Information Processing and Retrieval on WSNs for Plant Monitoring Using Natural Language

Justin Miguel A. Co
School of Electrical, Electronics, and Computer Engineering
Mapúa University
Manila, Philippines
jmaco@mymail.mapua.edu.ph

James Ryan A. Justo
School of Electrical, Electronics, and Computer Engineering
Mapúa University
Manila, Philippines
jrajusto@mymail.mapua.edu.ph

Analyn N. Yumang
School of Electrical, Electronics, and Computer Engineering
Mapúa University
Manila, Philippines
anyumang@mapua.edu.ph

Dionis A. Padilla
School of Electrical, Electronics, and Computer Engineering
Mapúa University
Manila, Philippines
dapadilla@mapua.edu.ph

2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE) | 979-8-3503-1469-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ECICE59523.2023.10383105

***Abstract*—Technology has improved so that everyone can obtain the information they need from a single prompt. Most data is contained in a database, and exploring the data we need in these databases requires the language and the database, which most users do not know. We created an NLP system that allows farmers to access their databases' information with simple prompts and queries. By entering natural language queries into the system, they are converted to structured language queries to access the database information. We used pre-processing techniques such as tokenization, lemmatization, parts of speech tagging, and parsing on the natural language, and a rule-based algorithm was created to translate the query. 106 queries from random respondents were used to test the algorithm. The developed system showed an accuracy of 91.51%. We converted simple queries to structured language queries and provided output from the database.**

Keywords—natural language processing, WSNs, rule-based algorithm, POS-tagging, structured query language

I. Introduction

Currently, data or digital information is ever-increasing due to the amount of data that has been generated since the start of the digital era. Therefore, it is only reasonable for people to learn to extract this massive amount of data to analyze and manipulate. One must acquire knowledge from a technical perspective such as querying a database. One example is Wireless Sensor Networks (WSN). Database querying is essential as data acquisition from sensor nodes to the base station (server) often uses a query-based approach to obtain data from the database [1]. WSN is defined as a wireless network composed of dispersed sensor nodes used to obtain data together, forming a network [2]. Wireless technology provides mobility convenience, and is cost-effective, scalable, in a wider range [3] for enhanced communication [4]. They solve the problem of expensive wiring, a wide range of monitoring sites, and changes in the work environment [5]. Three different methods are used to obtain data from sensor networks: event-based, query-based, and periodic sensing. In database query systems, Natural Language Processing (NLP) translates natural language to query language to retrieve data from sensor networks [6]. To obtain the data, these applications support efficient queries, allowing network communication. A study showed that to efficiently combine identifiers and decrease the labor needed to manually match records, NLP precisely identifies the names and addresses of providers. This allows for the execution of intelligent outreach. [7]. NLP translates different human language queries into somewhat understandable machine instructions [8]. Hence, a system to translate human language to database query can be done for non-technical people concerning data acquisition on WSNs. The user does not need to learn an artificial language such as structured query language, training is not essential, and using a Natural Language Interface to Database (NLIDB) has its benefits.

We implemented a system to integrate NLP with WSNs (WSN) in terms of information processing and retrieval from the server of the base station. Since the accuracy level of translating from natural language to SQL query [6] is already ideal, we improved their system by adding new features and added to the application of a WSN topology for a plant monitoring system. The study's general objective was to translate English queries into SQL queries to obtain information about the physical conditions of the environment for the plants. We added a speech-to-text feature into the data retrieval using a speech-to-text (STT) library and implemented NLP operations to the system such as tokenizing, lemmatizing, removal of stop word, and Part-of-speech (POS) tagging. For the translated text, the system translated natural language to SQL query using a rule-based architecture and user-defined algorithms. We evaluated the precision, recall, F1 score, and accuracy of the system.

II. Methodology

Natural language was translated to query language focusing on the accuracy of text translation to query language in SQL. Deep learning algorithms increase the efficiency of big data access and the translation accuracy of existing algorithms. We optimized data retrieval on WSNs by adding speech-to-text and text-to-speech features. We applied the system to WSNs for plant monitoring systems in agriculture. Using parameters from a different table as an argument for the translated query or the complicated use of the English language, we created a command for the system to retrieve data from the MySQL server. The conceptual framework is presented in Fig 1. The system takes user input through English prompts/queries in either speech or text format entered through the graphical user interface. English query in speech is used to provide another form of input for the user. The algorithm then performs natural language pre-processing techniques to filter the query such as tokenization and lemmatization. In this study, parts of speech tagging and

parsing were performed to understand the query by the computer better. A user-defined algorithm was used to obtain the information from the query and translate it according to the parameters given in the query. The output was then provided, which was the translated SQL query and the database data returned by the SQL database after entering the SQL query.

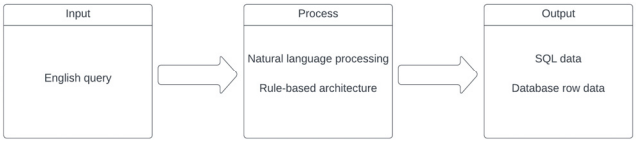

Fig 1. Conceptual framework.

A. Hardware Development

The block diagram of the wireless sensor node and the connections of all the sensor node devices are shown in Fig. 2. A WSN contained multiple sensor nodes [10] for which, in this study, four sensor nodes were built. The sensor node was composed of the ESP8266 ESP12-E as the microcontroller, a power supply, and the sensors: DHT11 sensor for detecting the temperature and humidity, BH1750 to detect the light level, FC-28 sensor which detects the soil moisture [11,12], and the MQ135 sensor which detects the presence of Ammonia, Mono-nitrogen oxides, Alcohol, Benzene, and carbon dioxide in the atmosphere [13,14]. Gas sensors detect different types of gases [15]. The power supply is connected to the sensors and the microcontroller. The sensors sent data to the ESP12-E microcontroller wirelessly to its router counterpart through a publish/subscribe protocol (Message Queuing Telemetry Transport, MQTT) hosted in the Raspberry Pi. The Raspberry Pi then served as a broker using Mosquito to obtain all the sensor nodes' data [13].

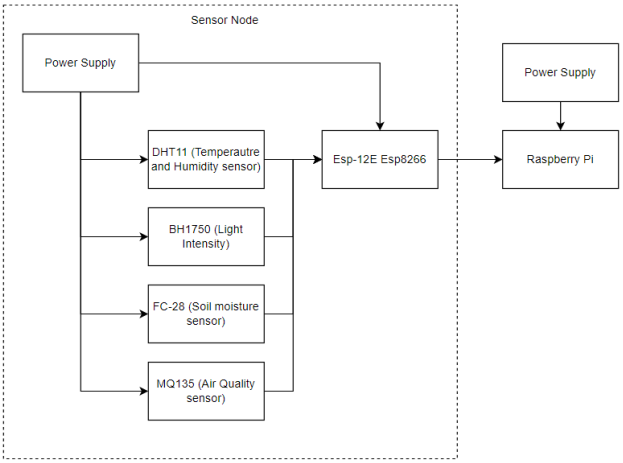

Fig 2. Block diagram.

B. Software Development

Figure 3 shows the flowchart of the system. The system started with an input to the Natural Language Interface to Database (GNLID). The query in text format was pre-processed for tokenization, removal of stop words, and lemmatization. The next step was to perform Parts-of-Speech (POS) tagging to assign speech tags to each word, such as nouns or verbs.

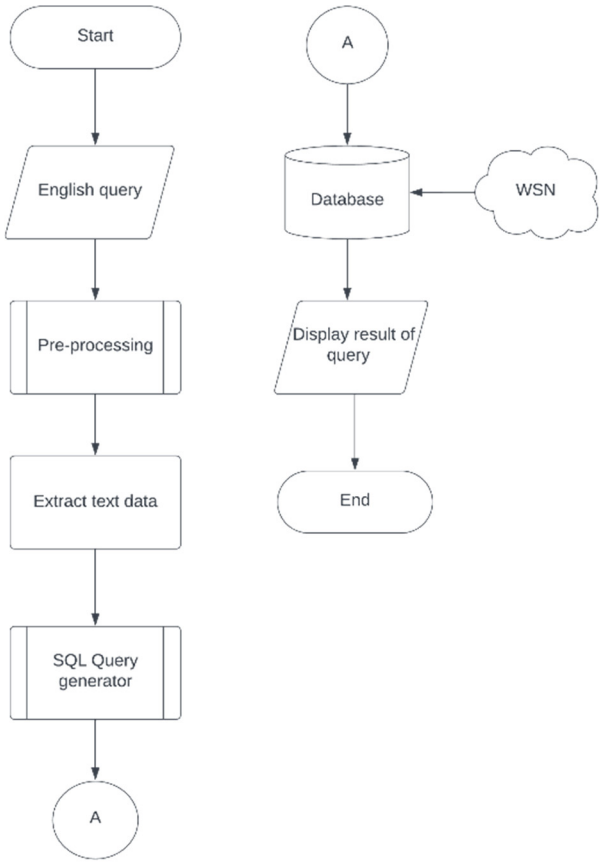

Fig 3. Program Flowchart of the Natural Language Interface to Database (GNLID)

Chunking was conducted to find the phrases containing the command, the indication of the parameter, and the statement of the date required if it existed. The chunked phrases were analyzed using the individual words with a rule-based algorithm to fit the specific needs of an agricultural plant database. The pre-processing module converted the text string to lowercase using the "lower()" function. It then removed punctuation using a combination of the NLTK pre-defined "replace()" function and Regular Expressions (RegEx). Next, the text was tokenized into smaller units called tokens using the "word_tokenize" module from the nltk.tokenize library. Stop words, which were irrelevant and occur frequently, were removed. Finally, the text data was lemmatized using the "lemmatize()" function from the WordNetLemmatizer class in the nltk.corpus. Stemming was not used as lemmatization provides more accurate results. This process refined the text, remove ambiguity, and sped up text data processing.

The SQL Query Generator module shows the query construction and refinement process in which a rule-based architecture is applied. The rule-based architecture uses various knowledge-based systems, using conditional statements to handle and execute logical decisions. This architecture evaluates the extracted entities and the user-defined schema in the figure. For this purpose, the process extracts the probable derivation of the input, which is a set of keywords or entities, and the user-defined schema, which contains the pre-defined table names, attributes, and clauses, was passed through the Classifier module to evaluate and map the supplied inputs. After mapping the inputs, the feasible text data in forming the SQL statement will be produced (table name, attribute, clauses, etc.).

The purpose of the Classifier module within the SQL Query Generator module is to evaluate and map the extracted entities and the user-defined schema from the SQL query generator module. Specifically, the inputs given to this module is processed as follows.

- Direct check – the inputs are checked and categorized into their corresponding SQL elements (table name, attribute, clauses, etc.)
- Concatenation check – the inputs are checked, if the direct check does not suffice, that is, the input is combined with the preceding or succeeding word, and then the combination is evaluated if it is a particular entity in the schema
- N-gram check – if the Concatenation check cannot rectify the problem, then the input undergoes more than two-word checks (3-to-many-word check) and is evaluated thereafter
- Synonym check – This is to determine if the extracted entities are synonymous with the elements in the user-defined schema; if so, it is replaced with the corresponding element. Otherwise, there is no operation.

Text data is processed and evaluated using either of the four methods. Tokens are compared with a data structure containing synonyms and a user-defined schema using if-else statements. If a match is found, the token is replaced with an SQL keyword and element. Unmatched tokens are discarded. The final output is in SQL format with SQL elements appended according to the user-defined schema. The system generates an SQL statement.

All of the above are designed for easy-medium type queries. However, for more complex queries, the steps above still apply with special keywords. The system uses keywords such as graph, average, and optimal. If a keyword is observed, it maps the corresponding SQL query along with the column and table name recognized by the previously mentioned algorithm. All transactions undergo a data class that maps the data model and the rows returned by the query from the database.

The application's front and back end are developed using the Kivy Python framework. Additionally, the database management system is MySQL, and the administrative tool for conducting the MySQL database is phpMyAdmin.

C. Experimental Setup

The experimental setup of the sensor node is shown in Fig 4, wherein the microcontroller, which is the ESP9266 ESP12-E, is connected to four sensors, which are the MQ135 air quality sensor, DHT11 temperature, and humidity sensor, the BH1750 digital light sensor, and the FC-28 soil moisture sensor the sensor is powered through a Universal Serial Bus (USB) adapter.

III. Results and Discussion

A. Data Gathering Process

The initial input data (before pre-processing) used will consist of one hundred six (106) English queries (via speech or text) as input to the GNLID obtained from a random sample of respondents. The datasets will also have two variations of accuracy level, which comprises whether the input data supplied is via speech or text.

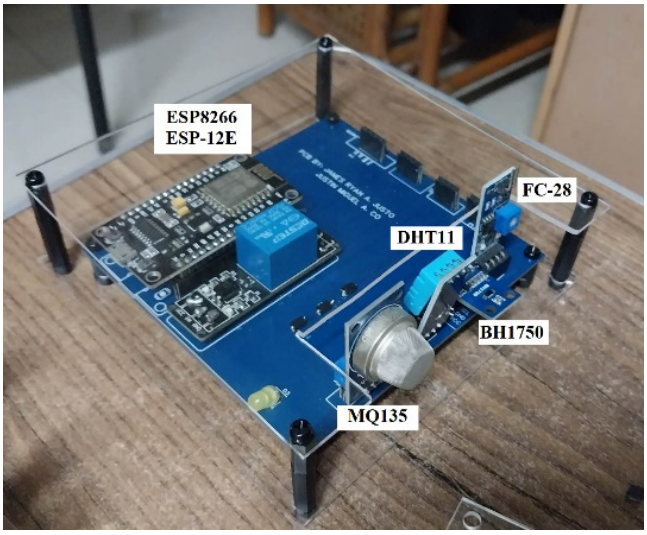

Fig. 4. Experimental setup of sensor node.

Table I shows the seven most common types of queries entered and translated in the testing with the rest having alterations in the parameters or use of words.

TABLE I. Expected SQL Query vs Translated Output

English query	Expected SQL Query	Translated Output
Show the temperature reading from the grape plant with temperature below 27 degrees.	SELECT Temperature, id, Date_n_Time FROM sensor_node_2_tb WHERE Temperature < 27	SELECT Temperature, id, Date_n_Time FROM sensor_node_2_tb WHERE Temperature < 27
Show the humidity reading from the corn plant with humidity below 50.	SELECT Humidity, id, Date_n_Time FROM sensor_node_4_tb WHERE Humidity < 50	SELECT Humidity, id, Date_n_Time FROM sensor_node_4_tb WHERE Humidity < 50
Show the light reading from the grape plant with light intensity above 200 lux	SELECT Light_Intensity, id, Date_n_Time FROM sensor_node_2_tb WHERE Light_Intensity > 200	SELECT Light_Intensity, id, Date_n_Time FROM sensor_node_2_tb WHERE Light_Intensity > 200
Show the soil moisture from the corn plant with soil moisture less than 50%	SELECT Soil_Moisture, id, Date_n_Time FROM sensor_node_4_tb WHERE Soil_Moisture < 50	SELECT Soil_Moisture, id, Date_n_Time FROM sensor_node_4_tb WHERE Soil_Moisture < 50
Show the soil moisture from the tomato plants with air quality more than 300 ppm.	SELECT Soil_Moisture, id, Date_n_Time FROM sensor_node_1_tb WHERE Air_Quality > 300	SELECT Soil_Moisture, id, Date_n_Time FROM sensor_node_1_tb WHERE Air_Quality > 300
Graph the temperature from the wheat plant.	SELECT Date_n_Time, Temperature FROM sensor_node_3_tb	SELECT Date_n_Time, Temperature FROM sensor_node_3_tb
Has the grape plant maintain good condition last year?	SELECT SUM(CASE WHEN p = 'TRUE' THEN 1 ELSE 0 END) AS count_temperature, SUM(CASE WHEN p2 = 'TRUE' THEN 1 ELSE 0 END) AS count_humidity, SUM(CASE WHEN p3 = 'TRUE' THEN 1 ELSE 0 END) AS count_soil, SUM(CASE WHEN p4 = 'TRUE' THEN 1 ELSE 0 END) AS count_air, SUM(CASE WHEN p5 = 'TRUE' THEN 1 ELSE 0 END) AS count_light	SELECT SUM(CASE WHEN p = 'TRUE' THEN 1 ELSE 0 END) AS count_temperature, SUM(CASE WHEN p2 = 'TRUE' THEN 1 ELSE 0 END) AS count_humidity, SUM(CASE WHEN p3 = 'TRUE' THEN 1 ELSE 0 END) AS count_soil, SUM(CASE WHEN p4 = 'TRUE' THEN 1 ELSE 0 END) AS count_air, SUM(CASE WHEN p5 = 'TRUE' THEN 1 ELSE 0 END) AS

B. Statistical Treatment

Among 106 total queries, 97 of the queries were correctly answered, resulting in an accuracy of 91.51%. Nine queries were incorrectly answered (8.49%), and zero queries were unanswered (Table II).

TABLE II. Overall Results of Translation

Total number of queries	Answered queries		Unanswered queries
106	Correctly answered queries	97 (91.51%)	0 (0.00%)
	Incorrectly answered queries	9 (8.49%)	

In the confusion matrix in Table III, among 106 total queries, 80 queries were correctly answered. Five were False Negatives, and four and seventeen were True Positives.

TABLE III. Confusion Matrix Result

106		PREDICTED	
		Positive	Negative
ACTUAL	Positive	80	5
	Negative	4	17

Precision is calculated using (1), which represents the percentage value of the correctness for the total number of queries. Precision is equal to (TP), which is the True Positive known to be the correctly translated queries divided by the sum of the True Positive (TP) and the False Positive (FP), which is known to be the incorrectly translated queries.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall was calculated using (2), which represents the percentage value of the correctness for the answered queries such that recall is equal to the (TP) True Positive divided by the (TP) True positive added to (FN) False Negative.

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

F1 score was calculated using (3). It is equal to precision multiplied by the recall divided by the precision added to the recall, with the result multiplied by 2.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (3)$$

The system's accuracy was computed based on (4), wherein the accuracy is equal to (TP+TN) the number of correct translated queries, which is the True Positive and True Negative all over (N) the total number of queries tested.

$$Accuracy = \frac{(TP+TN)}{N} \quad (4)$$

The total number of queries tested was 106. The performance metrics of the algorithm showed a precision of 94.12%, a recall of 95.24%, an F1-score of 94.67%, and an accuracy of 91.51%. The performance metrics are shown in Table IV.

TABLE IV. PERFORMANCE METRICS

Precision	0.9412
Recall	0.9524
F1-score	0.9467
Accuracy	0.9151

IV. CONCLUSION AND RECOMMENDATION

Using NLP rule-based architecture and user-defined algorithms, we translated simple natural language queries into SQL queries to access SQL databases. We applied a rule-based algorithm with NLP to convert natural language queries to SQL queries and entered them into an SQL database for agricultural data, wherein farmers entered simple queries into the system and obtained data from the sensor nodes stored in a database. Users queried through the Speech-To-Text feature (STT). We obtained a total accuracy of 91.51%. The algorithm could not process complex queries well, which needs to be improved by adding more code instructions. It is necessary to test and add functionalities for more queries of different needs. Applying the algorithm to different database applications, such as medicine, business, and education, is also recommended.

REFERENCES

[1] M. V. C. Caya, N. F. F. Losaria, M. J. C. Manzano, H. K. R. Tan, and R. V. Pellegrino, "Cashless Transaction for Resort Club Amenities Using RFID Technology," 2017.

[2] N. Ahmad, A. Hussain, I. Ullah, and B. H. Zaidi, "IOT based WSN for precision agriculture," *iEECON 2019 - 7th International Electrical Engineering Congress, Proceedings*, pp. 5–8, 2019, doi: 10.1109/iEECON45304.2019.8938854.

[3] J. R. Balbin *et al.*, "ZigBee and Power Line Communications interconnectivity applied to fuzzy logic controlled automated lighting system," in *Proceedings - 6th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2016*, Institute of Electrical and Electronics Engineers Inc., Apr. 2017, pp. 430–434. doi: 10.1109/ICCSCE.2016.7893612.

[4] R. V. Pellegrino, R. B. Lim, and A. N. G. Loceo, "Automated Wireless and Portable Measurement of Apnea-Hypopnea Index on Adult Patients with Obstructive Sleep Apnea Using Counter Based Algorithm," in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/HNICEM51456.2020.9400005.

[5] C. Tayo, N. D. Perez, and J. Villaverde, "Design and Development of a WSN for Water Quality Monitoring System of Shrimp Aquaculture," in *International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICECET55527.2022.9872980.

[6] S. Ali, H. Javed, A. Rauf, S. Mahfooz, and S. Khusro, "Natural Language Interface for Sensor Networks 1," vol. 19, no. 11, pp. 1563–1567, 2012, doi: 10.5829/idosi.wasj.2012.19.11.2470.

[7] M. Cook, L. Yao, and X. Wang, "A NLP Approach to Acquire Accurate Health Provider Directory Information," *Proceedings - 2018 IEEE International Conference on Healthcare Informatics Workshops, ICHI-W 2018*, pp. 76–77, 2018, doi: 10.1109/ICHI-W.2018.00027.

[8] A. N. Yumang, M. G. Abando, and E. P. M. De Dios, "Far-field Speech-controlled Smart Classroom with NLP built under KNX Standard for Appliance Control," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, ICST, Feb. 2020, pp. 219–223. doi: 10.1145/3384613.3384627.

[9] E. U. Reshma and P. C. Remya, "A review of different approaches in natural language interfaces to databases," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 801-804, doi: 10.1109/ISS1.2017.8389287.

[10] C. F. P. Abuda, M. V. S. Caya, F. R. G. Cruz, and F. A. A. Uy, "Compression of wireless sensor node data for transmission based on minimalist, adaptive, and streaming compression algorithm," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2018*, Institute of Electrical and Electronics Engineers Inc., Mar. 2019. doi: 10.1109/HNICEM.2018.8666320.

[11] A. N. Yumang, A. C. Paglinawan, L. A. A. Perez, J. F. F. Fidelino, and J. B. C. Santos, "Soil infiltration rate as a parameter for soil moisture and temperature based Irrigation System," in *Proceedings - 6th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2016*, Institute of Electrical and Electronics Engineers Inc., Apr. 2017, pp. 286–291. doi: 10.1109/ICCSCE.2016.7893586.

[12] J. L. C. Ison, J. A. B. S. Pedro, J. Z. Ramizares, G. V. Magwili, and C. C. Hortinela, "Precision Agriculture Detecting NPK Level Using a WSN with Mobile Sensor Nodes," in *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/HNICEM54116.2021.9732000.

[13] D. A. Padilla, G. V. Magwili, L. B. Z. Mercado and J. T. L. Reyes, "Air Quality Prediction using Recurrent Air Quality Predictor with Ensemble Learning," 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Manila, Philippines, 2020, pp. 1-6, doi: 10.1109/HNICEM51456.2020.9400051.

[14] M. V. C. Caya, A. H. Ballado, J. K. C. Asis, B. C. B. Halili, and K. M. R. Geronimo, "Air quality monitoring platform with the integration of dual sensor redundancy mechanism through internet of things," in *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment*

and Management, HNICEM 2018, Institute of Electrical and Electronics Engineers Inc., Mar. 2019. doi: 10.1109/HNICEM.2018.8666428.

[15] C. C. Hortinela, J. R. Balbin, J. C. Fausto, J. E. C. Espanillo, and J. K. P. Padilla, "Detection of Staleness in Raw Chicken Meat Due to Salmonella spp. And Escherichia Coli Bacteria Using Metal Oxide Gas Sensor with Support Vector Machine," in *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/HNICEM51456.2020.9399997 .