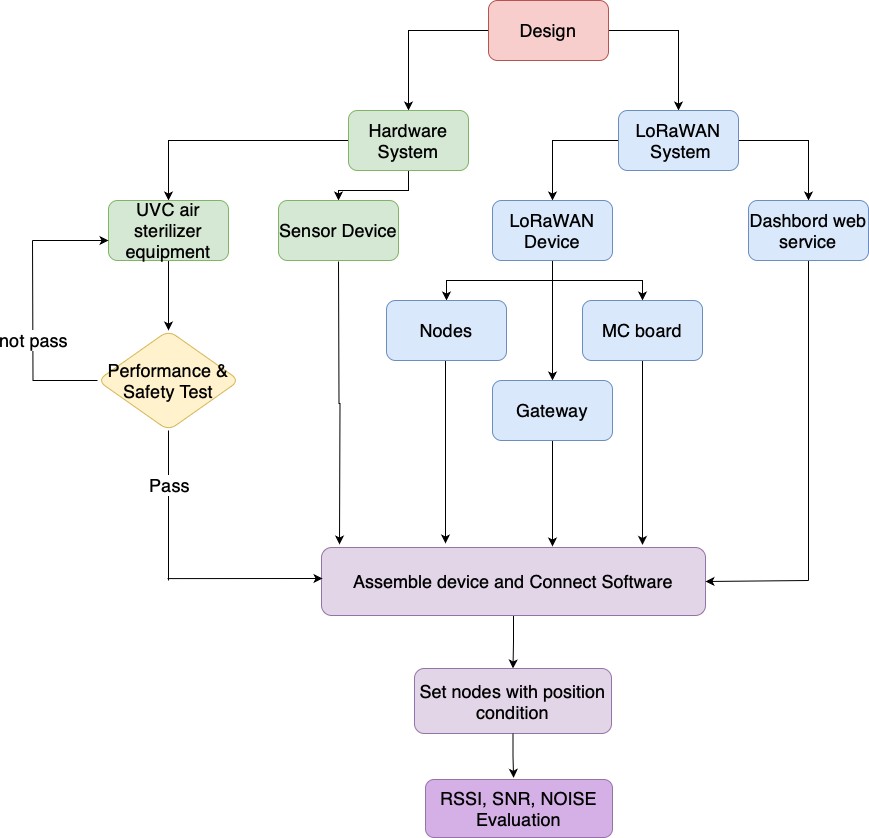


Fig. 2. System design of UVC air sterilizer with LoRaWAN.

As shown in Fig. 3, each component plays a vital role for the overall system functionality. Sensor devices were integrated with LoRaWAN technology to enable continuous monitoring and control capabilities. The implementation started from the sensor which captures critical data passing to the microcontroller for processing and preparing the data for transmission. The integration with LoRaWAN enables seamless data transfer to the gateway, which acts as a crucial intermediary, bridging the local network with the cloud infrastructure via the internet.

TABLE I. Quality rate of signal strength.

|  |  |  |  |
| --- | --- | --- | --- |
| **Quality** | **Signal Strength (dBm)** | **mW** | **Description** |
| Amazing | -30 dBm | 1/1,000th of 1 milliwatt | Max achievable signal strength  Not typical or desirable in the real world |
| Very Strong | -40 dBm | 1/10,000th of 1 milliwatt | Good Level for Wifi connecting |
| Very Strong | -50 dBm | 1/100,000th of 1 milliwatt | Very strong signal with minimal packet loss |
| Very Strong | -60 dBm | 1 millionth of 1 milliwatt | Good Level for LTE and 3G |
| Strong | -70 dBm | 1 ten-millionth of 1 milliwatt | Reliable signal with low to moderate packet loss Good Level for email and web |
| Fair | -80 dBm | 1 hundred-millionth of 1 milliwatt | Minimum signal strength for basic connectivity Packet delivery maybe unreliable |
| Weak | -90 dBm | 1 billionth of 1 milliwatt | Approaching to the noise floor |
| Very Weak | -95 dBm | Noise Floor | Under -90 dBm is no signal |

TABLE II. SNR quality levels.

|  |  |  |
| --- | --- | --- |
| **Quality** | **Range (dB)** | **Description** |
| Very good | >10 dB | Excellent signal quality with minimal interference |
| Good | 6-10 dB | Reliable signal with low interference |
| Fair | 0 - 6 dB | Usable signal, but some interference may occur |
| Poor | -10 - 0 dB | Weak signal with significant interference, communication might be unreliable |
| Very poor | < -10 dB | Very weak signal with high interference, unreliable for communication |

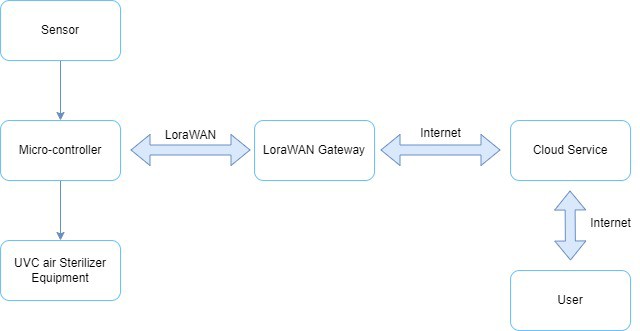


Fig. 3. Design of remote controlling of UVC air sterilizers.

1. *Experiment*

The experiment involved a setup where multiple UVC air sterilizers equipped with sensors were interconnected through LoRaWAN.



Fig. 4. Assembly devices.

These sensors collected various data points, such as operational status and efficiency metrics of the UVC lamps. Each sensor node is linked to a central gateway via LoRaWAN, enabling long-range and low-power communication. The setup aimed to validate the system's capability to function seamlessly in conditions that mimic real-world operational scenarios. The assembly devices are shown in Fig. 4*.*

* + *Location and Configuration*− The nodes were strategically placed in a designated indoor environment to mirror a typical installation in a public building or healthcare setting. The arrangement was designed to test the LoRaWAN network’s ability to handle signals in a space with potential physical and electronic interference, such as walls and other electronic devices that might impact signal quality. In this experiment, one unit of UVC equipment was installed covering the area not exceeding 50 square meters. Several factors need to be considered to ensure the system operates efficiently. The measure of Cubic Feet per Minute (CFM) is used to determine the volume of space and decide how quickly required for exchanging the air in a room. The equation is shown in Fig. 5. Several factors need to be considered to ensure the system operates efficiently, including:

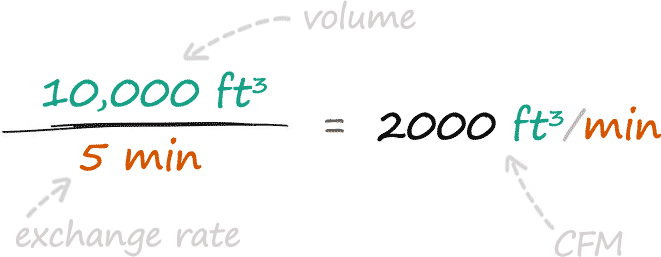


Fig. 5. CFM calculation equation.

− Device Positioning

The device mounted on the wall or ceiling should be at a point that allows the device to cover the entire usable area without any obstructions blocking the signal. The installation should be at a level that can distribute the signal to all important areas in the room effectively.

− Distance from the Gateway

The distance between the node and the gateway should not be too far, as it could cause the signal to drop out easily. The appropriate distance depends on the capabilities of the device and the environmental conditions within the room, including its ability to resist interference.

− Characteristics of Walls and Room Layout

The characteristics of the walls, such as the construction materials used, can affect the propagation of wireless signals. The wall made of concrete or metal may block or reflect wireless signals. Additionally, the layout and arrangement of the room are important as they can affect the distribution and reception of signals between the node and the gateway.

− Pre-installation Testing

Prior to permanent installation, the system should be tested to ensure that the signal from all devices can connect well with the gateway and that the system is stable. Fig. 6 illustrates the testing scenario of room size 5x8 meter containing eight UVC Lora nodes and a LoRa gateway.

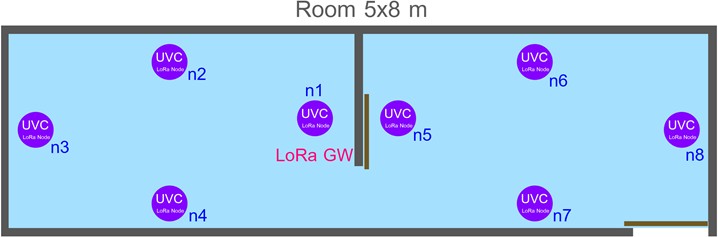


Fig. 6. Layout of LoRa gateway and nodes in experiment room.

1. *Evaluation*

The outcomes of the experiment provided insights into the practical deployment of IoT technologies in managing air quality control systems. The collected data helped in understanding the interaction between device placement and signal efficacy, which is crucial for deploying IoT solutions in complex environments. The measurement data gathered from each node, RSSI, SNR, and Noise, were used to compare with the threshold of each parameter to ensure that the system and each node function properly without any effect to decrease communication efficiency as well as to identify any potential drop in performance due to environmental factors.

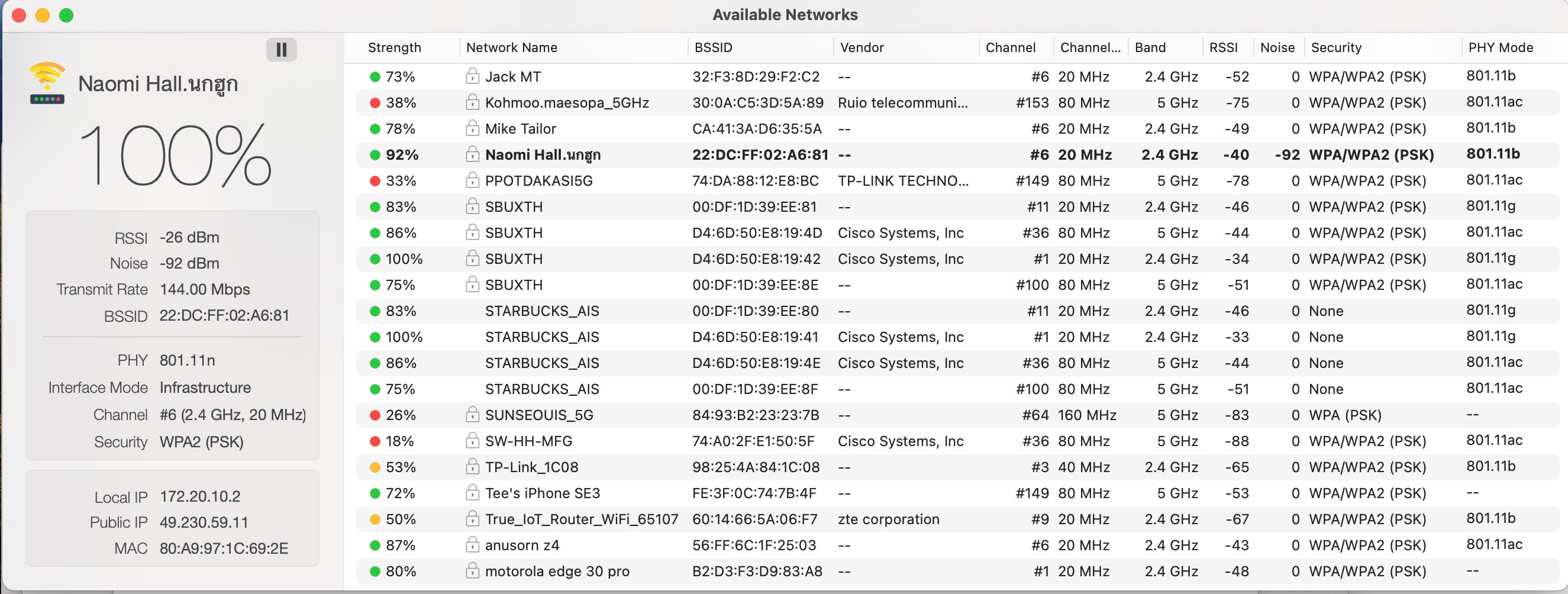


Fig. 7. Example information resulting from using WiFi signal strength analyzer application.

Fig. 7 reports the measurement data of RSSI and Noise by coding program on MC board or using application. RSSI is a measure of the power level that an individual device receives from the wireless access point or router. The measurement data gathered from the network “Naomi Hall” reported RSSI of -40 dbM (very strong signal) where -30 dBm (excellent) and -90 dBm (very weak). The value closer to zero denotes the stronger signal. The Noise level was reported -92 dBm, where typical values range from -90 to -120 dBm. Contrary to RSSI, the value of Noise closer to zero indicates more interference. The value of SNR was computed from the difference between the signal strength and the noise level. Here, the calculated value of SNR equals 52 dB considered excellent signal quality with minimal interference.

1. *Measurement Data*

Example measurement data of node2, 3, 4 with timestamp sending between MC board and nodes were collected. The graphs that visualize the measurement data of signal strength (RSSI), signal-to-noise ratio (SNR), and the presence of any noise interference affecting communication and visualized are shown in Fig. 8, 9, 10, respectively. These measurement data were useful for assessing the network's robustness and the reliability of data transmission across the system.

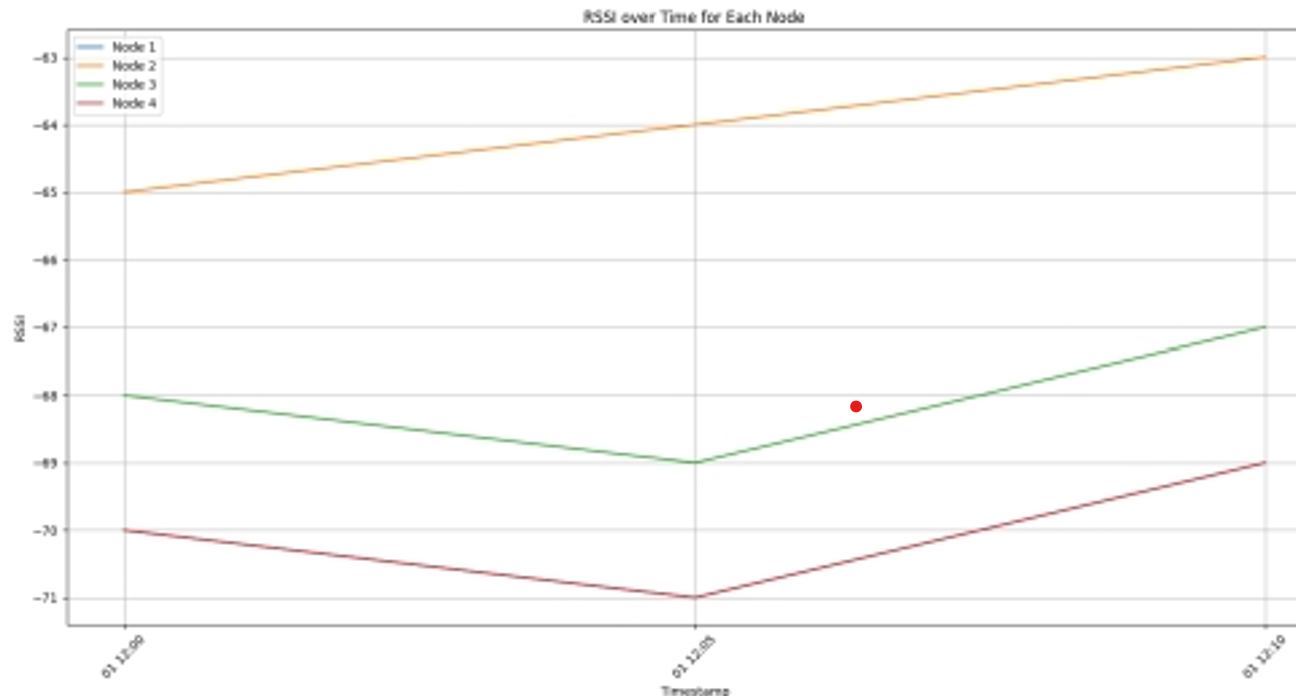


Fig. 8 RSSI over time.

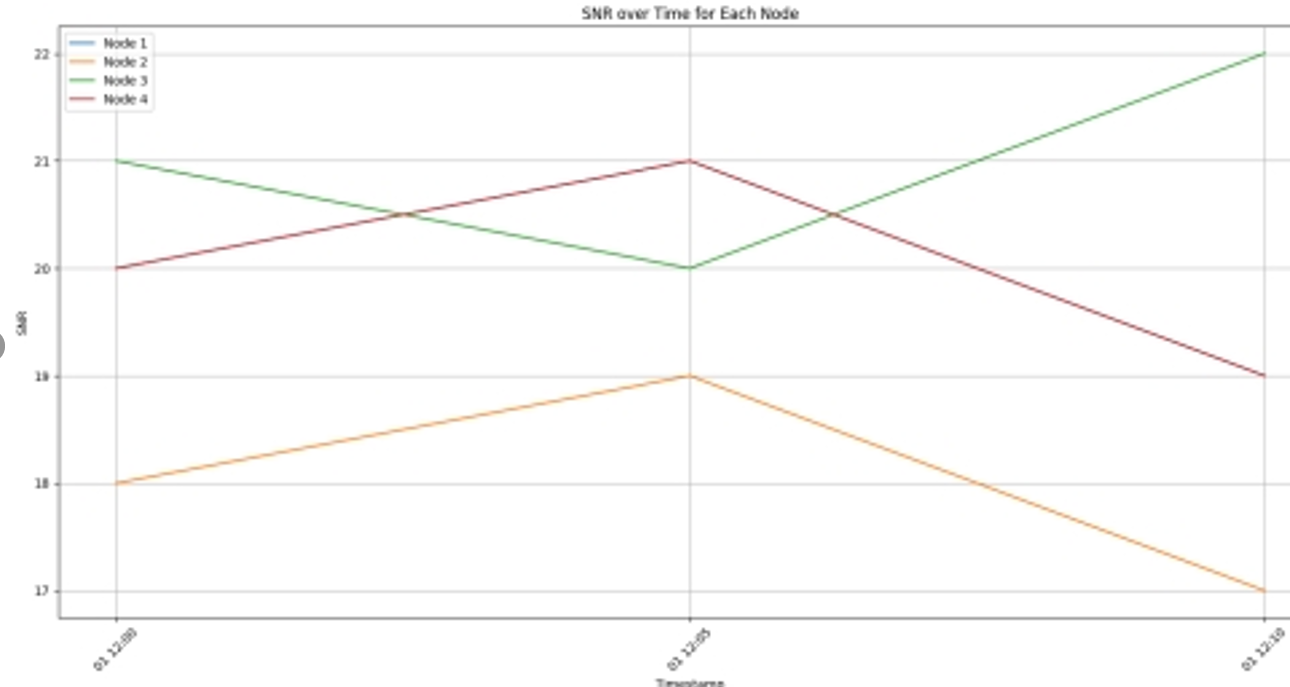


Fig. 9 SNR over time.

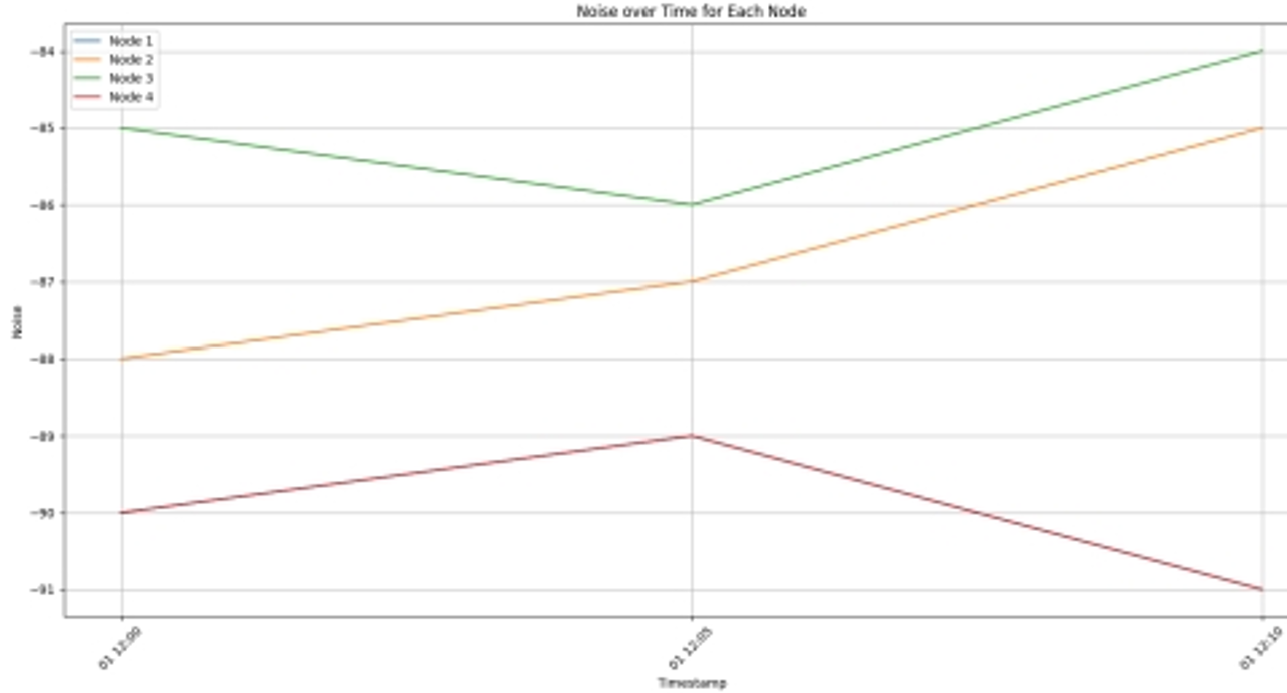


Fig. 10 Noise over time.

1. *Control and Notification*

The Control and Notification module is implemented with ChirpStack LoRaWAN for real-time monitoring the UVC air sterilizers. The functionality includes:

* + *Event Tracking:* Each event in the system is logged with a timestamp, indicating activities such as data transmissions and system checks. For instance, the system records at 2024-05-20 15:25:17 reporting successful uplinks, suggesting robust communication between the devices and the network server.
  + *Data Transmission:* The data columns indicate transmission of multiple data points, which likely include operational status, error codes, and possibly real-time UVC intensity levels. The data are crucial for ensuring the devices are operating safely and effectively, without exposing occupants to harmful UVC levels*.*
  + *Real-time Alerts:* Immediate notifications will be sent for warning any event of deviations from normal operating conditions, such as a drop in UVC intensity or system malfunctions. This proactive notification helps maintaining continuous sterilization effectiveness and immediate troubleshooting.
  + *System Logs and Updates:* The continuous logging of events ensures that all changes and incidents are recorded, allowing for detailed analysis and future improvements.

Fig. 11 illustrates an excerpt from the dashboard of control and notification module.

# CONCLUSION

This research underscores the potential of IoT technologies like LoRaWAN to revolutionize building management systems contributing to public health safety. The methodology demonstrates the effectiveness of using LoRaWAN cloud- based solution for remote monitoring and controlling of UVC air sterilizers by systematically addressing each phase, from design to evaluation. The proper system configuration, including the placement of LoRaWAN nodes and the tuning of signal-related variables such as RSSI, SNR, and Noise, potentially influence the performance and reliability of the communication links. The findings reported significant improvements in operational efficiencies and proactive maintenance capabilities that would lead to safer indoor environments.

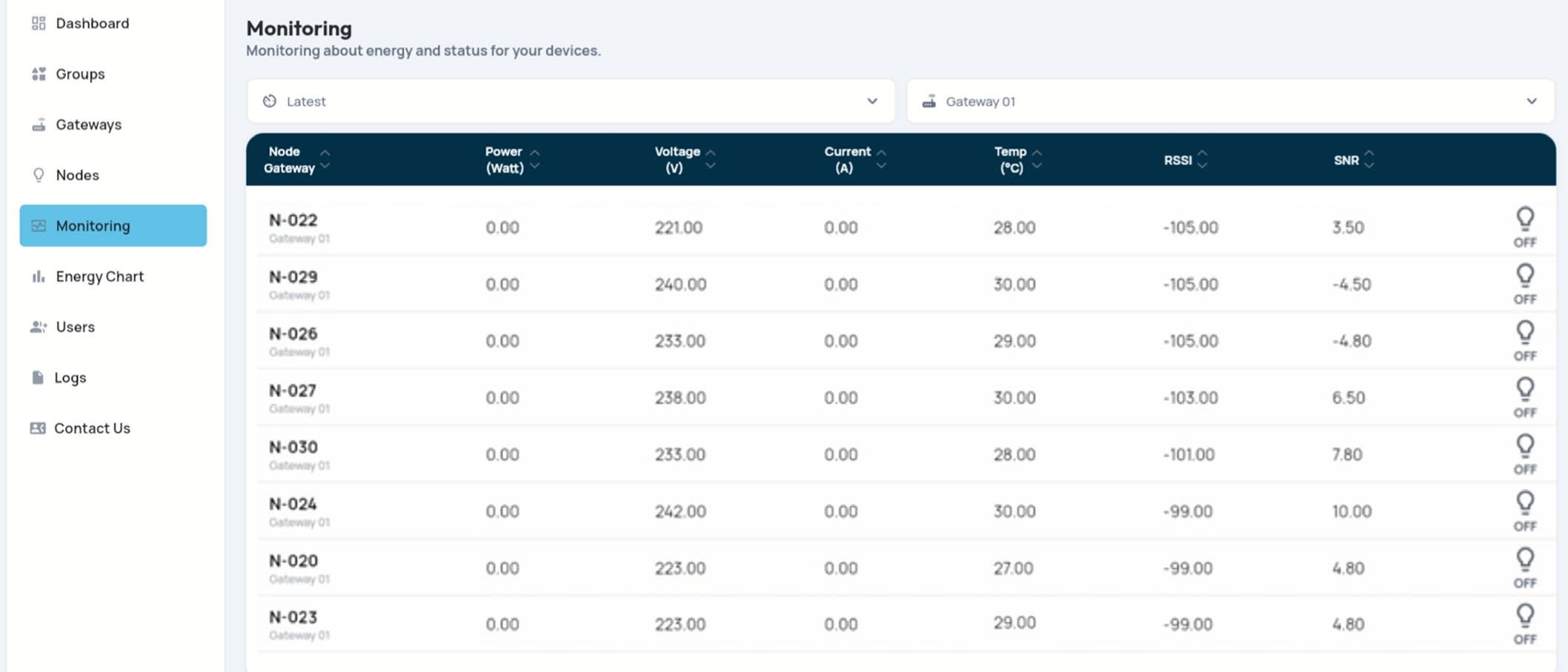


Fig. 11. Example dashboard of control and notification module.

The experiment successfully demonstrated the implementation and efficiency of a LoRaWAN-based control and notification system for UVC air sterilization in critical environments like hospitals and laboratories. Integrating LoRaWAN, a low-power, long-range communication technology, would enhance the ability to remotely monitor and manage UVC air sterilizers, ensuring sustainable air quality and reducing the risk of airborne infections.

Future direction would be optimizing the system architecture and extending the system's applicability to other types of environmental monitoring, in addition to exploring advanced data analytics to more accurately predict maintenance needs. For example, machine learning algorithms could provide deeper insights and predictive maintenance capabilities, further improving system performance and reliability.

REFERENCES

[1] T. G. G. M. O. P. Organization). "Thai Carbon Label." <https://www.tgo.or.th/2023/index.php/th/> (accessed January 15, 2025.

[2] O. Nexus. "OpenLCA Nexus Databases." <https://nexus.openlca.org/databases> (accessed January 15, 2025.

[3] I. P. o. C. C. (IPCC). "IPCC Emission Factor Database (EFDB)." <https://www.ipcc-nggip.iges.or.jp/EFDB/main.php> (accessed 15 January 2025).

[4] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," *Information processing & management,* vol. 41, no. 3, pp. 523-547, 2005.

[5] T. Grainger, D. Turnbull, and M. Irwin, *AI-Powered Search*. Simon and Schuster, 2025.

[6] M. Konda, *Elasticsearch in Action*. Simon and Schuster, 2024.