AIR QUALITY EVALUATION USING LOW-COST PORTABLE SENSOR NODE IN INDOOR ENVIRONMENT

by

IOANNIS DIAMANTAKIS

B.Sc. in Informatics University of Pireaus, 1998

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY

2023

Approved by:

Associate Professor
Spyridon Panagiotakis

Abstract

Air quality in large areas and especially in cities has been an issue of concern to the state and residents for decades, with the aim of ascertaining the situation and improving any problems that may arise, having developed common standards (AQI), surveillance and recording systems for decades. On the other hand, air quality in the places where we live, work, study, is not a key priority nor are there commonly accepted indicators by which the air of an enclosed space can be characterized as clean or harmful. Also, it is a fact that enclosed spaces differ significantly in air quality even, depending on their use even if they are very close. This creates the need for a large number of measuring stations and necessarily cheap ones to cover all spaces.

In this thesis, a record is made of the methods of air measurement, the elements they take into account and a method of assessing indoor air quality (IAQ) is proposed, taking into account the burden caused by human presence.

In addition, a device that uses the ESP32 microcontroller and a set of low cost, high accuracy sensors is designed, developed and tested. Its main purpose is to sense and send raw data of air metrics, using MQTT protocol, to a timeseries database for further processing. This device could be used to create a fog computing/sensor group, which could extend in a house or a larger building. It also proposes and develops a way of central monitoring (Grafana) of measurement stations and proposes a method of calculating the air quality index, with a centralized system and the use of machine learning algorithms like K-NN, that calculate the IAQ index of each station thus the room that is placed in, informing the inhabitants.

The proposal is interesting and shows prospects for implementation. In the future, it could be expanded by adding automatic positioning of a measuring station, immediate suggestions for improving air quality depending on the pollutant that is increased, or controlling devices that could improve air quality.

Table of Contents

Abstract iii

Table of Contents ii

List of Figures vi

List of Tables viii

Acknowledgements ix

Dedication x

Chapter 1. Introduction 12

1.1 Motivation 12

1.2 Objectives 13

1.3 Organization of Thesis 14

Chapter 2. Standards and Related work 16

2.1 Air Quality Indices 16

2.1.1 PSI Air Quality Index by EPA 16

2.1.2 Sydney RPI index 18

2.1.3 Bristol Air Pollution Indicator - BAPI 18

2.1.4 Indian Air Quality Index- INDAQI 19

2.2 Related methods of pollution measuring and work 19

2.3 Types of sensors 21

2.3.1 Electrochemical sensors 21

2.3.2 Metal oxide sensors 22

2.3.3 Photoelectric Detection Sensors 23

2.3.4 Pellistor 24

2.3.5 Field effect sensors 25

2.3.6 Piezoelectric sensors 26
2.3.7 Particle matter sensors 26
2.4 Machine Learning 27
2.4.1 Machine learning categories 27
2.4.1.1 Supervised learning 27
2.4.1.2 Unsupervised learning 28
2.4.1.3 Remedial Learning 28
2.4.2 Categories of machine learning algorithms 28
2.4.2.1 Classification 28
2.4.2.2 Regression 29
2.4.2.3 Clustering 29
2.4.3 Classification algorithms 29
2.4.3.1 K-NN algorithm 29
2.4.3.2 SVM algorithm 30
Chapter 3. Designing the system 33
3.1 General diagram of operation 33
3.1.1 Perception 33
3.1.2 Connectivity 34
3.1.2.1 MQTT 34
3.1.3 Data accumulation 35
3.1.3.1 InfluxDB 35
3.1.4 Visualization and data processing 35
3.1.4.1 Grafana 36
3.1.4.2 NodeRed 36
3.2 IAQ calculation technic 37
3.3 Operational states 40
3.4 Flow chart of data exchange 42

Chapter 4. System implementation....43
4.1 Selecting sensor components....43
4.1.1 Nova SDS011....43
4.1.2 Plantower PMS7003....44
4.1.3 Sensirion SPS30....45
4.1.4 Honeywell HPMA115S0....46
4.1.5 DHT22....47
4.1.6 Sensirion SGP30....48
4.1.7 CJMCU8128....48
4.1.8 Winsen ZE25-O3....50
4.1.9 MiCS Sensor working principles....51
4.1.10 CJMCU4541....54
4.1.11 CJMCU6814....55
4.2 Sensors Software Libraries....56
4.2.1 Βιβλιοθήκη CJMCU4541....57
4.2.2 Βιβλιοθήκη CJMCU6814....58
4.2.3 Βιβλιοθήκη CJMCU8128....60
4.3 IAQ Measuring station....60
4.3.1 Connection diagram....61
4.3.2 Software....61
4.3.3 Setting properties....62
4.4 Data storing....62
4.4.1 InfluxDB measurements....62
4.5 Calculating the IAQ Index using ML....64
4.5.1 Data Description....64

4.5.2 Data Preparation 65
4.5.2.1 Missing values 65
4.5.2.2 Removing Outliers 65
4.5.2.3 Export features 67
4.5.3 Algorithm Selection 68
4.5.3.1 K-nn evaluation 68
4.5.3.2 SVM evaluation 69
4.5.3.3 Results 70
4.5.4 ML Application in Node-Red 71
4.6 Displaying timeline data and geolocation of measuring station 72
4.6.1 Measurement period 72
4.6.2 Commentary on recordings 73
Chapter 5. Evaluation 78
5.1 Particulate Matter Sensors Experiment 78
5.2 Testing the rest sensors 82
Chapter 6. Conclusion 84
References 86
Appendix A - Software used during the development of this thesis and for the operation phase of the system 91

List of Figures

Figure 2-1 Health hazard and air Quality. Source EPA 16
Figure 2-2 Division of Air Quality. Source: EPA, 2012 17
Figure 2-3 Example of categorization with the K-NN algorithm (Wikipedia, n.d.) 30
Figure 2-4 SVM Concept (S. Raschka, 2017) 31
Figure 3-1 Generic system representation 33
Figure 3-2 Stages of operation and procedures of the measuring station 41
Figure 3-3 Flow chart of data routes 42
Figure 4-1 Sensor's proper placement 44
Figure 4-2 Nova SDS011 44
Figure 4-3 Plantower PMS7003 image and flow chart 45
Figure 4-4 Sensirion SP30 46
Figure 4-5 Honeywell HPMA115S0 operation description 46
Figure 4-6 DHT22 image and electrical connection diagram 47
Figure 4-7 Sensirion SGP30 image and operation flow chart 48
Figure 4-8 CJMCU8128 front and back image 49
Figure 4-9 CCS811 block diagram 49
Figure 4-10 BME280 block diagram 50
Figure 4-11 HDC1080 block diagram 50
Figure 4-12 ZE25-O3 image and pin numbering 51
Figure 4-13 MiCS operation cicruit 51
Figure 4-14 MiCS measurement voltage 52
Figure 4-15 Numeric evaluation 53
Figure 4-16 CJMCU4541 pinout diagram 54
Figure 4-17 RED sensor, continuous power ON, 25°C, 50% RH 55
Figure 4-18 OX sensor, continuous power ON, 25°C, 50% RH 55
Figure 4-19 CJMCU6814 operation cicruit 55
Figure 4-20 RED sensor, continuous power ON, 25°C, 50% RH 55
Figure 4-21 OX sensor, continuous power ON, 25°C, 50% RH 56
Figure 4-22 NH3 sensor, continuous power ON, 25°C, 50% RH 56

Figure 4-23 CJMCU4541 electrical circuit 57
Figure 4-24 CJMCU6814 electrical circuit 59
Figure 4-25 Sensor node electrical diagram. 61
Figure 4-26 Missing dataset values 65
Figure 4-27 PM values 66
Figure 4-28 O3, CO and NO2 values 66
Figure 4-29 CO2 and VOC values 66
Figure 4-30 Pearson correlation chart 67
Figure 4-31 Performance of knn on testing phase 68
Figure 4-32 Performance of knn on various testing percentages 69
Figure 4-33 Performance of Linear SVM 70
Figure 4-34 Performance of polynomial SVM on Testing 70
Figure 4-35 Calculating IAQ index 72
Figure 4-36 PM particle concentration recording, winter measurement period 73
Figure 4-37 CO2 and VOC concentration recording, winter measurement period 74
Figure 4-38 Ozone concentration O3, winter measurement period 74
Figure 4-39 Ozone O3 concentration, summer measurement period 75
Figure 4-40 NO2 concentration, summer measurement period 75
Figure 4-41 CO concentration, summer measurement period 75
Figure 4-42 IAQ value chart 76
Figure 4-43 Position of the sensor 76
Figure 4-44 Properties of the World Map add-on 77
Figure 5-1 Electrical circuit for the expirement sensor station 79
Figure 5-2 Block diagram of the experiment connections 80
Figure 5-3 Recorded values of different sensors 80
Figure 5-4 Scatter chart and trendlines 81
Figure 5-5 Connection diagram of the testing device 82

List of Tables

Table 2-1 Daily Value Threshold for AQI. Franceschini et al, 2005 17
Table 3-1 Proposed limits of each IAQ category..... 37
Table 3-2 Example average metrics..... 39
Table 3-3 Seperate indeces of pollutants and resulting IAQ 40
Table 4-1 Sensors, their type and the pollutants they detect..... 43
Table 4-2 Measuring station's components and cost 60
Table 4-3 Sensor station Data package..... 63
Table 4-4 Measurements and metrics in InfluxDB..... 64
Table 4-5 Part of dataset used to train ML model, IAQ included..... 64
Table 4-6 Dataset's variable description 67
Table 5-1 Experiment phases..... 79

Acknowledgements

I would like to express my graditude to the Professor Panagiotakis Spyros for the trust and patience he has shown towards me. His encouragement was instrumental in completing this work.

Dedication

This thesis is dedicated to Yianna, Giorgos, Katerina and Manolis.

Chapter 1. Introduction

1.1 Motivation

By "indoor air quality" we refer to the air in the interior of homes, workplaces, schools and any other indoor environment of buildings. Indoor air quality is influenced by pollutants, the majority of which come from sources within buildings; and some come from the external environment [1]. The most common pollutants are products combustion products such as carbon monoxide, particulate matter, combustion smoke, particulate matter, smoke from the cigarette smoke, building materials and building equipment. They also include substances radon, mould, lead, ozone and a host of other products. In recent decades, there has been an increase in the concentrations of pollutants in the in building interiors due to the use of synthetic construction materials, furniture and other materials etc [1].

The indoor room climate plays an important role for human wellbeing. The indoor air quality is worsening because of odors usually belong to volatile organic compounds (VOCs) which include gases emitted by humans. Those are not sensed by smell but only using sensor technologies. Additionally, the rapid spread of SARS-COV-2 has created many problems in humans and preliminary indications suggest that high concentrations of particles and CO_2 favor its spread especially indoors [2].

Buildings that have ventilation/air conditioning systems could automatically control the airflow in the building to reduce such gases but majority of buildings lack of those systems. The issue is constantly brought back to the fore because air, in addition to contamination by fine suspended dust particles (PM), contains gases such as O_3, NO_2. In small or larger concentrations can be toxic to the respiratory system of limiting the development of their lungs and creating more problems in their adult lives [3] [4].

The concentration of these gases depends on the location, whether it is a small or large city where human activity such as car traffic and the operation of craft/industrial units is significant [5] [6].

Data recording in every area of human activity using public measurement systems is not always possible because of their cost. In these cases, the solution can come with the

construction of low-cost air quality sensors which can be installed and managed with fewer resources and provide an opportunity to accelerate access to air quality information [7].

In recent years, such systems have been launched and gradually expanded, exploiting the potential of IoT technologies such as the Helsinki HOPE project [8], the Breathe London program IoT concerning the installation of corresponding sensors outdoors in cities.

1.2 Objectives

This thesis' main objective is the construction of a prototype portable measuring station for microparticles of size PM2.5 and PM10, other gaseous pollutants such as CO_2, NO_2, CO, O_3 and temperature and relative humidity conditions. The station will be able to record and send the above measurements to a time series database (InfluxDB), gathering them in a central data computing system for visualization and further processing.

The selection of the most suitable sensors will be made according to criteria such as:

1. market availability
2. reasonable cost of purchase to keep low the total development cost
3. comparison of their technical characteristics and operating specifications
4. their performance in a parallel trial period to determine which is the most accurate
5. the sensor manufacturer has a presence in the domain for several years to ensure support with operating libraries

The node will consist of a low-cost computing unit powered by a rechargeable battery. The data will be collected through communication protocols, depending on each sensor support (Serial, I2C, SPI) and send to the database using IEEE 802.11 network protocol.

Data visualization will be implemented using either InfluxDB's embedded tools. Data processing includes the interpretation of raw data to an air quality index (AQI) depending on sensor measurements, using a machine learning model that will be developed for this purpose. Those calculations could be implemented locally on every node concerning a short period or/and on the server side of the system, separately for each node. Each sensor, apart from displaying its' measurements, will also show AQI as room's state.

With this research we will be able to demonstrate that there are conditions which allow the construction of reliable low-cost sensors, measuring indoor air quality in order to improve it but also, to decrease the possibility of virus infections. Field of testing and application would be school buildings where the assessment of air quality, inside and outside classrooms, is an open issue of concern to the scientific community [9] [10] with direct effect on the health of pupils and teachers.

1.3 Organization of Thesis

The structure and content of this work is briefly described in the following paragraph in order to give a concise outline.

Starting with the chapter *Standards and Related work*, four key ambient air quality indicators widely used in different countries are presented and how their values are calculated and the quality ranges are described.

Below is a brief description of the methods of measuring air pollution and presents the categories of sensors that can be used for the purpose of this work. A comprehensive report is made of how they work, their positive and negative points, in order to understand the current capabilities and technical frameworks for pollutant detection.

In the following, recent implementations of similar devices and devices for measuring air quality are presented, in order to describe the design and development background of this work and its added value.

At the end of the chapter, basic points of Machine Learning are explained, relevant categories of algorithms that can be used such as classification, regression and clustering and how two algorithms, Knn and SVM, that were used during the implementation of this work are presented.

In the chapter *Designing the system* stages of data collection, flow to the storage system and their processing in order to monitor all sensors and calculate the Indoor Air Quality Index (IAQ). The calculation algorithm and characterization categories of the IAQ used in this work and the algorithm of operation of the measuring station;

In the *System implementation* chapter, there is a detailed presentation of all parts of the system that concern both hardware and software. There is a detailed presentation of

the technical characteristics of the selected sensors and software libraries that were incorporated into the code of the measuring stations.

The air quality measurement station is described, which is the main point of data collection, i.e. the connection diagram, its operating parameters and the operating phases of its software are presented.

Also, the process by which the machine learning model through which the IAQ air quality index is calculated is analyzed. The selection and preparation of the necessary ML dataset is described and the performance of two classification algorithms is compared in order to select the most appropriate model for the ML.

Finally, the system for monitoring the measuring stations, the settings and projection of its recordings is presented and some diagrams are commented on in order to highlight their usefulness.

In the Evaluation chapterEvaluation, PM sensors are evaluated through an experimental device developed and implemented in order to find the most accurate one to be used at the measuring station, while the benchmark was the DustTrack II Aerosol Monitor. Also, a similar comparison of the other sensors is presented, by recording their measurements during their periods of operation in various places and months. Data is annotated their problems, the problems that arose but also their positive elements, which led to the selection of some of them to be used in the measuring station.

In closing, the chapter *Conclusion* analyzes the conclusions that emerged from this work, regarding the feasibility of implementing a complete system for recording, calculating and presenting indoor air quality. The proposal of a similar IAQ indoor air quality index and its calculation using a machine learning model is a key element of this work. In conclusion, proposals are made for future expansion and study in this area.

Chapter 2. Standards and Related work

2.1 Air Quality Indices

2.1.1 PSI Air Quality Index by EPA

The PSI standard air pollution index was revised and transformed by the US Environmental Protection Agency (U.S. EPA) in 1999 and renamed the Air Quality Index (AQI). The AQI is calculated for each pollutant separately, based on the following equation:

$$AQI = \frac{(AQI_{Hi}) - (AQI_{Lo})}{(Conc_{Hi}) - (Conc_{Lo})} \times \left((\mathbf{Conc_i}) - (Conc_{Lo})\right) + (AQI_{Lo})$$

Where
$Conc_i$ = Input concentration for a given pollutant
$Conc_{Lo}$ = The concentration breakpoint that is less than or equal to $Conc_i$
$Conc_{Hi}$ = The concentration breakpoint that is greater than or equal to $Conc_i$
AQI_{Lo} = The AQI value corresponding to $Conc_i$
AQI_{Hi} = The AQI value corresponding to $Conc_i$

The purpose of the AQI Index is to help understand the potential public health impacts of air pollution at the local level. For this purpose it is divided into six different categories of ambient air quality which in turn are represented by six different colors, as shown below, corresponding in color to the potential health effects (US EPA).

AQI Value	Health Message	AQI Color
0 - 50	None	Green
51 - 100	Unusually sensitive people should reduce prolonged or heavy exertion	Yellow
101 - 150	Sensitive groups should reduce prolonged or heavy exertion	Orange
151 - 200	Sensitive groups should avoid prolonged or heavy exertion; general public should reduce prolonged or heavy exertion	Red
201 - 300	Sensitive groups should avoid all physical activity outdoors; general public should avoid prolonged or heavy exertion	Purple
301 - 500	Everyone should avoid all physical activity outdoors	Maroon

Figure 2-1 Health hazard and air Quality. Source EPA

AQI Value	AQI Category	AQI Color
0 - 50	Good	Green
51 - 100	Moderate	Yellow
101 - 150	Unhealthy for Sensitive Groups	Orange
151 - 200	Unhealthy	Red
201 - 300	Very Unhealthy	Purple
301 - 500	Hazardous	Maroon

Figure 2-2 Division of Air Quality. Source: EPA, 2012

Franceschini et al. [11] give the concentration values with their upper and lower limits, for each pollutant separately as well as the corresponding upper and lower values of the AQI air pollution index. After calculating the individual indices, for each pollutant separately, the daily value of the AQI index is considered to be the highest of the individual values.

AQI	0-50	51-100	101-150	151-200	201-300	301-500
PM_{10} $\mu g/m^3$ (24 hr)	0-54	55-154	155-254	255-354	355-424	425-604
$PM_{2.5}$ $\mu g/m^3$ (24 hr)	0-15.4	15.5-40.4	40.5-65.4	65.5-150.4	150.5-250.4	250.5-500.4
SO_2 ppm (24 hr)	0-0.034	0.035-0.144	0.145-0.225	0.225-0.304	0.305-0.604	0.605-1.004
CO mg/m^3 (8 hr)	0-4.4	4.5-9.4	9.5-12.4	12.5-15.4	15.5-30.4	30.5-50.4
O_3 ppm (8 hr)	0-0.064	0.065-0.084	0.085-0.104	0.105-0.124	0.125-0.374	(*)
O_3 ppm (1 hr)	-	-	0.125-0.164	0.165-0.204	0.205-0.404	0.405-0.604
NO_2 ppm (1 hr)	-	-	-	-	0.65-1.24	1.25-2.04

Table 2-1 Daily Value Threshold for AQI. Franceschini et al, 2005

There are seven sub-indicators. The daily value of the AQI index is the maximum value of these seven sub-indices, based on the ratio:

$$AQI = \max(I1, I2, I3, I4, I5, I6, I7)$$

Where I1 and I2 are the sub-index based on the mean 24-hour concentration values of the suspended particulate matter with an aerodynamic diameter of less than 10μm (PM10) and 2.5μm (PM2.5) respectively, I3 the sub-index based on the average 24-hour sulfur dioxide concentration values, I4 the sub-index based on the average 8-hour carbon

monoxide concentration values, I5 the sub-index based on the average 8-hour ozone concentration values, I6 the sub-index based on hourly ozone concentration values and I7 the sub-index based on hourly nitrogen dioxide concentration values. In order to calculate the index, it is required that are available data of at least three indeces and one of them should be a PM index.

2.1.2 Sydney RPI index

In Sydney, Australia, ambient air quality has been monitored since the 1960s, through the public authorities such as NSW Environment Protection Authority 1998, NSW Department of Health 2006 and NSW Department of Environment and Conservation 2006. Since the early 1980s, daily air quality reports have been issued daily, with an emphasis mainly on ozone pollutants and solid particulate matter. In 1993, following a more general review of air quality monitoring needs, the Government of New South Wales (NSW) upgraded its network.

The RPI is calculated twice a day. The morning report covers the period from 15:00 in the afternoon of the previous day, until 06:00 in the morning of the day the morning report is made. The second report of the value of the index covers the period from 06:00 am to 15:00 pm of the same day. In Sydney, Australia, daily reports of RPI values are based on measurements of ozone and nitrogen dioxide concentrations. The RPI index, for each pollutant separately.

RPI = Pollutant Concentration / Concentration Value Limit X 50

The RPI index can be categorized based on its values as low, medium and high, which means good, moderate and poor air quality, respectively.

More specifically, when the index takes values less than 25, it practically means that none of the measured pollutants show concentration values greater than 50% of their limit value.

2.1.3 Bristol Air Pollution Indicator - BAPI

The City of Bristol in the United Kingdom uses a Bristol Air Pollution Index (BAPI), a Bristol Air Pollution Index (BAPI), based on a number of studies on the effects

of air pollution on public health (Bristol City Council Air Quality 2006). The BAPI index is calculated based on the concentrations of air pollutants NO_2, SO_2, CO, O_3 and PM_{10}.

Specifically, the value of the index is calculated for each of the five pollutants examined and the daily value of the BAPI index is considered to be the highest of its five values, calculated in the total of the pollutants examined.

2.1.4 Indian Air Quality Index- INDAQI

The Government of India, in New Delhi, in an effort to measure air quality, has funded a project developed by the city's Pollution Control Center. Thus was created the Indian Air Quality Index (INDAQI). This index is based on roughly the same logic as the US Air Quality Index AQI, already mentioned [12]. The concentration values of air pollutants monitored and recorded by the National Atmospheric Air Quality Monitoring Program of India, concern the pollutants NO_2, SO_2, CO, O_3 and PM_{10}.

To determine the daily value of the index, it is necessary to record the concentration of at least three of the five pollutants under study. For each pollutant a sub-indicator is calculated based on the limits set for its concentration values. The daily value of the INDAQI index is considered to be the highest value of those of the individual indices.

2.2 Related methods of pollution measuring and work

Low-cost air quality monitoring systems are a good solution to the difficulties of setting up a dense and extensive pollution monitoring network with expensive instruments. The data, based on their established flow to date, are collected mainly by monitoring stations in accordance with the legislation and international standards of methods. The installation, operation and supervision of such stations is done by governments, industry and some specialized research organizations.

To estimate the levels of gaseous pollution, various methods have been developed for the sampling and calculation of the concentration of pollutants in the atmosphere. Especially for suspended particles, the various techniques are based on filtration, impact, thermal and electrostatic properties. The reference methods developed are based on the gravimetric determination of the suspended particles. However, in addition to gravimetric determination, methods for continuous recording of the concentration of suspended

particles have been developed, which have the advantage of giving the concentration of PM as a function of time.

The use of these automated recording methods is constantly expanding, but they are not as accurate as gravimetric analysis methods. Low-cost particle measurement techniques are based on optical measurements, which typically use a low-power light source (infrared LED or laser), where particles passing through the device through forced flow scatter light measured by a light-detecting device (light), usually a light emitting diode [13].

Another issue that concerns researchers is the performance of low-cost sensors, whether the measurements they give are accurate and whether they are affected by the climatic conditions of the space in which they are located. In the research of Demanega et. al. [14] The performance of sixteen sensors is evaluated through their placement and operation in an experimental room, under controlled ventilation conditions and temperature in order to achieve the test conditions (heat & high humidity, cold & low humidity). By implementing eight experiments that simulated different human activities, their response was tested when measuring PM2.5, PM10, CO_2, TVOC. The experiments highlighted the reliability of some sensors such as SCD40 and SHT31 in pollutant measurements, while no significant performance difference was observed compared to different climatic conditions.

The work of Martinez A. et.al. [15] is a proposal for the implementation of low-cost pollutant measurement systems, so that their implementation and use is economically feasible. Using the Arduino Mega 2560 rev3, components for wireless networking, GPS positioning, batteries, photovoltaic panels for charging, and four sensors (PM2.5, PM10, CO_2, SO_2, O_3, Humidity, Temprature) for measurements, they built a device for no more than $400. Following the same reasoning, Lopez et.al. [16] compare low-cost sensors with equivalent commercial devices, over a period of seven months, indoors. In summary, their conclusions are that there is a burden on air quality depending on the space and activity of people, which can be measured with sufficient accuracy with economic sensors, which showed R^2 values from 0.87 to 1.

Particularly important are the conclusions of the research of Narayana et.al. [17] on the long-term evaluation of low-cost sensors which, as they point out, is lacking, while a specific framework describing calibration procedures has not been identified. A

disadvantage is that their measurements are affected by the environmental conditions of the installation area, resulting in calibration during their construction to reduce their accuracy, if not repeated at regular intervals. This could increase their uptime, which is relatively short and does not exceed 18 months. It is important to evaluate them by comparing the transmutations with reliable sensors, for each type of pollutant.

2.3 Types of sensors

2.3.1 Electrochemical sensors

The technology utilized by this type of sensor is one of the oldest and is widely used for measurements of gaseous pollutant concentrations. separated by some kind of electrolyte. The electrode at which the oxidation or reduction of the gas to be measured takes place, is called the working electrode. The concentration of the respective gaseous pollutant is determined by measuring either the potential difference or the current intensity between the electrodes. The electrochemical process is due to the transfer of charge from one electrode to another. During the operation of the sensor, the gas for which it is made and which is located near the working electrode is ionized and through the special electrolyte that is in between the ions are led to the cooperating electrode located on the opposite side. In this way an electric current is created by the measurement of which it is possible to determine the concentration of the gas. The choice of material of the working electrode, the type of electrolyte to be used as well as the potential difference between the two electrodes determine the type of gas that the sensor will monitor. These types of sensors according to the presentations of [18] and [19] have the following advantages:

- They have relatively good selectivity that is sufficient for several applications
- They have quite good sensitivity as they can determine concentrations from 1 to 10 ppm (parts per million)
- They are ideal for mobile devices as they are quite compact (usually have a typical dimension of about 20mm) and do not require large power consumption (most of the energy they use is for signal amplification)

- They have fast response which is linear and limited only by the diffusion of the gas into the working electrode because due to the diffusion the rate at which the gas molecules react with the electrode surface is reduced.

Some disadvantages include:

- Short lifespan, ranging from a few months to a maximum of 2-3 years. This is due to the consumption of the electrolyte, especially in cases where the gas appears in high concentrations or when it is constantly present
- They can not take measurements in an environment with high temperatures because the electrolyte dries
- They are expensive compared to other sensors and this is due to the complex how they are made
- They do not have much room for improvement. This is primarily due to the fact that by reducing the size of the sensors their lifespan is limited and secondly their construction is automated with the result that no further reduction of the construction cost can be achieved.

2.3.2 Metal oxide sensors

The basic principle of operation of this type of sensor according to the presentations of [18] and [19] is that during the absorption / reaction of the gas for which this sensor is made, with the corresponding semiconductor layer of the sensor causes changes in the electrical properties of the latter with the main change in its conductivity. Formerly they were made as solid elements consisting of the heat source placed inside a ceramic housing around which was the metal oxide semiconductor in which the reactions with the surrounding gases took place.

For the measurement, electrodes on the surface of the semiconductor were used. Today, with the advancement of technology, their construction is mostly done by printing microelectronic circuits on silicon plates, thus allowing them to become smaller in size and cheaper in their construction, since in this way mass production is achieved. The materials for the construction of semiconductors, so in order for the sensor to work, it must be

sensitive to various gases, for which either organic semiconductors or more commonly metal oxide semiconductors with the addition of special catalysts are used [13].

The most commonly material used for the construction of metal oxide semiconductors is Tin oxide as it reacts with a number of gases and the changes it shows in its resistance are intense. Aluminum oxide with dispersed platinum elements acting as heaters and electrodes is commonly used as the substrate material. To accelerate the reactions that occur on the surface of the semiconductor, usually a heat source is installed in the sensor in the ways mentioned above [19].

The most widely accepted operating model of this technology is that tin oxide forms granules that affect the conductivity of the material: since there is plenty of oxygen in the surrounding air, the oxidizing gas molecules react by trapping electrons on its surface. semiconductor and thus creates a positively charged space that acts as a barrier to electrical conduction. The advantages of these sensors are [19]:

- the long service life they show (> 10 years)
- the low production cost and the high sensitivity they show since they can detect concentrations that are much less than 1ppm.

Some disadvantages are, however, that:

- they need to operate at temperatures from 200 ° C to 800 ° C in order to function properly
- the reactions that take place in the semiconductor are relatively short, with the result that the energy consumption is significant.
- low selectivity they show due to the plethora of oxidizing gases with which the oxide used reacts and as a result the measurements are incorrect due to "interference"
- they do not show much stability in their operation.

2.3.3 Photoelectric Detection Sensors

From the presentations of [18] and [19] it is understood that the main parts of this sensor are a source of infrared radiation, the measuring chamber where the air sample enters and exits, the optical filter and the infrared detector. During sensor operation, the

source generates light in the infrared range and directs it to the detector through the measuring chamber into which the ambient air sample flows. The beam of infrared light causes oscillations in the air molecules, and this results in the absorption of some wavelengths of light.

These wavelengths will be absorbed but depends each time on the gaseous pollutants flowing into the measuring chamber. The absorption range of wavelengths is a characteristic of any gas that is unique and separates it from other gases. The absorption caused by the respective gas can be described by Beer-Lambert's law. By placing the appropriate optical filter in front of the infrared detector, all the infrared light inside the chamber is blocked, with the exception of the wavelengths that have the ability to absorb the gaseous pollutant to be measured. The concentration of gaseous pollutants is determined by measuring the attenuation of light waves incident on the detector. [19]

Advantages of these sensors:

- They are characterized by high stability in their operation over time as well as high accuracy in their measurements
- The good selectivity they can achieve in detecting specific gaseous pollutants. [18]

Disadvantages:

- these sensors need correction in terms of the effects of temperature, humidity and atmospheric pressure. [18]

2.3.4 Pellistor

The name Pellistor comes from the words Pellet and Resistor. Sensors of this type consist of small pellets of a catalyst located on a ceramic surface which encloses a heat source. The heat source is also used as a heat sensor. These sensors are used to detect gases that are either flammable or have significantly higher thermal conductivity than air. The basic principle of their operation according to the presentation of [18] is that as the flammable gas to be detected comes in contact with the surface of the catalyst, a chemical reaction takes place which results in an increase in the temperature of the sensor. The operating temperature of the catalyst ranges from 200 ° C to 600 ° C.

The signal used for measurements is either the change in the temperature of the catalyst having the energy given by the source constant in the form of heat or the change in the energy given by the source so that the temperature remains constant. The second method performs faster measurements and at high concentrations of flammable gas does not damage the sensor. These sensors are characterized by [13]:

- a short lifespan of not more than 6 months
- limited gas selectivity and low sensitivity with a small detection limit.

2.3.5 Field effect sensors

In this type of sensor according to the presentation of [18] the property of controlling the electric current from an electric field is used, with the help of a suitable electronic field device (Field Effect Transistor) . In the case of a gas sensor, the electrical resistance of the electronic device is determined by dipoles created due to the absorption of the gas whose concentration is to be estimated. Thus, for example, according to the work of [20] for the construction of a sensor in order to determine the concentration of hydrogen, a palladium (Pd) electrode gate is used on the surface of which the hydrogen decomposes. This means that on the surface of the electrode are the positively charged protons resulting in the rest of the electrode being negatively charged and thus creating the dipole layer.

These sensors are only sensitive to the hydrogen and gases they contain, and this makes it a serious limitation to their wider use. To address this limitation various types of gates are used such as perforated or porous electrode gates or gates made of of silicon carbide-based materials. Also, with the use of s a substrate material it is possible to operate the sensors at quite high temperatures and adverse environmental conditions. The advantages of these sensors are as follows [19]:

- For operation at room temperature their energy consumption is low
- By separating the chemical recognition of gases (using different materials for the gates) and the signal output used for the measurements, offer a simultaneous optimization as well as the creation of a wider range of different gas sensitive mattresses utilizing a simple technology

- They are suitable for their direct integration into microelectronic systems offering, using amplifiers, high sensitivity and thus lay the foundations for the production of systems low cost.

In addition to these advantages there are some disadvantages which are [21]:

- Due to their irreversible absorbency their stability in the long run is not good (requires periodic cleaning of their surface by heating it to high temperatures)
- Ambient temperature has an effect on their operation, for example at high temperatures there is an increase in their sensitivity, stability and consumption.

2.3.6 Piezoelectric sensors

From the presentation of [18] we deduce this type of sensor consists of a piezoelectric mechanical resonator that in most cases is a quartz. They also include a coating of material that is sensitive to the gas to be detected and on which the gas is absorbed. This absorption results in a change in the mass charge on the coating. This change also determines the resonant frequency of the piezoelectric resonator whose behavior is described by the Sauerbrey3 equation. The principles on which all the different constructions of these sensors are based are the principle of the "Surface Acoustic Wave" and the principle of the "Acoustic Wave" (Bulk Acoustic Wave).

2.3.7 Particle matter sensors

Optical particle counters are most used to calculate the concentrations of suspended particles in the atmosphere, according to the book by [22]. From the work of [19] the main parts of the sensors are the particle detector, the light source (LED or laser), the light trap and the mirror. During the operation of the sensor, the source sends a continuous laser beam to the light trap in front of it and the flow of atmospheric air passes perpendicular to this beam.

Each time a ray strikes a particle it is reflected and redirected to the mirror which in turn redirects it to the detector where it calculates the number of suspended particles based on the intensity of the scattered light and then converts that number to mass

concentration per cubic meter, based on an assumption about density. Optical particle detectors respond very quickly to changes in concentration [13].

Also, thanks to the principle on which their operation is based, they have the ability to detect and count particles in concentrations of at least 1, provided that they can perceive their diameter. Finally, the cost of these sensors starts from 30 €, which makes them a good choice of sensors for the construction of a low-cost PM measuring device [21].

2.4 Machine Learning

Machine learning is a subset of artificial intelligence (AI) that can solve tasks that are impractical or very difficult to deal with with more traditional programming languages. In 1959, Arthur Samuel defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" [23]. Despite the variety and effectiveness of machine learning algorithms, it is becoming clear that data is arguably more important than any chosen algorithm.

2.4.1 Machine learning categories

Machine learning, based on the feedback of the system under development, is divided into supervised, unsupervised and reinforcement learning.

2.4.1.1 Supervised learning

Supervised learning is the process of training an algorithm that learns from our dataset that could be abstractly considered as a teacher overseeing the learning process. The correct answers are known, the algorithm makes repeated predictions about the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. In practice we see that the majority of machine learning practices use supervised learning

In a mathematical (which is the true nature of machine learning) approach we would say that in supervised learning we have a set of input variables (xdata) and an output variable (Y, the so-called output) and an algorithm is used to calculate the matching

function from input to output Y = f(X). This function is the machine learning model. The goal is for the model to approach the mapping function so well that when new input (x) comes it is possible to predict the output variables (Y) for that data.

2.4.1.2 Unsupervised learning

In unsupervised learning there is only input (X) and there are no corresponding output variables. The goal for unsupervised learning is to model the underlying distribution of the dataset in order to learn more about it. They are called unsupervised learning, because unlike supervised learning above there are no correct answers and there is no teacher.

2.4.1.3 Remedial Learning

Third category is reinforcement learning, which is a type of dynamic programming that trains algorithms using a reward and punishment system. A reinforcement learning algorithm learns by interacting with its environment (e.g. an autonomous driving system). The algorithm is rewarded for executing and predicting correctly, while it is penalized otherwise. The algorithm learns without intervention from a human, maximizing their reward and minimizing their penalty.

2.4.2 Categories of machine learning algorithms

Based on the result extracted by the model, machine learning algorithms are divided into 3 categories: Classification, Regression and Clustering algorithms.

2.4.2.1 Classification

Classification is a supervised learning technique, with the help of which a data that comes to the entrance is classified into one of the categories into which the model has learned to separate. By design, all models are binary classifiers, which means that they distinguish whether an input data belongs (decision^{1}) to a category, or whether it does not belong (decision^{-1}). They were then extended to multiclass classifiers.

2.4.2.2 Regression

Regression is a supervised machine learning technique with the help of which a trained model will be able to predict a specific value for the future based on previous data. An example of regression is predicting the price of a particular stock, not whether it will rise or fall but how much it will be.

2.4.2.3 Clustering

The third category is clustering, which is an unsupervised learning technique for grouping similar data together. The clustering algorithms put the data into different clusters without knowing its nature. After splitting, sorting algorithms are used to separate the input data.

2.4.3 Classification algorithms

2.4.3.1 K-NN algorithm

The K-nn method is a nonparametric method of classification and regression. It belongs to the category of "Instance Based Learning" and to the more general family of "Case Based Reasoning". It is also referred to as the "Lazy classifier" because it does not require learning. It is one of the simplest machine learning models. The algorithm uses the dataset dataset, also called a code book.

Its basic parameters are the K which indicates the number of neighbors it is going to consider.

Another parameter is p which indicates the kind of norm to be used to calculate distance. More precisely, the calculation of the distance between two vectors is given by equation below. Calculation of distance between two vectors X and Y. The values of p also determine the type of distance, as shown below:

P=1, Manhatan Norm
P=2, Eucledian Norm
P=3, Chebyshev Norm

$$d_p(X,Y) = \| X - Y \|_p = \sqrt[p]{(\sum_{i=1}^{m} | (x_i - y_i |)^p}, p \geq 1$$

Calculation of distance between two vectors X and Y

The K-NN algorithm is as follows

1. Input data shall be secured, which are:
   code book
   metric distance (p)
   number of neighbors K (caution avoid even number to avoid ties).
   input data to be categorised
2. Calculate the distances of each point in the code book with the input data
3. Order of distances in ascending order
4. The first K data decides and assigns the input data to any of the existing classes.

For a better understanding we can take a look at Figure 2-3.

The purpose is to categorize the given input of the green dot into any of the existing classes, red triangle or blue square. If 3 neighbors are selected, the shape will be categorized as a red triangle. If 5 neighbors are selected, it changes and becomes a blue square. The choice of the parameter k is therefore crucial. [24]

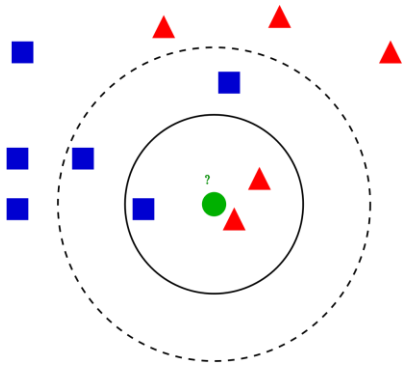


Figure 2-3 Example of categorization with the K-NN algorithm (Wikipedia, n.d.)

2.4.3.2 SVM algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for either classification or regression purposes. The basic idea of SVM is to find a hyperplane that better divides a dataset into two categories (binary classification) [25].

As we can see in Figure 2-4, and more specifically in its left part, a simple sorting problem appears with three different possible separation lines. All three of the lines drawn separate the two categories, with the feeling that all 3 are "correct". However, if we had to choose one of the rows as a classifier for a test dataset, we would have to ask ourselves by what criteria to select it and also which of the rows is better than the others.

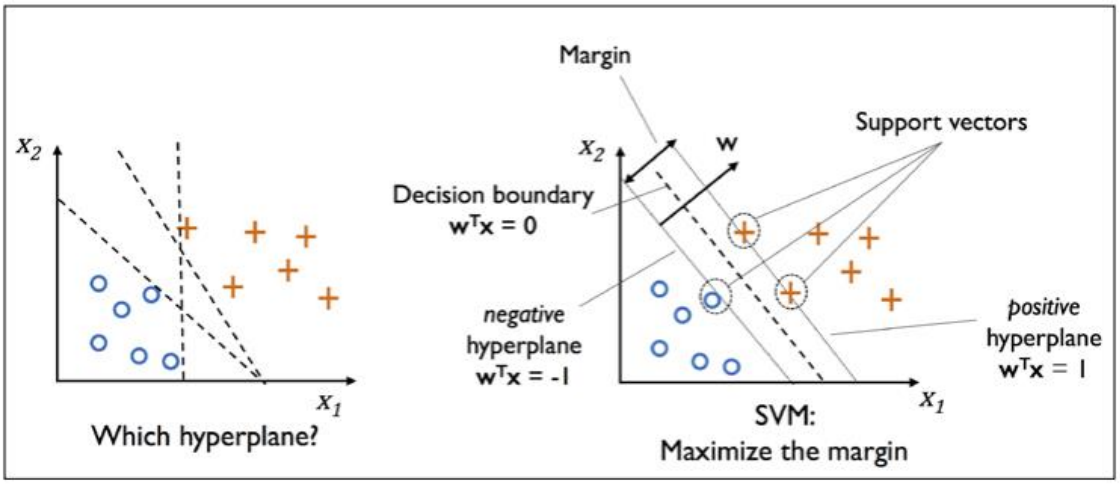


Figure 2-4 SVM Concept (S. Raschka, 2017)

If we select the rows that are glued to either the left or right subset of the data then there is a chance that an entry point will be categorized in the wrong category. That's because we've put the line tight on some of the data points we have on the training set. Thus, a point that will enter will be very close, e.g. to the left category, but the line places it to its right, therefore, it belongs to the right category.

The line in the middle does not have this problem. So we have to choose the line in the middle as the separator. This is the basic idea in SVM. Here we need to define some basic concepts.

- The line in the middle of the separation is called the decision bound
- The area defined by the decision line, if we travel left and right of it until we encounter data, is called margin. The margin is symmetrical around the decision line.
- The data closest to and essentially defining margin separation lines is called support vectors.

One of the requirements is to have the maximum possible margin, so that our algorithm generalizes better to the data that is to come, resulting in them being classified correctly.

An important element of SVM is that support vectors are the most useful data points. After the training we can throw away all the data except the support vectors, which will be used to create the model that will carry out the categorization.

Chapter 3. Designing the system

Existing air quality control regulations mainly apply to cities and open areas, but no system has been established for the interior of buildings where people are crowded and lived. The aim of this work, as mentioned, is to design and implement a low-cost system that can properly measure, record and inform about indoor air quality.

3.1 General diagram of operation

The general plan of the system with the combination of different technologies for receiving, storing, processing and displaying information is presented in Figure 3-1**.**

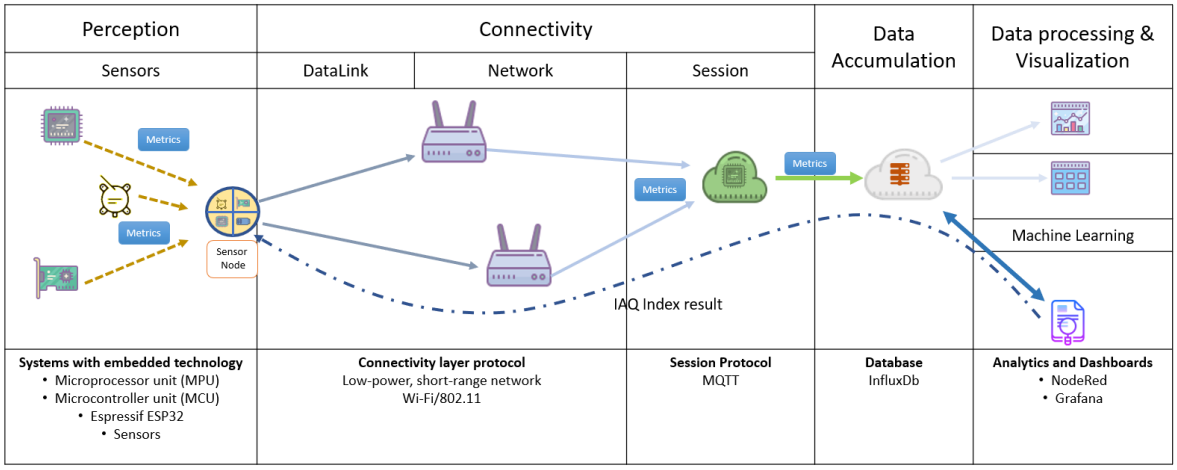


Figure 3-1 Generic system representation

3.1.1 Perception

The combination of sensors, actuators, hardware, and software designed to sense the environment where it exists, using minimum power consumption, creates a measuring station (sensor node) that uses some of the next device types.

Microcontroller unit (MCU) are small computers are embedded on microchips and contain CPUs, RAM, and ROM like ESP32. Although they contain the elements needed to execute simple tasks, microcontrollers are more limited in power than microprocessors.

Microprocessor unit (MPU) houses the functions of CPUs on single or multiple integrated circuits. Although microprocessors require peripherals to complete tasks, they greatly reduce processing costs because they only contain a CPU.

Sensors setect changes in their environment and create electrical impulses to communicate. Sensors commonly detect environmental shifts like changes in temperature, chemicals, and physical position.

3.1.2 Connectivity

In the IoT technology stack, devices connect either through gateways or built-in functionality, usually in short distance nodes and gateways connect a group of sensor nodes to the internet. Sensors' data collected from IoT devices moves through a gateway, gets preprocessed at the edge, and then gets sent to the cloud. In this project we make use of Wi-Fi/802.11 which has low operation cost and is a standard across homes and offices and other small environments. MQTT is also used as a session messaging protocol designed for lightweight machine-to-machine communication and primarily used for low-bandwidth connections to remote locations.

3.1.2.1 MQTT

MQTT [26] is a Client Server publish/subscribe messaging transport protocol. It is light weight, open, simple, and designed to be easy to implement. These characteristics make it ideal for use in many situations, including constrained environments such as for communication in Machine to Machine and Internet of Things (IoT) contexts where a small code footprint is required and/or network bandwidth is at a premium.

The protocol runs over TCP/IP, or over other network protocols that provide ordered, lossless, bi-directional connections. Its features include:

- Use of the publish/subscribe message pattern which provides one-to-many message distribution and decoupling of applications.
- A messaging transport that is agnostic to the content of the payload.
- Three qualities of service for message delivery:
    - "At most once", where messages are delivered according to the best efforts of the operating environment. Message loss can occur. This level could be used, for example, with ambient sensor data where it does not matter if an individual reading is lost as the next one will be published soon after.
    - "At least once", where messages are assured to arrive but duplicates can occur.

- "Exactly once", where messages are assured to arrive exactly once. This level could be used, for example, with billing systems where duplicate or lost messages could lead to incorrect charges being applied.

- A small transport overhead and protocol exchanges minimized to reduce network traffic.
- A mechanism to notify interested parties when an abnormal disconnection occurs.

3.1.3 Data accumulation

This layer determines the type of a database that should be used, which of sensors' data should be persisted, how the data are appropriately organized for the required storage system, if data should be recombined or recomputed or aggregated with previously stored information, some of which may have come from non-IoT sources. Considering all that, in this project we use InfluxDB, a timeseries database, for data storage.

3.1.3.1 InfluxDB

InfluxDB [27] is a high-performance database written specifically for time series data. It allows high-volume data entry, compression and real-time search. It provides write and search capabilities with a command line interface, built-in HTTP API, a set of client libraries (e.g. Go, Java, JavaScript) and plugins for well-known data management applications such as Telegraf, Graphite, Collectd and OpenTSDB. InfluxDB works with InfluxQL, an SQL-like search language for interacting with data. It has been created to be easy to use for those familiar with other SQL environments, while providing features specifically for storing and analyzing long time series data. InfluxDB can handle millions of data points per second.

3.1.4 Visualization and data processing

The last stage is where information interpretation occurs. Software at this level interacts with new and at rest data. Applications vary based on the nature of device data and user needs. In this application will focus on monitoring device data using Grafana and

combining them using NodeRed and Machine Learning to calculate the IAQ index for every measuring device (sensor node). The IAQ index will then follow the opposite direction of the sensors' data packets, to the InfluxDB and finally the sensor node.

3.1.4.1 Grafana

Grafana [28] is an open source interactive data-visualization platform, developed by Grafana Labs, which allows users to see their data via charts and graphs that are unified into dashboard for easier interpretation and understanding. Users can also query and set alerts on their information and metrics from wherever that information is stored, whether that's traditional server environments, Kubernetes clusters, or various cloud services, etc. Additionally users are able to analyze the data, identify trends and inconsistencies, and ultimately make processes more efficient. Grafana was built on open principles and the belief that data should be accessible throughout an organization, not just to a small handful of people. This fosters a culture where data can be easily found and used by anyone who needs it, empowering teams to be more open, innovative, and collaborative.

3.1.4.2 NodeRed

Node-RED [29] is a flow-based programming tool part of the OpenJS Foundation. Flow-based programming is a way of describing an application's behavior as a network of black-boxes, or "nodes" as they are called in Node-RED. Each node has a well-defined purpose; it is given some data, it does something with that data and then it passes that data on. The network is responsible for the flow of data between the nodes.

It is a model that lends itself very well to a visual representation and makes it more accessible to a wider range of users. If someone can break down a problem into discrete steps they can look at a flow and get a sense of what it is doing; without having to understand the individual lines of code within each node.

Node-RED consists of a Node.js based runtime within the browser where you create applications by dragging nodes from your palette into a workspace and start to wire them together. The flows you create can be easily shared as JSON files.

3.2 IAQ calculation technic

The calculation of the IAQ index is based on a mathematical model of calculating the individual indicators for each gas through a function and then selecting the largest value, i.e. the worst characterization . With this paper we propose a reformed mathematical function, based on existing formulas as described in paragraph *Air Quality Indices*, aiming to simplify the calculations.

The measurements stored in the InfluxDB are used to derive the characterization of air quality with the corresponding color category and index value, proportional to the color code for outdoor air, as already described in the theoretical part of the work.

However, these measurements are averages of 24 or 8 hours, depending on which pollutant we are measuring or reporting. In indoor spaces of buildings this does not make much sense due to limited space, smaller air volume that can collect more pollutants, degrading quality faster. Therefore, we propose that the average air pollutants be calculated in rolling periods of ten minutes in order to more directly identify changes in the index [30].

We must also take into account that the presence of people indoors affects quality with the increase of volatile organic compounds (VOC) and carbon dioxide (CO_2). Therefore, their concentration limits and the characterisation of air quality for each category corresponding to the other pollutants should be established. Building on the proposals presented by [30], the following categories and limits of pollutants as presented in Table 3-1, were created and used. The cases where the upper value limit is not specified mean that all the values greater than lower limit are accepted.

INDEX DESCRIPTOR	INDEX VALUE	INDEX COLOR	PM2.5 (µg/m3)	PM10 (µg/m3)	O3 (ppb)	CO (ppm)	NO2 (ppb)	CO2 (ppm)	VOC (ppb)	
GOOD	1	GREEN	0	0	0	0	0	400	0	Lower value
			12	54	54	4,4	53	500	87	Upper Value
MODERATE	2	YELLOW	12,1	55	55	4,5	54	501	88	Lower value
			35,4	154	70	9,4	100	1000	160	Upper Value
UNHEALTHY FOR SENSITIVE GROUPS	3	ORANGE	35,5	155	71	9,5	101	1001	161	Lower value
			55,4	254	85	12,4	360	1500	261	Upper Value
UNHEALTHY	4	RED	55,5	255	86	12,5	361	1501	262	Lower value
			150,4	354	105	15,4	649	2000	350	Upper Value
VERY UNHEALTHY	5	PURPLE	150,5	355	106	15,5	650	2001	351	Lower value
			250,4	424	200	30,4	1249	3000	430	Upper Value
HAZARDOUS	6	MAROON	250,5	425	201	30,5	1250	3001	431	Lower value
			500,4	604	-	50,4	2049	-	-	Upper Value

Table 3-1 Proposed limits of each IAQ category

As described in the Air Quality Indices section in combination to the Table 3-1, for each gaseous pollutant, a scale has been created that divides the possible values that its concentrations may have, into six groups. The groups correspond to the value of the air quality index, with distinct values from 1 to 6, the corresponding rating (Good, Moderate, Unhealthy for sensitive groups, Unhealthy, Very unheathy, Hazardous) and the corresponding color. Each group has lower and upper concentration limits so that the average of the measurements over the 10-minute period is placed in the correct group and the pollutant index I_x is obtained. Therefore, the following mathematical function is obtained that gives the value of the air quality index I_x for the gaseous pollutant X.

$$I_x = \begin{cases} 1, & X_{l_v_1} \leq \text{Conc}_x \leq X_{u_v_1} \\ 2, & X_{l_v_2} \leq \text{Conc}_x \leq X_{u_v_2} \\ 3, & X_{l_v_3} \leq \text{Conc}_x \leq X_{u_v_3} \\ 4, & X_{l_v_4} \leq \text{Conc}_x \leq X_{u_v_4} \\ 5, & X_{l_v_5} \leq \text{Conc}_x \leq X_{u_v_5} \\ 6, & X_{l_v_6} \leq \text{Conc}_x \leq X_{u_v_6} \end{cases}$$

Equation 1 Pollutant index equation

- X : Type of pollutant
- $X_{l_v_i}$: Lower value of the corresponding group i of pollutant X, $i \in [1,2,3,4,5,6]$
- $X_{u_v_i}$: Upper value of the corresponding group i of pollutant X, $i \in [1,2,3,4,5,6]$
- Conc_x : Concentration of pollutant X

After repeating the calculation of I_X for every pollutant, we take the maximum value of the individual indicators to obtain the general index and the characterization of air quality thus the function is shaped as

$$IAQ = \max(I_{PM2.5}, I_{PM10}, I_{O3}, I_{CO}, I_{NO2}, I_{CO2}, I_{VOC})$$

Equation 2 Function of IAQ index value

For example, the measurements of Table 3-2 are obtained during sensors' testing period, every row represents a set of metrics that recorded by the same station at the same time. Our goal is to calculate the IAQ of each row.

ROW	PM2.5	PM10	O3	CO	NO2	CO2	VOC
1	8.2	9.4	15	1.7	14	495	86
2	13.1	10.9	16	1.7	14	493	136
3	52	0.2	55	1.5	12	540	79
4	97	265	16	1.6	17	405	72
5	21	22	107	1.5	17	431	320
6	228	602	16	1.6	20	510	2896

Table 3-2 Example average metrics

Row 1 - Step 1

Firstly, for every pollutant in row 1, we apply the Equation 1 in conjuction with the Table 3-1 and thous we get:

- $PM2.5_{l_v_1} \leq \text{Conc}_{PM2.5} \leq PM2.5_{u_v_1}$ => $0 \leq 8.2 \leq 12$ => $I_{PM2.5} = 1$
- $PM10_{l_v_1} \leq \text{Conc}_{PM10} \leq PM10_{u_v_1}$ => $0 \leq 9.4 \leq 54$ => $I_{PM10} = 1$
- $O3_{l_v_1} \leq \text{Conc}_{O3} \leq O3_{u_v_1}$ => $0 \leq 15 \leq 54$ => $I_{O3} = 1$
- $CO_{l_v_1} \leq \text{Conc}_{CO} \leq CO_{u_v_1}$ => $0 \leq 1.7 \leq 4.4$ => $I_{CO} = 1$
- $NO2_{l_v_1} \leq \text{Conc}_{NO2} \leq NO2_{u_v_1}$ => $0 \leq 14 \leq 53$ => $I_{NO2} = 1$
- $CO2_{l_v_1} \leq \text{Conc}_{CO2} \leq CO2_{u_v_1}$ => $400 \leq 495 \leq 500$ => $I_{CO2} = 1$
- $VOC_{l_v_1} \leq \text{Conc}_{VOC} \leq VOC_{u_v_1}$ => $0 \leq 14 \leq 53$ => $I_{VOC} = 1$

Row 1 - Step 2

The next and final step is to apply the Equation 2 to get max index as IAQ.

$$IAQ = \max(I_{PM2.5}, I_{PM10}, I_{O3}, I_{CO}, I_{NO2}, I_{CO2}, I_{VOC}) = \max(1,1,1,1,1,1,1) = 1$$

Row 2 - Step 1

Repeating the procedure for data in row 2 we get:

- $PM2.5_{l_v_2} \leq \text{Conc}_{PM2.5} \leq PM2.5_{u_v_2}$ => $12.1 \leq 13.1 \leq 35.4$ => $I_{PM2.5} = 2$
- $PM10_{l_v_1} \leq \text{Conc}_{PM10} \leq PM10_{u_v_1}$ => $0 \leq 10.9 \leq 54$ => $I_{PM10} = 1$
- $O3_{l_v_1} \leq \text{Conc}_{O3} \leq O3_{u_v_1}$ => $0 \leq 16 \leq 54$ => $I_{O3} = 1$
- $CO_{l_v_1} \leq \text{Conc}_{CO} \leq CO_{u_v_1}$ => $0 \leq 1.7 \leq 4.4$ => $I_{CO} = 1$
- $NO2_{l_v_1} \leq \text{Conc}_{NO2} \leq NO2_{u_v_1}$ => $0 \leq 14 \leq 53$ => $I_{NO2} = 1$
- $CO2_{l_v_1} \leq \text{Conc}_{CO2} \leq CO2_{u_v_1}$ => $400 \leq 493 \leq 500$ => $I_{CO2} = 1$
- $VOC_{l_v_2} \leq \text{Conc}_{VOC} \leq VOC_{u_v_2}$ => $88 \leq 136 \leq 160$ => $I_{VOC} = 2$

Row 2 - Step 2

The next and final step is to apply the Equation 2 to get max index as IAQ.

$$IAQ = \max(I_{PM2.5}, I_{PM10}, I_{O3}, I_{CO}, I_{NO2}, I_{CO2}, I_{VOC}) = \max(2,1,1,1,1,1,2) = 2$$

The above two-step process is repeated for every set of metrics available in general and for each row of Table 3-2 which, in short, creates the Table 3-3 of pollutant indeces and the resulting IAQ value in the last column.

ROW	PM2.5	PM10	O3	CO	NO2	CO2	VOC	IAQ
1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	2	2
3	3	1	2	1	1	2	1	3
4	3	4	1	1	1	1	1	4
5	2	1	5	1	1	1	4	5
6	5	6	1	1	1	2	6	6

Table 3-3 Seperate indeces of pollutants and resulting IAQ

In addition to the above proposed index and calculation technic, it is proposed to study whether is efficient to train an appropriate machine learning model that will categorize each set of measurements, recorded by a measuring station, into one of the six categories-characterizations of air quality (IAQ), as shown in Table 3-1**.** This procedure is described and implemented in paragraph Calculating the IAQ Index using ML.

3.3 Operational states

The measuring station (sensor node) performs a set of checks for the existence and operation of necessary systems and the appropriate choice of its mode of operation. Below are defined in detailed the stages and in certain order of execution:

1. Power On
2. Check for known Wifi connection
3. If available, set ESP32 in station mode and connect to Wifi network
   3.1. Begin sensors' initialization
   3.2. Repeat every hour

3.2.1. Update date-time using NTP

3.3. Repeat every minute

3.3.1. Measure Temperature and Humidity

3.3.2. Measure PM2.5 , PM10

3.3.3. Measure VOC, CO_2, O_3, CO, NO_2

3.3.4. Show new data on sensor's screen

3.3.5. Send data using MQTT to InfluxDB

3.3.6. NodeRed calculates and sends back new IAQ index

3.3.7. Get the new IAQ and change the LED color accordingly

4. Else set ESP32 as Access Point mode and create wifi ESP32XXXX network

4.1. Start local webserver in ESP32 to show and store new attributes (for wifi connection, MQTT, InfluxDB) using a webpage

4.2. A user can connect to ESP32's wifi network and access webpage on a specified ip address to configure it's properties

4.3. After saving new properties, restart device

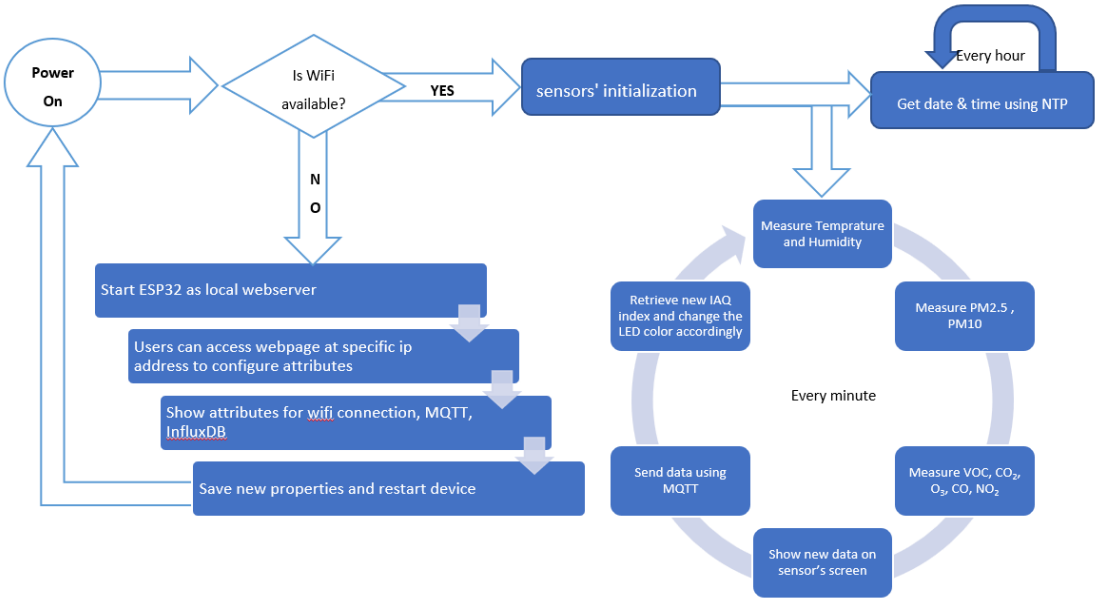


Figure 3-2 Stages of operation and procedures of the measuring station

3.4 Flow chart of data exchange

The recording of sensor data (raw data) alone is not enough get a conclusion about the air quality in the space where they are located. Equally important is the development of systems for their storage, processing and display. Every minute, each measurement is sent to a server where a time series database operates.

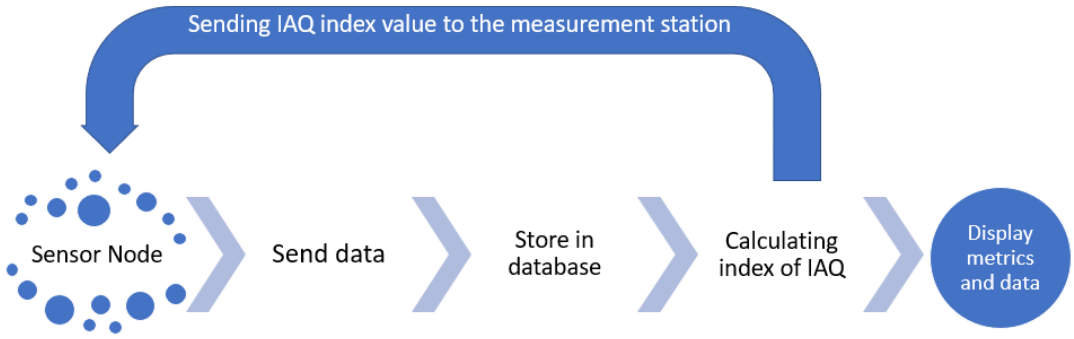


Figure 3-3 Flow chart of data routes

This is the point where two different flows start as shown in Figure 3-3. In the first, data are extracted, in groups of the last ten minutes, from which, using a machine learning system, the air quality index is sent to the measurement system from which the measurements originated. In the second, the data is retrieved and displayed as graphs through a website, depending on the settings and filters applied by the visitor and it is possible to view graphs per sensor or overall. In addition, a map allows us to geographically determine the location of each measurement system and see the current air quality index.

Chapter 4. System implementation

4.1 Selecting sensor components

The sensors cover a wide range of aerosol concentration measurements and provide instant information. Their prices show great variations due to the different quality characteristics of their construction and operation.

One of the goals of our study is to keep the cost of the measuring station low. For this reason, at least two low-cost sensors per aerosol category were selected for comparison and depending on market availability.

The aerosols detected by the station and the corresponding sensors and their types are listed in the Table 4-1

Pollution factor	Sensor		Other detectable gas
	Manufacturer	Model	
Particulate Matter (PM)	Sensirion	SPS30	
	Plantower	PMS7003	
	Nova	SDS011	
	Honeywell	HPMA115S0	
TVOC	Sensirion	SGP30	
	SGX Sensortech	CJMCU 8128	
CO_2	Sensirion	SGP30	
	SGX Sensortech	CJMCU 8128	
CO	SGX Sensortech	CJMCU 6814	NH_3, H_2S, CH_4, C_2H_5OH, C_3H_8, C_4H_{10}
	SGX Sensortech	CJMCU 4541	NH_3, H_2S, CH_4, C_2H_5OH
NO2	SGX Sensortech	CJMCU 4541	NO, H_2
	SGX Sensortech	CJMCU 6814	NO, H_2
O_3	Winsen	ZE25	

Sensor	Type
SPS30	Particle matter
PMS7003	Particle matter
SDS011	Particle matter
HPMA115S0	Particle matter
SGP30	Metal oxide
CJMCU 4541	Metal oxide
CJMCU 6814	Metal oxide
CJMCU 8128	Metal oxide
ZE25	Electrochemical

Table 4-1 Sensors, their type and the pollutants they detect

4.1.1 Nova SDS011

Using the laser scattering principle, the SDS011 can detect particle concentrations in the air between 0.3 and 10 micrometers. It is dependable and stable due to its internal fan and digital output. Its characteristics are:

1. Accurate and Reliable: steady, good consistency laser detection.
2. Prompt reaction: when the situation changes, the reaction time is under 10 seconds.
3. Simple integration: fan built-in, UART output.
4. High resolution: 0.3 μg/m3 resolution.

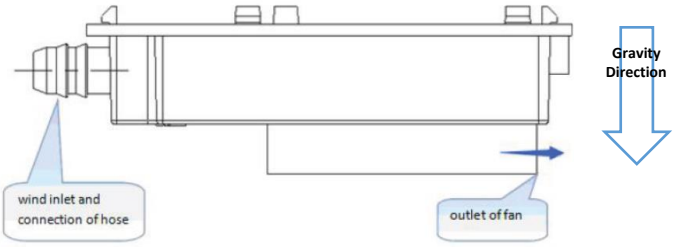


Figure 4-1 Sensor's proper placement



Figure 4-2 Nova SDS011

This sensor's high-quality laser diode has a service life of up to 8000 hours. The default setup, which measures at the frequency of once per second, might be used if real-time data is required. When the real-time demand is low, it can be switched to the discontinuous measurement mode to extend the service life [31].

4.1.2 Plantower PMS7003

A general-purpose, digital particle concentration sensor, which in order to give accurate concentration data over time, this sensor may be included into a variety of tools that gauge the amount of airborne suspended particles or other environmental improvement tools [32].

The laser scattering concept is also used in the sensor, and it can calculate the equivalent particle diameter and the quantity of different-sized particles present in a given volume using a microprocessor.

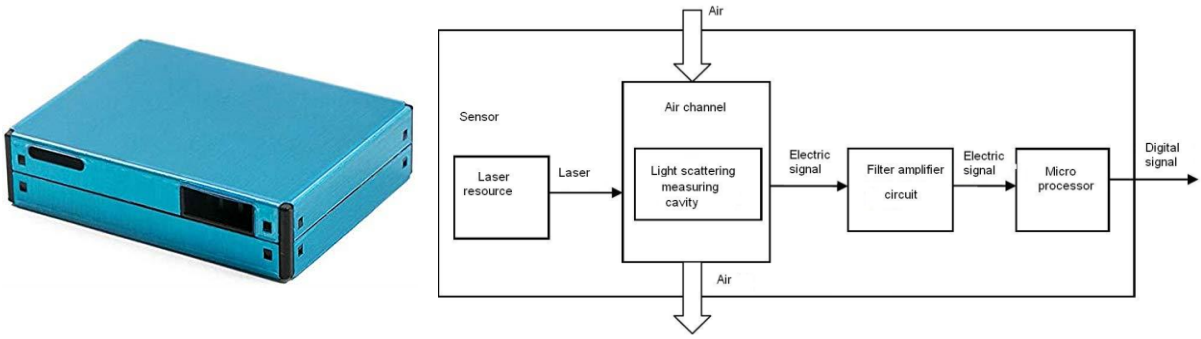


Figure 4-3 Plantower PMS7003 image and flow chart

Particle number is measured in volume units of 0.1L, whereas mass concentration is measured in units of $\mu g/m^3$. The quality and number of particles of various sizes per unit volume are the main metrics used to assess the output.

The output of a digital device might be passive or active. Upon startup, the default setting is activated. The host would automatically receive serial data from the sensor in this mode. Stable mode and fast mode are the two sub-modes of the active mode. The sensor would operate in steady mode with a true interval of 2.3 seconds if the concentration change was slight. Additionally, the sensor will switch to fast mode instantly in the event of a significant change, with an interval of 200–800 ms; the higher the concentration, the shorter the delay.

4.1.3 Sensirion SPS30

The SPS30 [33] particulate matter uses laser scattering and both UART and I^2C communication protocols. Supports Sensirion's contamination-resistance technology, that permits exact measurements from the device's initial use and for the duration of its more than ten-year lifetime. This is achieved with an automatic fan-cleaning procedure which is triggered periodically following a defined cleaning interval. This will accelerate the fan to maximum speed for 10 seconds to blow out the dust accumulated inside the fan. The default cleaning interval is set to one week, however if the sensor is often triggered or deactivated, a manual cleaning cycle should be start at least once per week.



Figure 4-4 Sensirion SP30

The major difference compared to the others, is that this sensor can count particles and send results as number of particles in a cm^3 (number concentration $\#/cm^3$). It's algorithm is capable of distinguish particles in categories as PM0.5, PM1.0, PM2.5, PM4.0, PM10 and calculate the typical particle size as an indication on the average particle diameter in the sample aerosol. Such output correlates with the weighted average of the number concentration bins measured with a TSI 3330 optical particle sizer. Consequently, lighter aerosols will have smaller TPS values than heavier aerosols.

All data for PM2.5 particles are calibrated to TSI DustTrak DRX 8533 Ambient Mode for mass concentration and to TSI OPS 3330 for number concentration. Tryner, Mehaffy et al., found that SPS30 sensors were, in fact, more contaminant resistant than PMS5003 or SM-UART-04L sensors, possibly due to the properties listed above [34].

4.1.4 Honeywell HPMA115S0

The HPMA115S0 is a HPM series PM2.5 particle sensor with UART output that uses light scattering method to detect and count particles in concentration range of 0 $\mu g/m^3$ to 1000$\mu g/m^3$ in each environment. A laser light source illuminates' particle as it is pulled through the detection chamber and is recorded on the photo or light detector. The light is then analyzed and converted to an electrical signal providing particulate size and

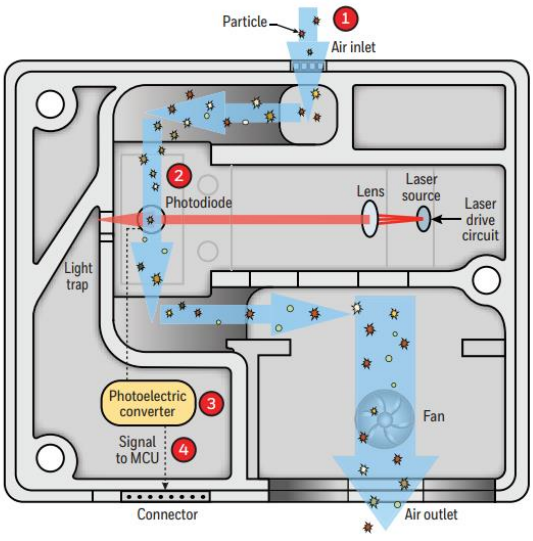


Figure 4-5 Honeywell HPMA115S0 operation description

quantity to calculate concentrations in real time. Signal is transmitted to micro control unit where a proprietary algorithm processes the data and supplies outputs for the density of the particulate ($\mu g/m^3$) [35].

It works very steadily in automatic data transmission mode with data being sent every 1 sec, as it was observed during testing periods. In query mode operation, a command wakes up the sensor and asks for the current measurement. The response was very slow compared to other sensors due to higher warm-up time and many measurements were lost or error messages were sent by the sensor. Also, it shows delays that are not related to exogenous factors (humidity, contamination agents, temperature etc) which were increased as the runtime period was longer, without any obvious reason.

4.1.5 DHT22

The DHT22 is a low-cost digital temperature and humidity sensor with a single wire digital interface. It uses a capacitive humidity sensor and a thermistor to measure the surrounding air and spits out a digital signal on the data pin (no analog input pins needed) [36].

The sensor is calibrated and doesn't require extra components so you can get the right to measuring relative humidity and temperature.

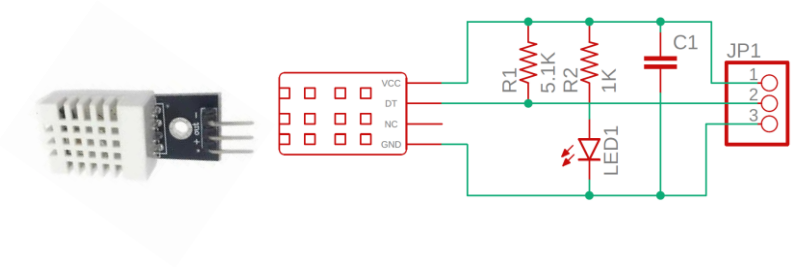


Figure 4-6 DHT22 image and electrical connection diagram

Higher relative humidity decreases the resistance between the electrodes, while lower relative humidity decreases the resistance between the electrodes. The lower relative humidity increases the resistance between the electrodes. thermistor. A thermistor is a thermal resistance - a resistance that changes resistance with temperature. Technically, all resistors are thermistors - their resistance changes slightly with temperature - but the change is usually very small and difficult to measure [37].

4.1.6 Sensirion SGP30

The SGP30 is a digital multi-pixel gas sensor that offers a complete sensor system on a single chip, featuring a digital I^2C interface, a temperature controlled micro hotplate, and two preprocessed indoor air quality signals. It's measurements are based on ethanol and H_2 signals which, with the usage of a baseline compensation algorith, are transformed to a TVOC and a CO_2eq signal [38].

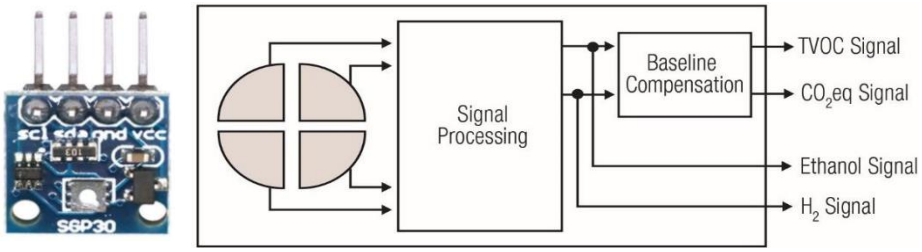


Figure 4-7 Sensirion SGP30 image and operation flow chart

The SGP30 is exceptionally stable and the constant baseline compensation technique preserves the accuracy of the readings over time. By incorporating an external humidity sensor to provide humidity correction to your measurements, you may further refine them. The sensor shows best results when a 1Hz sampling rate is used.

4.1.7 CJMCU8128

This sensor consists of the three different purpose other sensors, having some benefits but some disadvantages too. More specifically, includes an AMS CCS811 sensor for CO_2 and TVOC, a Bosch BME280 for temperature, relative humidity, air pressure and a T.I. HDC1080 for relative humidity and temperature. I^2C communication protocol is used as it is the common between all three sensors.

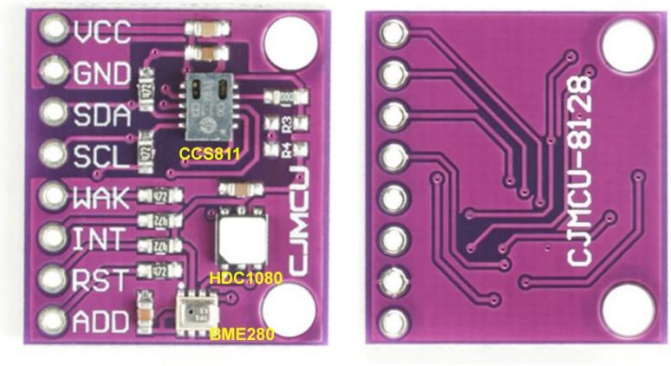


Figure 4-8 CJMCU8128 front and back image

The CCS811 is a metal oxide sensors like the MiCS sensors, and as it is mentioned in the paragraph *MiCS Sensor working principles* in detail, measurements depend on sensor layer's baseline resistance R_0. However, metal oxide sensors do not give absolute readings as the resistance Rs varies from sensor to sensor due to manufacturing variation, from use-case to use-case, and over time. To mitigate this problem, the output of the sensor is normalized: Rs is divided by R_0. The value of R_0 is known as the baseline. R_0 cannot be determined by a one-time calibration and it is maintained on-the-fly in software. This process is known as baseline correction. This sensor provides automatic baseline correction after initial device operation. As the air quality is expected to vary in a typical environment so the minimum time over which a baseline correction is applied is 24 hours [39].

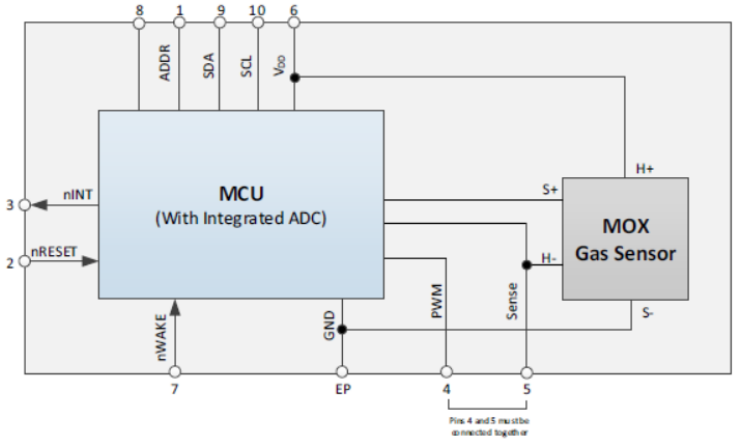


Figure 4-9 CCS811 block diagram

The BME280 achieves high performance, accuracy and response in measurements due to the algorithm of calculations and the embedded raw data filtering capabilities, called IIR which can be configured to different coefficients [40].

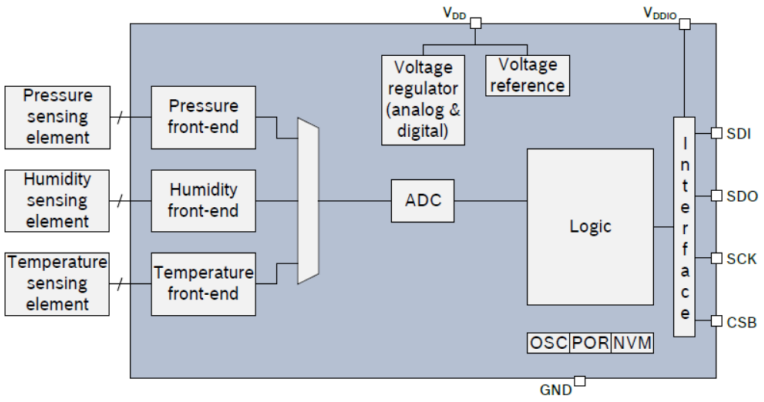


Figure 4-10 BME280 block diagram

The HDC1080 is also a high resolution and accuracy sensor, a bit more than the BME280 and maybe this is the reason why it is part of the CJMCU8128 [41]. More specifically, it exhibits an accuracy of ±0,2 ºC (±0,5 ºC o BME280) at temperatures from 10 ºC to 50ºC and deviation in relative humidity by ±2% (±3% o BME280) in the range from 20% to 60% relative humidity.

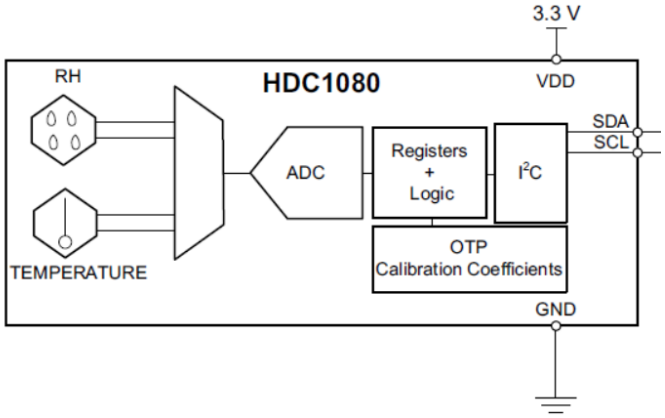


Figure 4-11 HDC1080 block diagram

4.1.8 Winsen ZE25-O3

ZE-O3 is a general-purpose electrochemical formaldehyde detection module for the ozone in open-air or in-door environments with detection rate that varies between 0 – 10000 ppb. It is highly sensitive to ozone concentration changes with resolution of 10 ppb and uses UART communication protocol [42].

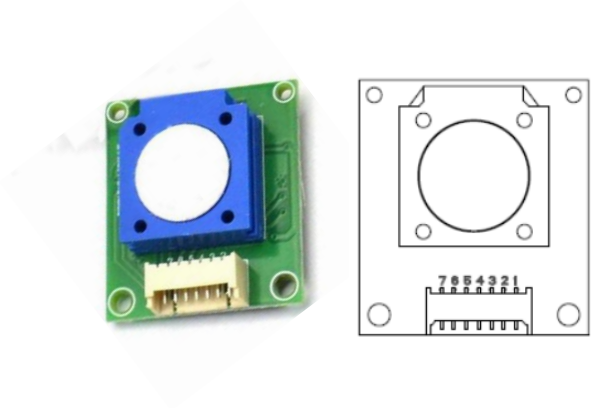


Figure 4-12 ZE25-O3 image and pin numbering

In order to be used, it has to be warm-up for at least 5 minutes in an outdoor, clean environment and at every power on, the sensor requires 90 seconds to send the first measurements. The sensor has two operation modules, active as a default that sends gas concentrations every second and query.

The sensor is valnurable to excessive vibrations and can not be exposed in high concentration organic gas for a long time.

4.1.9 MiCS Sensor working principles

The MiCS sensor is a metal oxide sensor which its operation relies on an output signal (Vo) that is usually read between the sensitive layer resistance R_S and a load resistor. In order to obtain the best readings, some preinstallation work has to be done, beginning by powering the sensor in a clean without pollutants environment for at least 60 minutes in order the readings are stabilized [43].

The resistors are linked in series and powered by 5V (Figure 4-13). The easiest way to track sensor resistance changes, is to add a load resistor to Rsensor and measure the voltage drop in this resistor while the bridge (Rsensor + Rload) is generally operated between 5V and GND. After that the sensor output (Vo) is measured with the aid of a microcontroller, an analog to digital converter or an analog output indication.

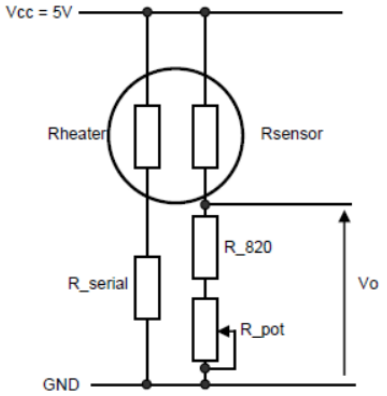


Figure 4-13 MiCS operation cicruit

For measurements to be accurate and consistent over time, the current going through the sensor layer must be restricted. This is the rationale for the use of a two-resistor

load circuit made up of a series connection between a 500kOhm potentiometer (Rpot) and an 820 Ohm initial load resistor (R820). The minimum value needed to restrict sensor current is R820 (Figure 4-13).

After choosing a proper value of Rpot and preheating the sensor, we can observe in Figure 4-14 that stabilized Vo value is around 4.5V in these test conditions. To obtain the best resolution it is advised to work around the mid-operating range (i.e. Vcc/2 = 2.5V). To adjust Vo at 2.5V, load resistance has to be reduced. This is done by decreasing the resistance value of the potentiometer. This action has been done at time t0+100 min and is shown in the Figure 4-14.

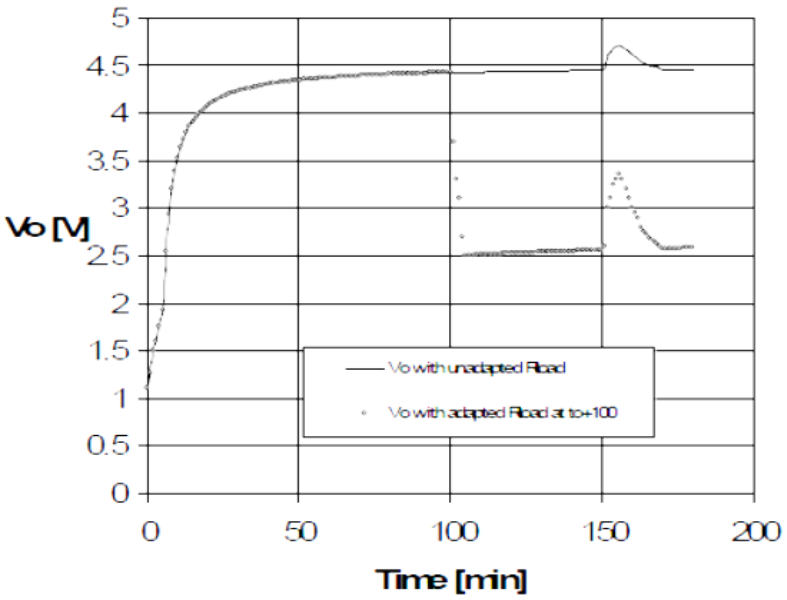


Figure 4-14 MiCS measurement voltage

At +150 min, a pollution event, represented by the opening of a permanent ink pen in the vicinity of the sensor for example, is represented. The increase of Vo is representative of the increase in "gas detected" (drop of sensor resistance value). One can observe that the voltage change is then higher for the signal adjusted at 2.5V baseline than for the one at 4.5V.

	Baseline adjusted around Vcc/2	Baseline at 80%of Vcc	
Vout_0	2.56	4.45	[V]
Vout_event	3.35	4.70	[V]
Delta Vout	0.79	0.25	[V]
Delta Vout/Vout_0	31%	6%	
Rsensor_0	51.6	51.6	[kOhms]
Rsensor_event	26.6	26.6	[kOhms]
Δ Rsensor	-25.0	-25.0	[kOhms]
Δ Rsensor/Rsensor_0	-48%	-48%	

Figure 4-15 Numeric evaluation

As the specifications declare, the absolute sensitivity of 48% is identical for all load resistances, tests should be done with a lot more accuracy if the load resistor is changed to have a baseline measurement range of about Vcc/2.

$R_{sensor} = R_{LOAD} \times (Vcc - Vo) / Vo$

R_{LOAD}: load resistor (Rpot + R820)

Vcc = 5.0 V

Vo = voltage detected on pin SensorOut

Then, sensitivity may be calculated using S = (Rsensor-Rsensor0)/Rsensor0.

Rsensor0: resistance value prior to an event of pollution

Rsensor: resistance when a pollutant is present

With an 10-bit ADC and 5V as the reference voltage, we may use 1024 measurements with a resolution of 5V/1024 = 4,88 mV. When the baseline is changed to 80% of Vcc, there is a measurement discrepancy of 4,70-4,45 = 0,25V. Hence 250mV/4,88mV = 51 ADC steps.

The observed difference at a baseline adjustment of roughly 50% of Vcc is 3,35-2,56 = 0,79V. Thus 790mV/4,88mV = 161 ADC steps.

This increase of ADC steps employed for measurement, reflects the improved precision and detecting capabilities of the measurement circuit. Using an uneven load resistor may lead to erroneous sensor data for pollution detection and inaccurate performance forecasts for the sensor in your application [43].

To adapt the scale to the resistance range, the microcontroller may switch two or three resistors in series. The number of resistors that must be changed depends on the ADC's precision. Two switching resistors may be sufficient with a 10-bit ADC to span a resistance detection range, such as 2kOhms to 1MOhms.

4.1.10 CJMCU4541

This sensor is based on MiCS 4514 dual sensor for hydrocarbons and oxidising gases with independent heaters and sensitive layers, has separate input for on-demand sensor preheat, high sensitivity and wide detection range. It sends a separate signal for each gas group (hydrocarbons RED, oxidising gases NOX) in the form of a power voltage that has to be transformed to a gas concentration value using the data that specifications sheet provides and the instructions as described in the previous section of MiCS Sensors [44].

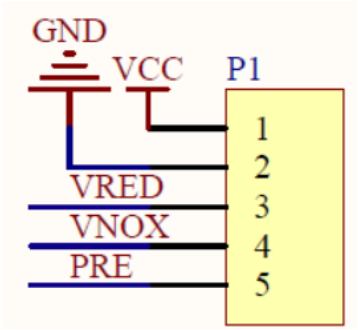


Figure 4-16 CJMCU4541 pinout diagram

The concentration is calculated by following the charts provided by the manufacturer as seen in Figure 4-17 and Figure 4-18.

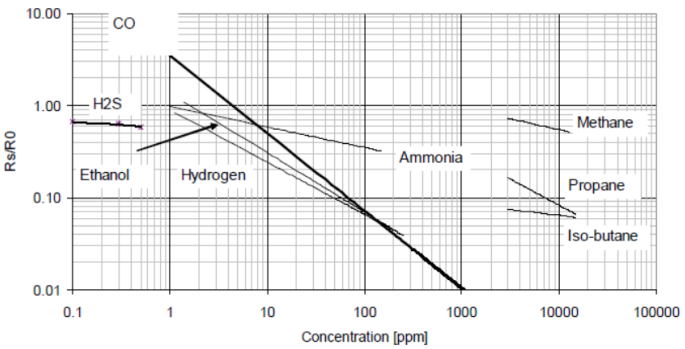


Figure 4-17 RED sensor, continuous power ON, 25°C, 50% RH

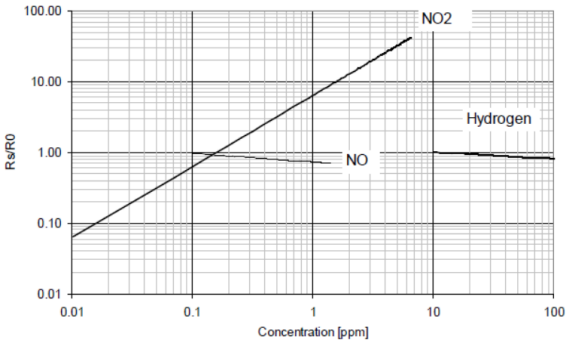


Figure 4-18 OX sensor, continuous power ON, 25°C, 50% RH

4.1.11 CJMCU6814

This sensor is based on MiCS 6814 triple sensor for hydrocarbons, ammonia and other oxidising gases with independent heaters and sensitive layers. It sends a separate signal for each gas group (hydrocarbons RED, oxidising gases NOX, ammonia NH_3) in the form of a power voltage that has to be transformed to a gas concentration value using the data that specifications sheet provides and the instructions as described in the previous section of MiCS Sensors.

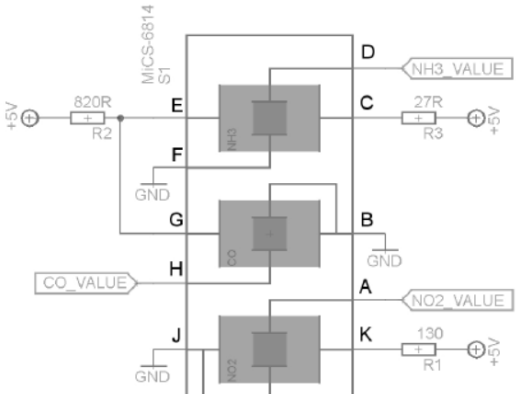


Figure 4-19 CJMCU6814 operation cicruit

The concentration is calculated by following the charts provided by the manufacturer as seen below in Figure 4-20, Figure 4-21 and Figure 4-22. It has to be placed outside for 48 hours for calibration according to its specs [45].

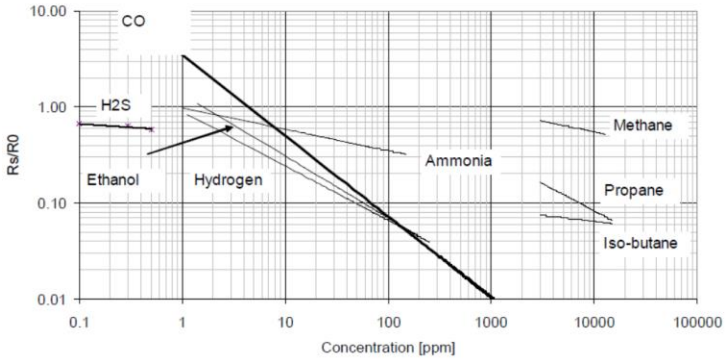


Figure 4-20 RED sensor, continuous power ON, 25°C, 50% RH

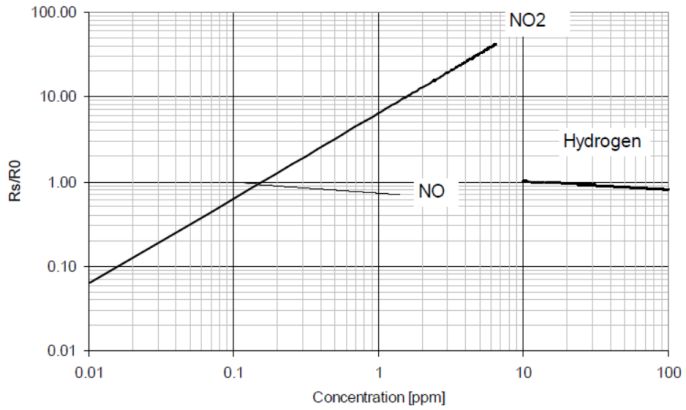


Figure 4-21 OX sensor, continuous power ON, 25°C, 50% RH

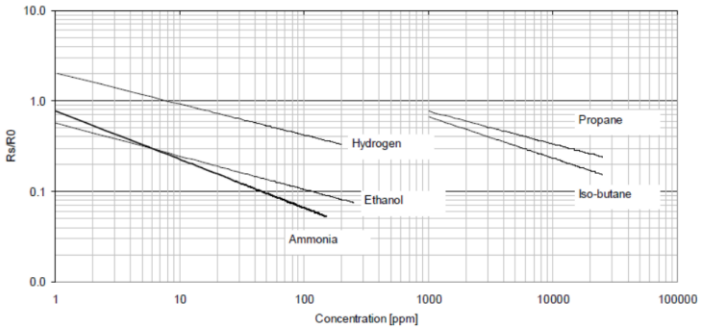


Figure 4-22 NH3 sensor, continuous power ON, 25°C, 50% RH

4.2 Sensors Software Libraries

An important issue for the operation of sensors is the availability and proper operation of software libraries for communication with microcontrollers and in our case with ESP32. Their manufacturers do not have the same policy and support towards end users, so they are divided into two categories. In first group are these who provide a complete guide to the commands supported by the sensor and libraries built into the Arduino IDE including SDS011, SPS30, PM7003, HPMA115S0, Winsen ZE25 and SGP30.

In the second category of manufacturers of those who do not provide libraries but should be developed separately for each one or shared online by other users, belong CJMCU4541, CJMCU8128 and CJMCU6814. In this case, of course, they should not be used immediately without testing the reliability and stability of communication and sensor measurements.

In addition, the results depend on the current voltage and the microcontroller used in the circuit. This, as described in the technical specifications and instructions of the manufacturers, is due to the fact that each sensor needs a different resistance to adjust the voltage V of the signal bearing the measurement, in the middle of the interval. That is, if an ESP32 microcontroller operating at 3.3V is used, the output voltage of the sensor must be set to 1.7V, by the procedure mentioned in *MiCS Sensor working principles*.

For the needs of this work, libraries were developed for the two aforementioned sensors which are presented below.

4.2.1 Βιβλιοθήκη CJMCU4541

The sensor, as depicted in Figure 4-23, is linked to the application through a pull-up device employing two resistors, albeit not explicitly outlined in the manufacturer's specifications. However, in order to obtain correct measurements, the sensor must be initialized in fresh air to obtain the base resistance R0, so that the ratio Rs/R0 gives the concentration of the measured gas, using the pin 5 to preheat it.

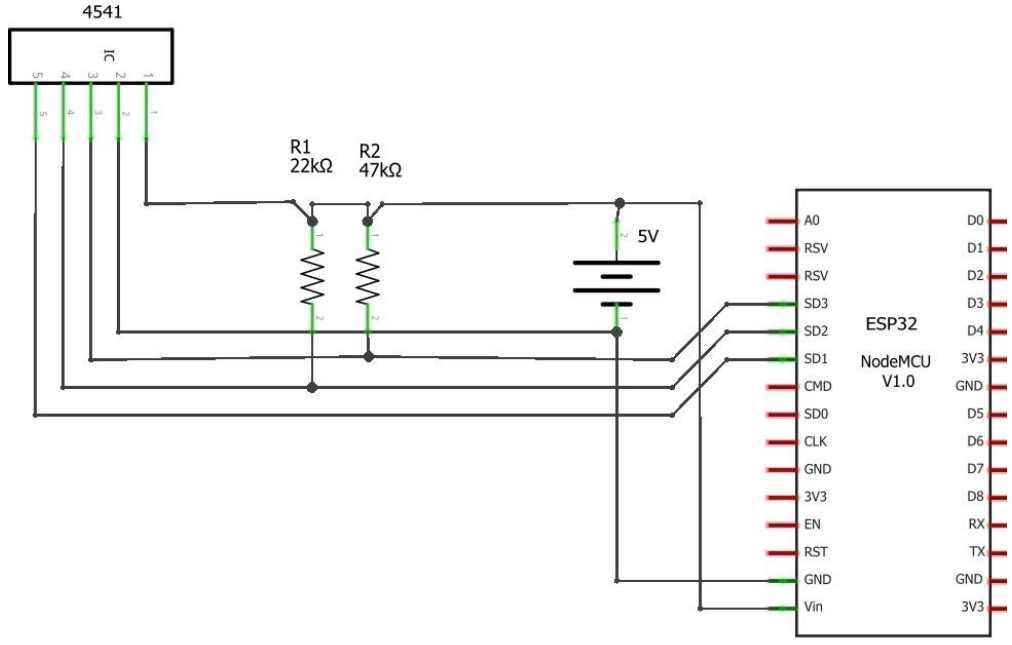


Figure 4-23 CJMCU4541 electrical circuit

The resistor values were selected at 22KΩ and 47KΩ for NO2 and CO respectively because, as explained in *MiCS Sensor working principles*, the output voltage of each sensor must be in the middle of the operating range in order to have greater accuracy.

The functions for the correlation of gas concentration and measured resistance ratio were then determined according to the graphs provided by the manufacturer and referred to in paragraph *CJMCU4541* in Figure 4-17 and Figure 4-18**.**

Based on these, the following were derived for each gas measured:

CO: $ppmco = 4.4574*x^{-1.177}$, $R^2 = 1$
NO2: $ppmNO2 = 0.1532*x^{1.002}$, $R^2 = 0.9999$

which will be used in the sensor library to obtain the concentrations.

The library has two main functions:

- calculating the base resistance R0 for each calibrate() gas, each time the sensor is activated, and
- measuring the concentration of a specific gas at the request of the user, measure(gas_t gas) that returns the ppm of the gas gas [CO, NO2] given as a parameter.

4.2.2 Βιβλιοθήκη CJMCU6814

The sensor, as shown in Figure 4-24, is connected to three resistors in a pull-up way, although this is not clearly stated in the manufacturer's specifications. The correct values of resistors and the circuit connections were concluded after several tests, according to the manufacturer's instructions, during which measurements that were not reasonable, were constantly produced, such as, for example, a reduction in gas concentration values when opening containers with ammonia (NH3) near the sensor.

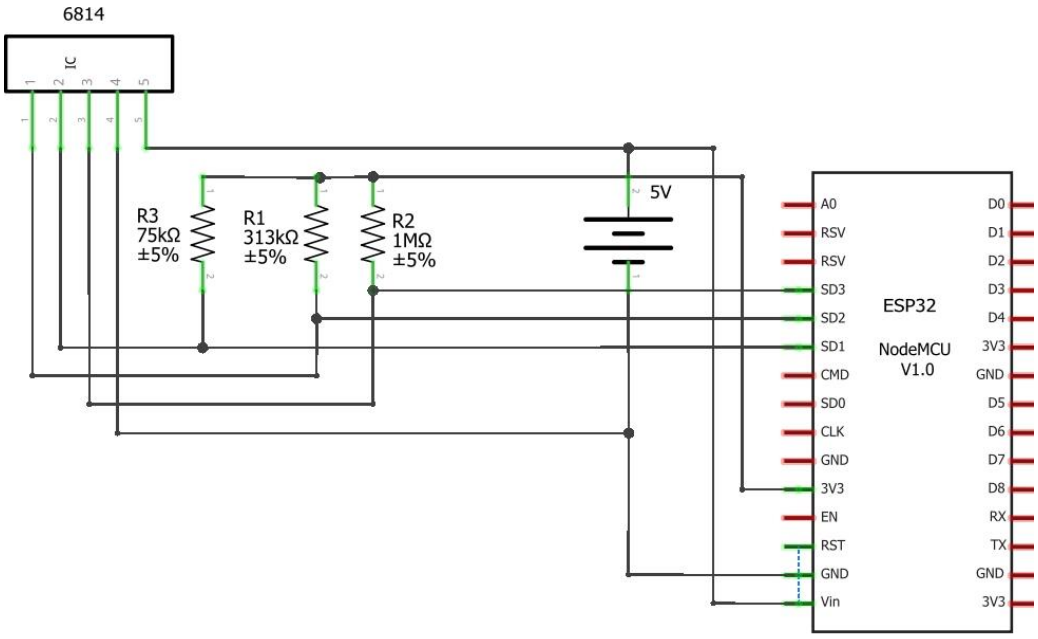


Figure 4-24 CJMCU6814 electrical circuit

The resistor values were selected at 75KΩ, 313KΩ and 1MΩ for NH3, CO, NO2 respectively because, as explained in *MiCS Sensor working principles*, the output voltage of each sensor must be in the middle of the operating range in order to have greater accuracy.

The functions for the correlation of gas concentration and measured resistance ratio were then determined according to the graphs provided by the manufacturer and referred to in paragraph *CJMCU6814* in Figure 4-20, Figure 4-21 and Figure 4-22**.**

Based on these, the following were derived for each gas measured:

CO: $ppmco = 4.4638*x^{-1.177}$, $R^2 = 0.9999$
NO2: $ppmNO2 = 0.1516*x^{0.9979}$, $R^2 = 1$
NH3: $ppmNH3 = 0.6151*x^{-1.903}$, $R^2 = 0.9995$

which will be used in the sensor library to obtain the concentrations.

The library has two main functions:

- calculating the base resistance R0 for each calibrate() gas, each time the sensor is activated, and
- measuring the concentration of a specific gas at the request of the user, measure(gas_t gas) that returns the ppm of the gas gas [CO, NO2, NH3] given as a parameter.

4.2.3 Βιβλιοθήκη CJMCU8128

Like CJMCU's previous sensors, there isn't a library that takes advantage of all three sensors included. In other words, they are treated as three separate sensors and for the needs of this work the ccs811, Adafruit_BMP280 (for BME280) and Adafruit_Si7021 (for HDC1080) libraries included in the Arduino IDE library catalogue were used.

4.3 IAQ Measuring station

The study of the recordings of the sensors and the comparison of their technical characteristics as described in *Evaluation* chapter, led to the selection of suitable components for the construction of the air quality measurement station. The following elements were evaluated and taken into account for the selection (the order of reference does not indicate weight):

- Sensor preparation and initialization time
- Microcontroller communication protocols
- Performance in accordance with its technical characteristics
- Low number of errors during test sessions

Based on them, the following components were selected:

Component	Price
• **Microcontroller ESP32**	5€
• **Nova SDS011**	23€
• **Sensirion SGP30**	10€
• **DHT22**	0,5€
• **ZE25**	30€
• **CJMCU6814**	20€
• **OLED Screen 1306**	2€
• **RGB Led**	0,1€
• **Battery LiPo 2000mAh, 3.7Volt**	3€
• **Box**	3€
Total Cost	3€
	99,6€

Table 4-2 Measuring station's components and cost

In terms of cost, this depends on the country of supply of the components, their availability and of course their number. In the case of this work, the total cost of all the

above amounted to almost 100 euros which is characterized as low for the construction of a single measuring station.

4.3.1 Connection diagram

The Figure 4-25shows the proposed connection scheme of the measuring station to the ESP32 microcontroller, the selected sensors, the power supply for operating and charging the battery, in case there is no continuous power supply.

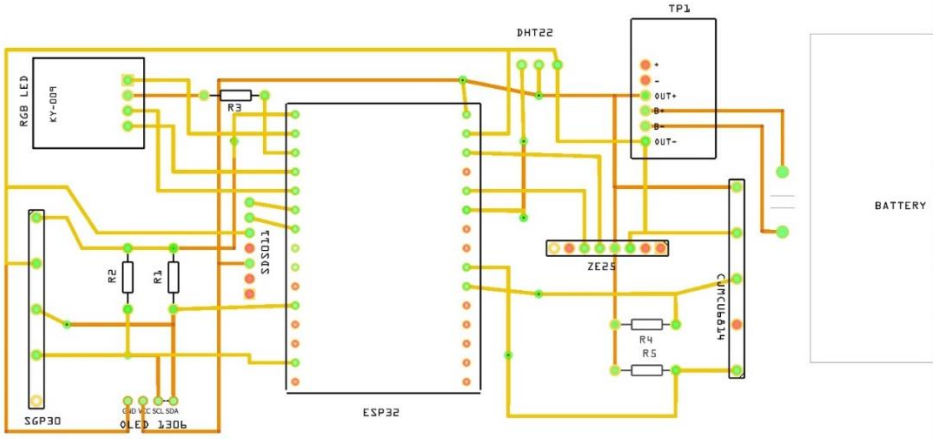


Figure 4-25 Sensor node electrical diagram.

4.3.2 Software

The ESP32 microcontroller software was developed using the Aduino IDE environment in C language with integration of the appropriate libraries for each component/sensor as provided by the manufacturer or developed for this task.

The Setup() function forms the ID of the node using the MAC address of ESP32 and loads various parameters for connecting to the WiFi network and for communicating and exchanging data. If the network connection fails, the local web server starts to configure the node, as described in the *Setting properties* paragraph.

The loop() function calls the functions for each component of the node and sends the measurement packets via the MQTT protocol, in JSON format, by publishing to topic *metrics*. The package structure is described in the InfluxDB measurements section.

The IAQ index is downloaded, as it has emerged through machine learning, by subscribing to the topic *iaqidex*.

4.3.3 Setting properties

Studying the *Operational states* paragraph, one can see that the station can be in perameter adjustment mode. This is done when it fails to connect to the existing wireless network or when the configuration files have not been sent when uploading a new version of its operating software.

In such a case, the station creates a temporary wireless network with an SSID in the form of ESP32XXXXXX and starts its embedded web server. The user connects to the open wireless network and via IP: 192.168.4.1 is given the ability to configure the following parameters:

- WiFi SSID
- WiFi password
- MQTT username
- MQTT password
- MQTT server IP
- MQTT server port
- InfluxDB database name
- Device's Latitude
- Device's Longitude

After setting up and saving the new settings, the device must be restarted.

4.4 Data storing

4.4.1 InfluxDB measurements

The data of each station are sent and stored in an InfluxDB time series database that consists of two measurements called **metrics** and **iaqindex.**

The measurement **metrics** stores all sensor data and the serial number of the measuring station specified during its construction and generated using the serial number of the ESP32 microcontroller used.

Each data package that is sent through MQTT protocol has the following structure shown in Table 4-3.

In the data label of the sensors (capital letters), the location (latitude, longitude) of the device has been incorporated by the microcontroller for the purpose of spatial determination and creation of a map of all sensors. This information has to be configured during station set-up, whenever it changes its location.

Data package	Unit of measurement
[{ "PM25":*FloatValue*,	µg/m^3
"PM10": *FloatValue*,	µg/m^3
"O3": *IntegerValue*,	ppb
"CO": *FloatValue*,	ppm
"NO2": *FloatValue*,	ppb
"CO2": *IntegerValue*,	ppm
"VOC": *IntegerValue*,	ppb
"TEMPRATURE": *FloatValue*,	ºC
"HUMIDITY": *FloatValue*,	% rel.
"PRESSURE": *FloatValue*,	hPa
"latitude": *FloatValue*,	
"longitude": *FloatValue*	
},{	
"sensor": *StringValue*	
}]	

Table 4-3 Sensor station Data package

The time of the data recorded is not sent by the device as it is not uniform for all sensors and moreover the whole system is not time-critical. The data package incorporates from the database its registration time, which is essentially the measurement time.

The measurement **iaqindex** stores data regarding the IAQ index number and color that is calculated for a specific sensor station and the station's ESP32 microcontroller serial number.

To summarize, below on Table 4-4, are presented the two measurements of the project's database.

METRICS	
Data label	**Data type**
Time	
PM25	Float
PM10	Float
O3	Integer
CO	Float
NO2	Float
CO2	Integer
VOC	Integer
TEMPRATURE	Float

IAQINDEX	
Data label	**Data type**
Time	
Color	String
Iaqcolor	Integer
Sensor	String

HUMIDITY	Float
PRESSURE	Float
LATITUDE	Float
LONGITUDE	Float
SENSOR	String

Table 4-4 Measurements and metrics in InfluxDB

4.5 Calculating the IAQ Index using ML

4.5.1 Data Description

The data used to train the model came from a larger dataset as recorded by an airQ sensor for a project of the Aachen University of Applied Sciences [46]. Their recording period extends from 22/3/2023 to 11/7/2023 and the sensor was mounted in a laboratory room and an one-room house. The sensor was recording a total of 31 air characteristics, of which only a subset that is suitable for this project is selected, in 2 minute intervals and a total of 63633 recordings were made.

The set of above features includes several pieces of information that go beyond the goals of this work. As described in the *Data Preparation* section, a subset of them was selected, from which the IAQ index of each measurement was calculated as described in the sectionIAQ *calculation technic* and incorporated into the dataset (Table 4-5).

	PM2.5	PM10	O3	CO	NO2	CO2	VOC	IAQ
1	10.1	10.9	16	1.7	24	493	136	2
2	36	43.7	16	1.6	38	507	342	4
3	34	40.8	16	1.6	37	505	334	4
4	9.6	11.2	13	1.5	37	458	358	5
5	157.5	162.3	13	1.6	37	466	7263	6
6	7.2	9.7	9	1.5	2	422	11	1
7	10.1	11.6	17	1.7	23	485	136	2
8	8	8.8	8	1.5	15	492	18	1
9	21	22	25	1.5	37	431	430	5
10	14.7	15.7	16	1.6	50	510	2896	6
11	11.6	12.2	23	1.4	37	469	262	4
12	0	0	11	1.5	38	527	405	5
13	14.9	15.8	17	1.5	49	483	2017	6
14	0.9	1.3	17	1.5	10	436	155	2
15	10.2	10.5	23	1.4	37	471	269	4
16	0	0.2	5	1.5	22	540	179	3
17	8.2	9.4	15	1.7	34	495	86	1
18	10.1	11.4	16	1.7	24	484	139	2
19	0	0	5	1.5	22	538	175	3
20	0	0.4	5	1.5	22	537	173	3

Table 4-5 Part of dataset used to train ML model, IAQ included

The segment is selected to display all possible IAQ values. As we can observe in lines 6,8,17 all measurements are in the lowest category of the index (Green, IAQ=1), according to **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**. In lines 1,7,14,18 the indicator has a value of 2 (Yellow) due to the increased concentration of VOC. Applying the same method, we can deduce the pollutant that determined the IAQ value for all its values.

4.5.2 Data Preparation

The use of the above data requires a preliminary work to find possible misrecords, missing certain values from a recording or values outside the limits of the sensor or negative.

4.5.2.1 Missing values

Using python data processing tools, the following map was obtained where there is a lack of NO_2 and O_3 data, resulting in the corresponding 597 records being removed.

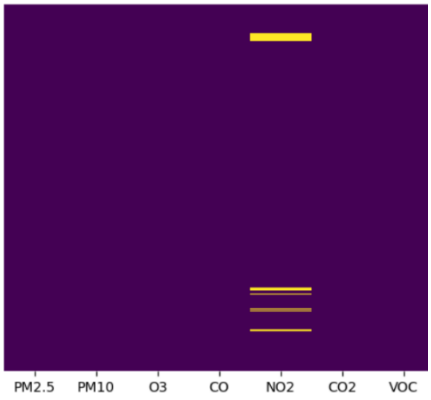


Figure 4-26 Missing dataset values

4.5.2.2 Removing Outliers

An important role in the performance of a system is also played by finding and excluding measurements that exceed the limits of sensors or are considered unrealistic. Visualizing the data as shown in the graphs below, we observe that PM particles show certain peaks with values exceeding the limit of 500μgr/m^3 for PM2.5 and the limit of 604μgr/m^3 for PM10.

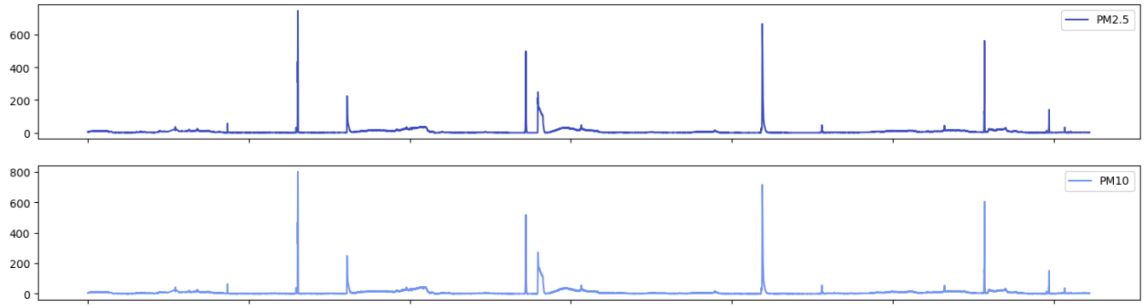


Figure 4-27 PM values

Examining the gases O_3, CO and NO_2 we observe that they are all within normal limits, as shown in Figure 4-28**.**

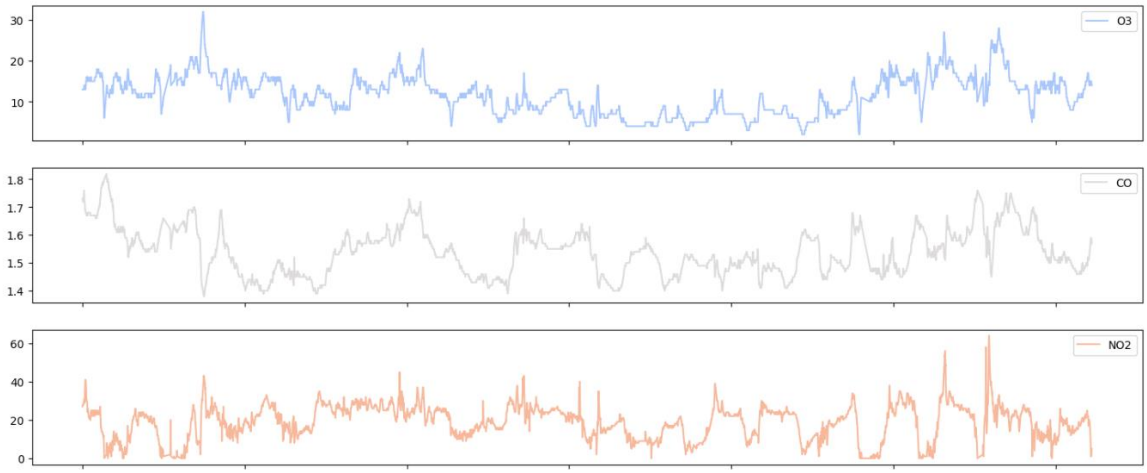


Figure 4-28 O3, CO and NO2 values

Excessive values appear in VOC measurements in many cases exceeding 8000 ppb while reaching up to 60000ppb which is not justified at such a height. These and those of PM are excluded from the dataset.

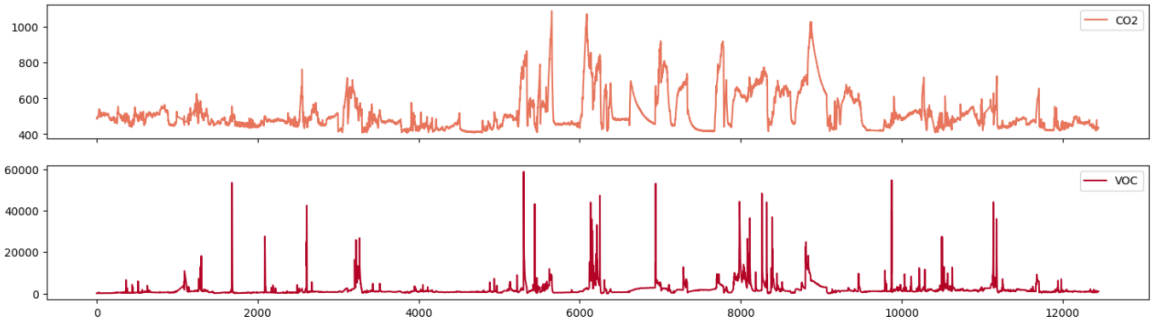


Figure 4-29 CO2 and VOC values

4.5.2.3 Export features

The Pearson correlation coefficient is a way to see how two things relate to each other. It tells us if they are connected in a straight line and how strong that connection is. If the correlation is -1, it means that they are connected in a straight line but in opposite directions. If the correlation is +1, it means that they are connected in a straight line and in the same direction.

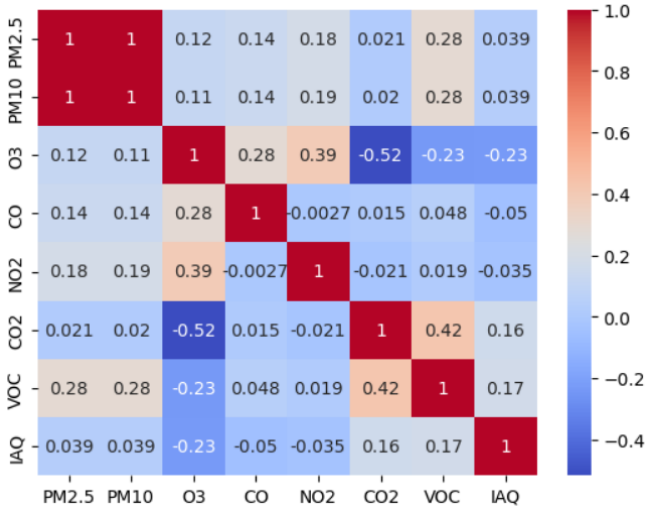


Figure 4-30 Pearson correlation chart

Using the data set after subtractions of the data in the previous stages and using the IAQ indicator, we create the above chart in Figure 4-30.

The following is the selection of those characteristics that will judge the value of the IAQ index and will be used as a whole to train our model, summarized in Table 4-6.

Variable name	Description
PM2.5	Represents the concentration of particulate matter 2.5 micrometers or less in diameter, measured in (µg/m3).
PM10	Represents the concentration of particulate matter 10 micrometers or less in diameter, measured in (µg/m3).
O3	Represents the concentration of ozone (O3), measured in (ppb).
CO	Represents the concentration of carbon monoxide (CO), measured in (ppm).
NO2	Represents the concentration of nitrogen dioxide (NO2), measured in (ppb).
CO2	Represents the concentration of carbon dioxide (CO2), measured in (ppm).
VOC	Records the total volatile organic compounds (TVOC) concentration, measured in (ppb).

Table 4-6 Dataset's variable description

4.5.3 Algorithm Selection

To develop the ML model, two classification algorithms, K-NN and SVM, were implemented in order to compare their results, to evaluate and choose the best, as the IAQ is classified into six classes. The models were developed using Python and its sklearn libraries.

4.5.3.1 K-nn evaluation

In order to evaluate th K-nn algorithm, were performed some with different values of the K-neighbor parameter and the separation rates of the Train and Test dataset. It was initially tested with K=5, 70% train set and 30% test set, which resulted in an accuracy of 0.99421 and a KappaScore = 0.9920549.

Following, the value of K fluctuated in between [2..89] with appreciable droping in the value of accuracy and the highest accuracy value is achieved when K=7 , as shown in Figure 4-31.

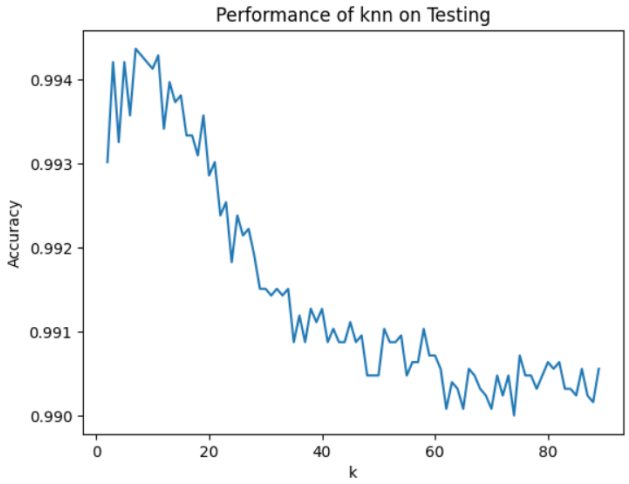


Figure 4-31 Performance of knn on testing phase

Continuing, different ratios of Training and Testing sets were tested where the percentages of the test set ranged from 10% to 50%, with the results shown in Figure 4-32**.** In this we observe that the highest accuracy is achieved when we have 10% test set and 90% training set, followed by a slight drop, although the accuracy does not fall below 0.99 in any case, indicating the correctness of the algorithm and the success of predictions.

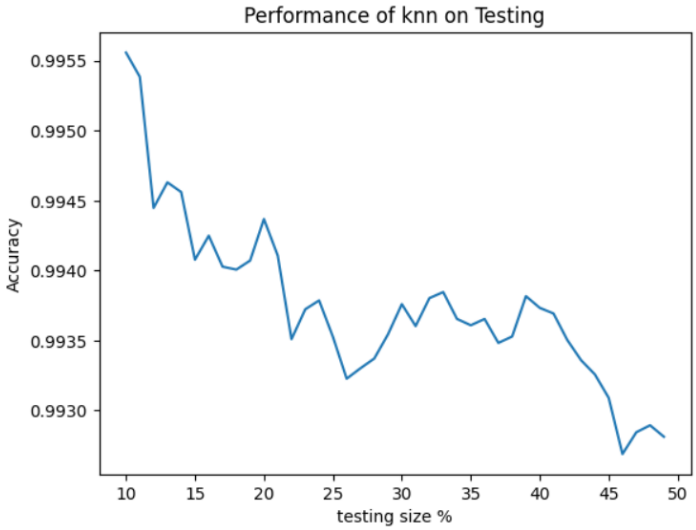


Figure 4-32 Performance of knn on various testing percentages

4.5.3.2 SVM evaluation

The training of the ML model using the SVM algorithm was done using the same initial parameters to separate the dataset with the K-nn algorithm, i.e. 70% train set and 30% test set and was selected as parameter C=1.

The SVM algorithm uses various cores to extract the results, such as Linear, Gausian, Polynomial. Starting with Linear we get as initial accuracy 0.544 and KappaScore equal to 0.47644, which are not satisfactory. By testing a different test and train set rate, there is little improvement when the separation is 20% - 80% respectively, as shown in Figure 4-33. Applying this improvement we see that accuracy improves, with a value of 0.663 and KappaScore at 0.488.

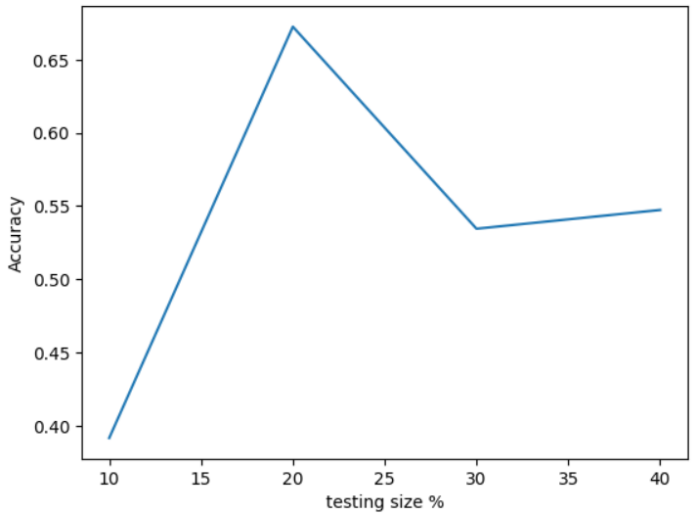


Figure 4-33 Performance of Linear SVM

Also, an application was performed using polynomial SVM, starting with the same initial parameters of the dataset, obtaining the results accuracy=0.7931 and KappaValue=0.7155. With additional changes in its parameters, it is observed that the performance drops quite a bit, as shown in Figure 4-34.

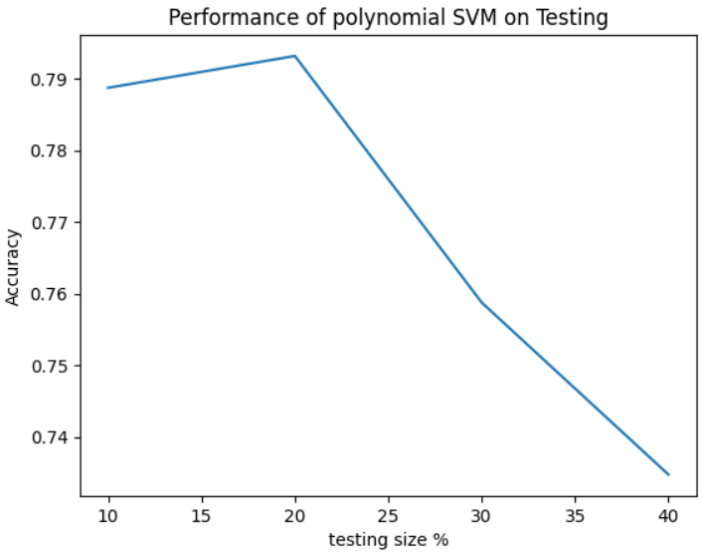


Figure 4-34 Performance of polynomial SVM on Testing

4.5.3.3 Results

Significant and expected improvement of model performance using the polynomial SVM algorithm, as the classes are 6 to 7 characteristics for classification in them. But its performance does not come close to that of the K-nn algorithm which is almost ideal. In

conclusion, therefore, the choice of the latter is a one-way street with such performance indicators.

4.5.4 ML Application in Node-Red

The selection of the most appropriate ML algorithm and its implementation as described in the previous paragraph, helps us move on to the next step, which is the extraction of the model and the ability to access it, in order to extract the IAQ index of each measurement station.

Using the joblib library of Python, the model is exported and can be used by the NodeRed script developed for this purpose as shown in Figure 4-35. The script is controlled by a trigger that is triggered every 1 minute, requesting from the measurement metrics in InfluxDB, the averages of the last 10 minutes of station measurements, by executing the query:

```
SELECT mean(PM25) as PM25, mean(PM10) as PM10, mean(O3) as O3,
mean(CO) as CO, mean(NO2) as NO2, mean(CO2) as CO2, mean(VOC) as
VOC FROM metrics WHERE time >=now() - 10m  GROUP BY sensor
```

The data is then divided into independent messages, one for each measuring station with the *Split Results per Sensor* node function. In the node *Create the array of data*, the layout of the averages of each message is modified, so that it is in array format [PM2.5,PM10,O3,CO,NO2,CO2,VOC]. This is required by the ML model as in this form it must be its input to function properly. They are then led to a queue from which, every 20 seconds, a data packet is directed to the *Predict IAQ node* that loads the ML model to calculate the IAQ pointer. The result is an integer [1..6], appended to the message with the rest of the measurement station data on the *Append IAQ 2* node.

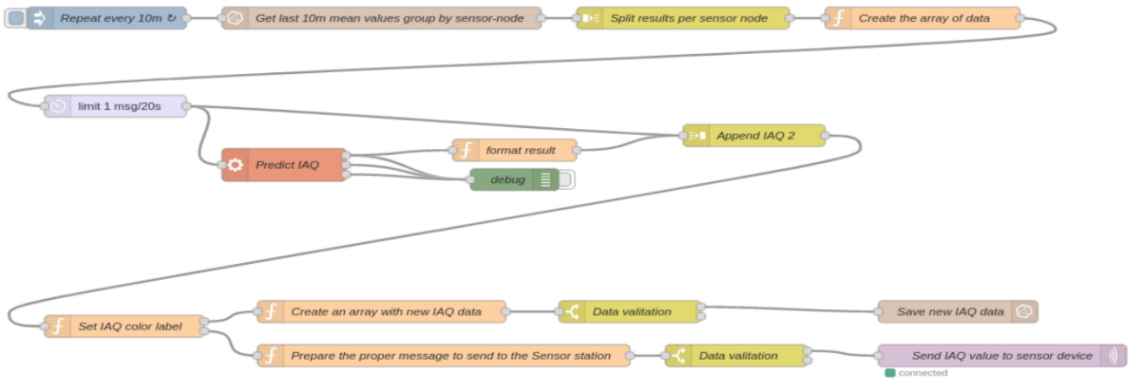


Figure 4-35 Calculating IAQ index

The *Set IAQ color label* function specifies the color of the index according**Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**, and in *Create an array with new IAQ data*, the data are organized in array format to be sent to the database. This is done in the *Save new IAQ data* node, which sends it to the *iaqindex* measurement. In the function *Prepare the proper message to send to the Sensor Station*, the message is configured in an appropriate format and is going to be sent to the measuring station whose IAQ has just been calculated, using the *Send IAQ value to sensor device* node that sends the message via the MQTT communication protocol. Finally, the *Data validation* function performs a final check for the correctness of the data before it is sent to its destinations.

4.6 Displaying timeline data and geolocation of measuring station

4.6.1 Measurement period

The data of all nodes, once collected and stored in InfluxDB, must be made visible and accessible to oversee the proper operation of the nodes and indoor air quality. For this reason, a platform for their supervision and presentation was created, using the Grafana tool and some additional functions it provides.

The tests were conducted in a room of a house where the family gathers, initially in winter season, for 12 consecutive days and then in summer period for 5 consecutive

days. The choice of these periods was made because they present great meteorological differences but also differences in human habits.

During the winter season, ventilation openings (doors, windows) are almost always closed, central heating (radiators) works and residents stay more hours inside the house. On the contrary, in the summer season, residents are outdoors, do not use heating and have the windows open almost all day.

The measurement area had an area of about 40m^2, a ceiling height of 3m, had four windows and two doors, one of which was external, and there was a closed type fireplace that was used a few times during the mentioned period.

4.6.2 Commentary on recordings

Below are some characteristic snapshots of the measurements, as taken during the trial periods of the measuring station and the calculation of the air quality index, which are interesting.

The Figure 4-36 characteristically shows the change in the concentration of PM particles indoors in relation to the activity of the inhabitants of the house. As it seems, there is a gradual increase from 25μg/m^3 at 06:30 when activity begins with a peak of 53μg/m^3 at about 10:00, with the next period de-escalating as there was their departure and at the same time opening windows with the outside air cleaning the interic

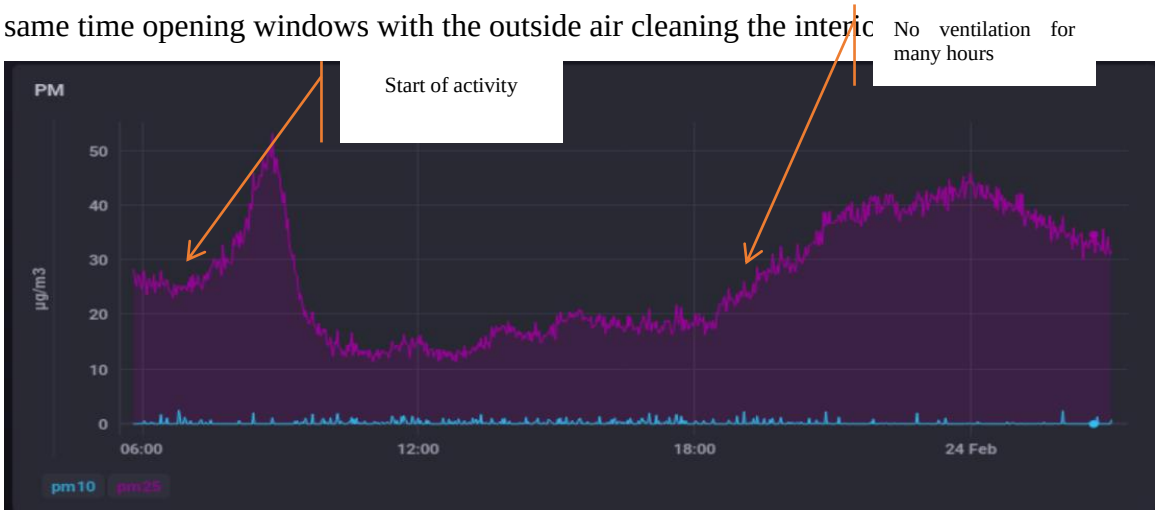


Figure 4-36 PM particle concentration recording, winter measurement period

Similar recordings are observed in CO_2 and VOC measurements that clearly show the hours of human presence in space, with their simultaneous increase in the aforementioned hours, as shown in Figure 4-37**.** Also towards the end of the day there is a gradual decline in values, indicating the departure of people from the space, their transition

to the bedrooms, without there being a sharp drop in measurements as happened in the morning hours.

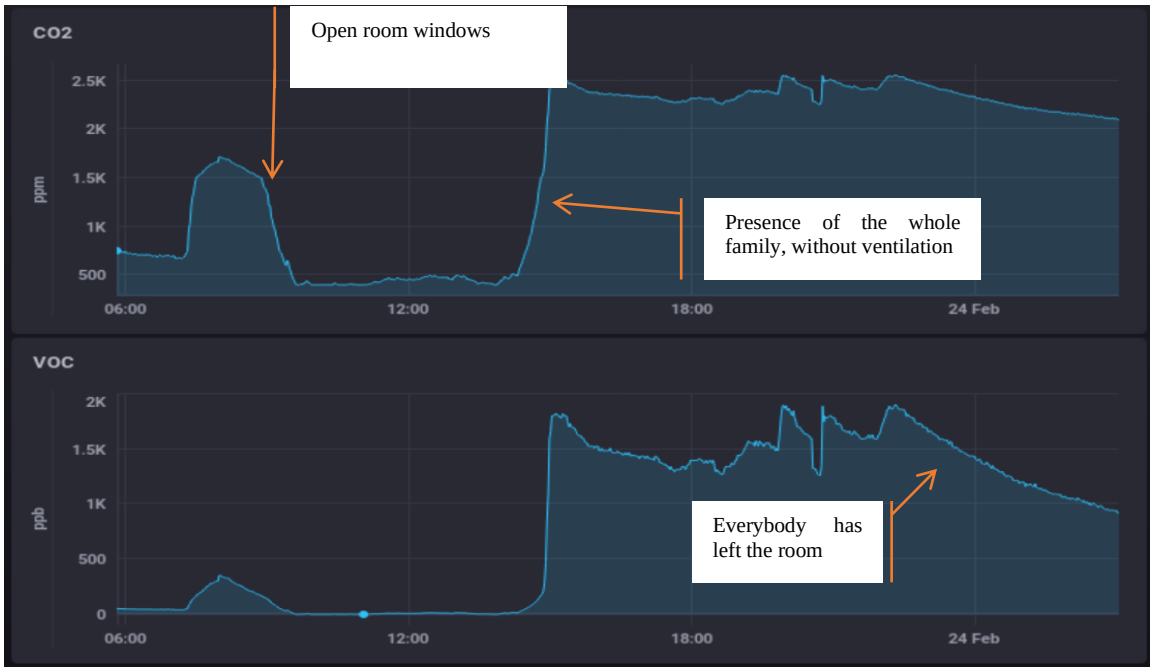


Figure 4-37 CO2 and VOC concentration recording, winter measurement period

Regarding the rest of the gases, we can note that there are no records of NO_2 and CO concentration in indoor spaces, while O_3 ozone values move at levels characterized from ideal (20 – 55 ppb) to moderate (56 – 75 ppb) as shown in the Figure 4-38**.**



Figure 4-38 Ozone concentration O3, winter measurement period

On the contrary, during the summer season, as shown in the Figure 4-39, we observe that ozone levels are much higher, moving at values from 25 ppb to just over 100 ppb, affecting the IAQ index as the upper values indicate that air quality is unhealthy for sensitive groups or for everyone.



Figure 4-39 Ozone O3 concentration, summer measurement period



Figure 4-40 NO2 concentration, summer measurement period



Figure 4-41 CO concentration, summer measurement period

The IAQ index is presented in two ways, as a chart (Figure 4-42) to show its change and as a circle of variable diameter on a map (Figure 4-43), showing the location of the measuring station, its characterization according to the color code as mentioned in the paragraph *IAQ calculation technic*.

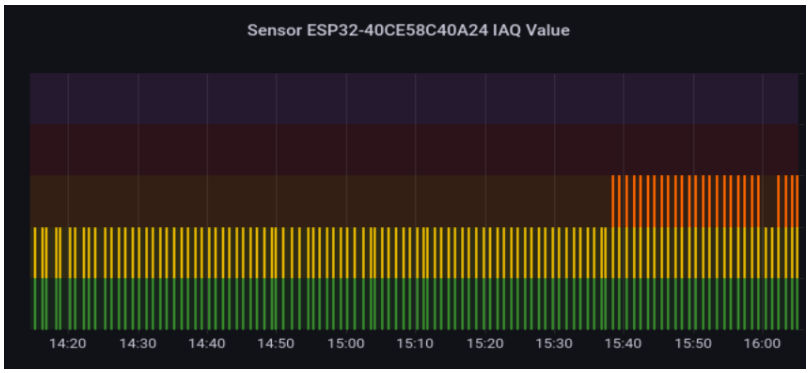


Figure 4-42 IAQ value chart

Grafana allows the installation of plugins such as World Map that uses data from OpenStreetMap and with proper data manipulation can be configured to show both location and IAQ rating. More specifically, by setting the size limits of the circle to 2 to 15 and setting the Threshold options to as many distinct values as the possible IAQ values, the circle will change in size according to its value and will vary its color according to its specifications.

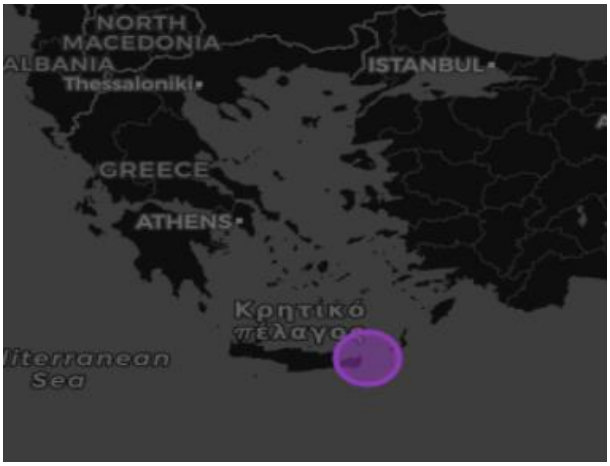


Figure 4-43 Position of the sensor

The position is determined by deriving the coordinate values of the measurement station from the InfluxDB database *with the query SELECT last("latitude") AS "Lat", last("longitude") AS "Lon" FROM "metrics" WHERE ("sensor" =~ /^$sensor$/) AND $timeFilter GROUP BY "sensor"* and with the corresponding setting of Latitude and Longitude Fields.

The pointer value is pulled back from InfluxDB with query *SELECT last("iaqcolor") AS "IAQ" FROM "iaqindex" WHERE ("sensor" =~ /^$sensor$/) AND $timeFilter GROUP BY "sensor"*.

The two queries are filtered (WHERE *("sensor" =~ /^$sensor$/)* to obtain only the metering station-specific data selected by the user from the station list in Grafana. These names are unique and cannot be changed by the measuring station property setting because they are automatically defined according to the ESP32's MAC address.

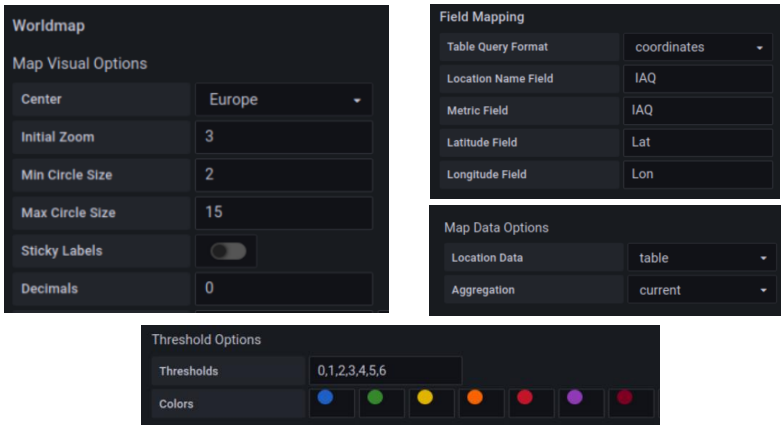


Figure 4-44 Properties of the World Map add-on

Chapter 5. Evaluation

5.1 Particulate Matter Sensors Experiment

The choice of a measuring device should be based on objective data which can be obtained by comparing it with a corresponding, more accurate and certified for the accuracy of its measurements. For this reason, an experimental schema was created consisting of an airtight container, an air supply pump and the precision measuring device, as show below. The experiment conducted at the Department of Chemistry of the University of Crete, which provided us with his equipment for use that was consisted by:

- a DustTrak II Aerosol Monitor 8530 which is a desktop data-logging, single-channel, light-scattering laser photometer that gives real-time aerosol mass readings and collects a gravimetric sample. It uses a sheath air system that isolates the aerosol in the optics chamber to keep the optics clean for improved reliability and low maintenance and can measure aerosol contaminants such as dust, smoke, fumes, and mists.
- A container filled with PSL spheres 200nm to create aerosol particles and simulate PM pollution.
- A qubic airtight container with 0,5 m acme
- An air supply pump used to push the PSL spheres into the container

For the purposes of our testings, a device for storing the measurements of the PM sensors, that we want to compare, was also constructed which consisted of the following:

- ESP32
- Honeywell HPMA115S
- Plantower PMS7003
- Nova SDS011
- Sensirion SPS30

All four pm sensors were connected as shown in the Figure 5-1 below.

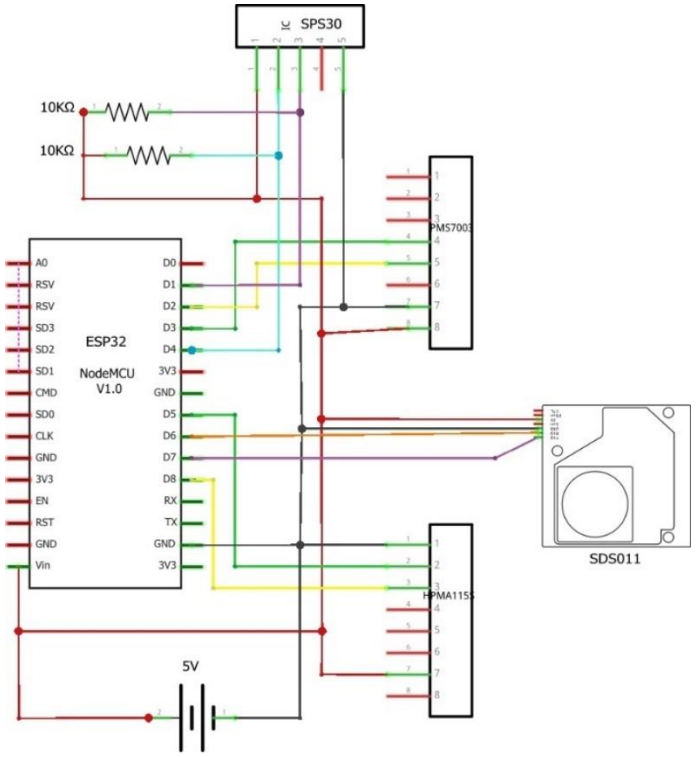


Figure 5-1 Electrical circuit for the expirement sensor station

The first experiment stage consists of releasing the mixture in the chamber, in order to gradually increase pm concentration. When the concentration of pm has exceeded maximum concentration in environmental conditions, the pump and the chemical mixture are disconnected. By this way, the air is only allowed to exit gradually, passing the DustTrak sensor, until the chamber air is empty.

Experiment time (sec)	Experiment Phases
0	Release of mixture and initiation of forced air flow into the chamber passing through the chemical mixture container
1380	Disconnect mixture and stop forced air intake into the chamber. Air is only allowed to exit gradually
3720	Emptying chamber's air. Measurement is complete.
3900	Chamber empty and instrumentation zeroing
4320	End of experiment

Table 5-1 Experiment phases

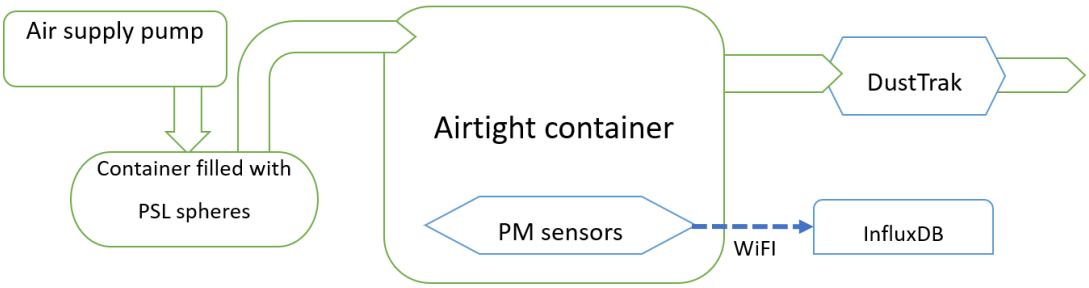


Figure 5-2 Block diagram of the experiment connections

Every sensor measures pm concentration every 10 seconds and stores the values in a InfluxDB timeseries database. The DustTrak on the other hand, records data in a csv file every second. All data were extracted to be compared as shown in the Figure 5-3.

The value in y axis is the measured pm concentration as μgr/m^3 and in x axis is the experiment duration in seconds.

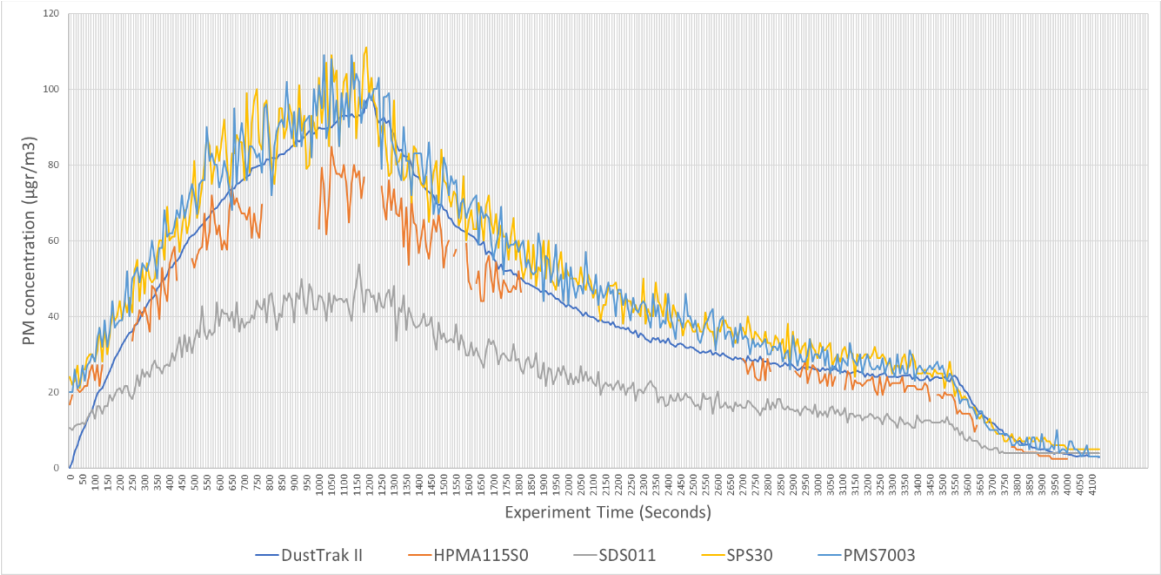


Figure 5-3 Recorded values of different sensors

The DustTrak is the benchmark of our measurements as a highly accurate instrument, with the measurements of the other sensors being compared to it.

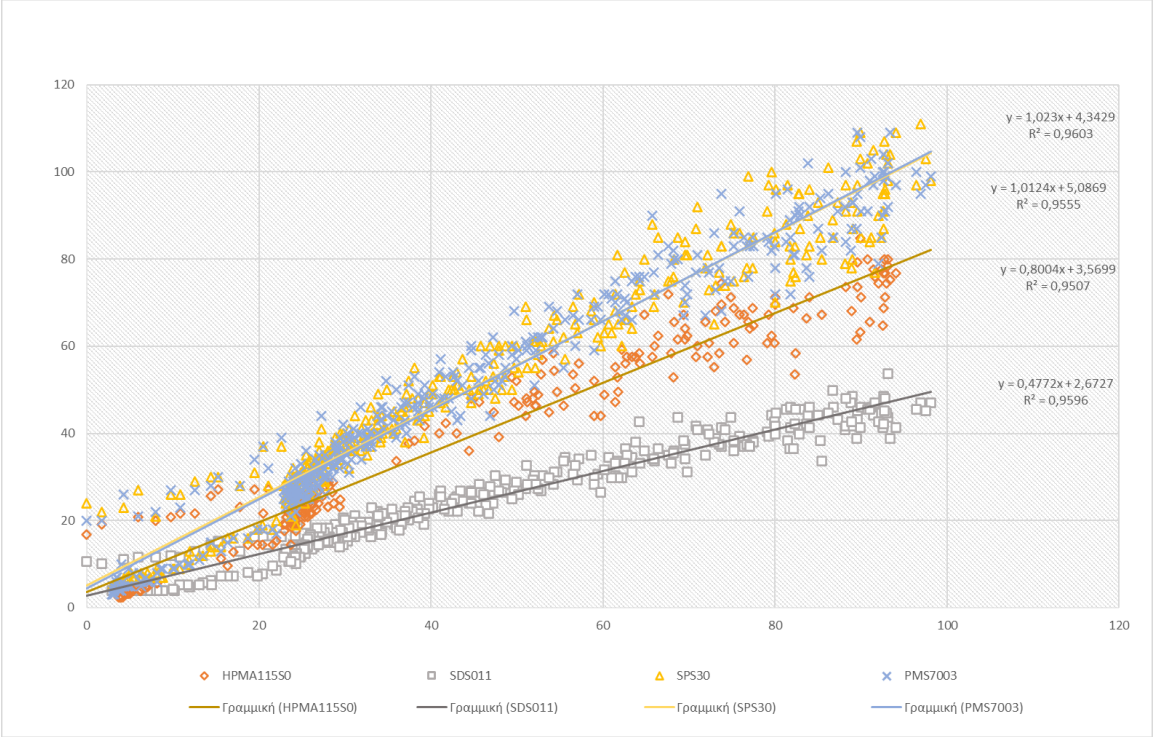


Figure 5-4 Scatter chart and trendlines

The scatter chart and the linear trendlines of the sensors was created in conjunction with DustTrak (Figure 5-4).

As we can observe, there is a significant dispersion of the recorded values from the sensors PMS7003, SPS30 and HPMA115S as the analysis showed that they had a standard deviation of 28.3, 28 and 23.5 respectively, although their values are certainly closer in absolute number with the measurements of DustTrak. It was also observed that the sensor HPMA115S often showed an inability to measure as shown in figure 5-3, it did not send data in 41% of the number of measurements of the experiment, but that was not justified by the recordings of its library nor by checking the error file of the ESP32 software. The SDS011 measurements are much lower, however, shows remarkable consistency in measurements, with data deviation of 13.2, much better than the others. This shows the quality of the sensor and the algorithm for calculating the concentration that its microcontroller "runs".

For the above reasons we conclude that the most reliable sensor is the Nova SDS011, which we will choose to incorporate into the air quality measurement device.

5.2 Testing the rest sensors

A measuring station was created to test how sensors measuring gases of the type VOC, CO_2, CO, NO_2, O_3 work. As said in the paragraph *Sensors Software Libraries*, the appropriate library of each component is not always provided by manufacturers so it has to be created from scratch, or they do not provide libraries for reading data using the ESP32 microcontroller. It took several tests and experiments to achieve stable communication with the sensors, select the correct communication protocol (Serial /I2C /Analog) and correctly read the data by ESP32.

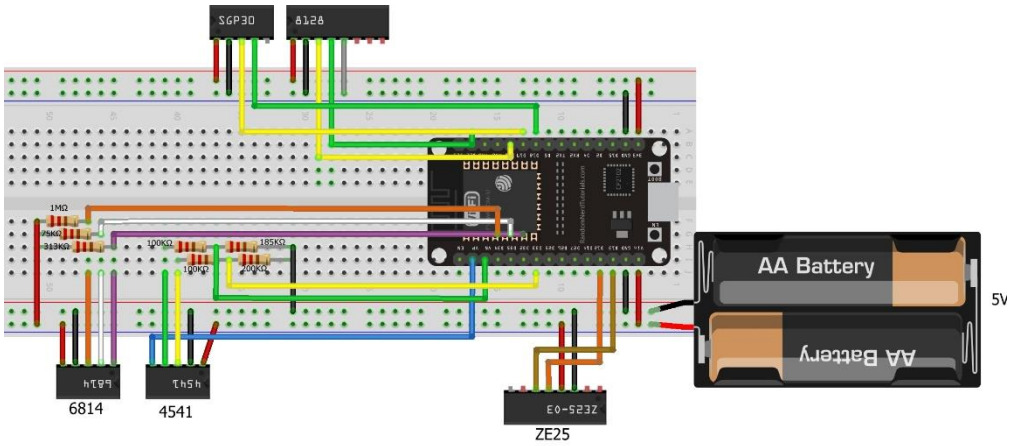


Figure 5-5 Connection diagram of the testing device

The tests were conducted in various indoor environments (home, classroom) for a few days to observe their functionality. Due to the manufacturing method as metal oxide sensors, they should be in fresh air during their first phase of operation which can last many hours (about 24 to 48). Their measurements, as stated in the specifications of the manufacturers, are not completely accurate numbers as they do not attribute the concentration of the measured gas in space but the change in its concentration. For the following reason, it is recommended to calibrate the sensor with a certain amount of resistance, which will vary and depends on the place (a house, a mall, a cinema, a club). [45] [44].

The CJMCU6814 was connected to the ESP32 with a resistor in a pull-up configuration and set each output pin to give half of the 3.3V operating voltage as it is explained in *MiCS Sensor working principles*. For this purpose trimmer potentiometers were placed on CO, NH_3 and NO_2 pin. The datasheet suggests measuring the voltage

between the sensor and ground [45], but this method could not be made to function, leading to the conclusion that the pins need to be raised.

The SGP30 sensor proved to be the easiest to connect to the ESP32 and read its measurements because the manufacturer provided it with the corresponding library in C. In addition, it runs an internal algorithm that calculates VOC and CO_2 concentrations based on sensor's raw signals, taking moisture into account.

As regards the CJMCU8128 sensor, it took about 24 hours to get the readings down to 400 ppm from the initial 1000 plus value, while it was placed outside to be calibrated. This sensor includes the CCS811, an oxidizing sensor that emits heat when it is on, affecting the temprature and humidity measurements of the HDC1080 and BME280. To come to this conclusion, the data were compared with those of a DHT22 sensor which was placed next to it and a +4 ^{o}C deviation was observed.

Also this sensor family has an oxidation/reduction chemical reaction with the VOC gases. However, the sensor's sensitivity to CO_2 is minimal because of the measurement that is based on the incorrect assumptions that humans are the only CO_2 producers and that the sensor is located within a structure. The gas sensor analyzes volatile organic compounds, assumes they are produced by humans, converts that to the number of people present, and then converts that to the amount of CO_2 they would produce .

Finally, the ZE25-O3 sensor, which measures ozone concentration, showed stable operation and easy connectivity. The measurements showed variations both in a short period of time and with the passage of some months and the seasonal change of meteorological conditions.

Chapter 6. Conclusion

This thesis aimed to study the possibility of constructing an indoor air quality monitoring system and to propose new ways of calculating the quality index, according to the available software tools and at low cost. Existing indicators refer to outdoor spaces, usually the focus is cities, and the reference intervals from which the Air Quality Characterization (AQI) is derived are from eight to twenty-four hours. Indoors this does not make sense to apply as air quality can be affected by more factors such as outdoor openings, air conditioning, room volume, machinery/appliances in operation and the number of people.

The proposal of this thesis is the creation of an Indoor Air Quality Index (IAQ) based on the existing standard categories and color codes of the US EPA but differentiated in the calculation time which is rolling ten minutes, in order to detect changes faster. In addition, it is proposed to add the measurement of CO_2 and VOC, which are indicators of human presence which, when high, affects air quality.

It was found that there are many sensor options that could be studied and used, at an affordable cost, from which a small subset of them was selected, covering a wide range of measurements of pollutants and other gases that may exist inside buildings or other enclosed spaces.

The study of these showed that in terms of their technical characteristics there are not great differences and what is observed is not a reason for rejection of some. However, the support and extra features given by the manufacturers of some are a criterion for giving preference over others. MiCS sensors are concerned about the long warm-up time it takes to give consistent results and the lack of an official library of operating software.

Sensors such as the CJMCU8128, which incorporates three smaller ones, were observed to give temperature data that had been affected by CCS811 which, as a metal oxide (MOX) gas sensor, must be preheated to function properly.

Through the experiments and test periods of the PM sensors, a qualitative difference was found between them, with the HPMA115S0 showing errors and incomplete recordings, while SDS011 stood out with its internal algorithm for calculating microparticle concentrations. For this reason it was the choice in this category and was incorporated into the measuring station.

Using the MQTT protocol and the InfluxDB time series database, sending, receiving and storing data is very simple and fast. Build a page that displays the recorded data of each node, its IAQ index and its position on a map using geographical coordinates. In case of multiple measuring stations, it is possible to select from a list of some, by dynamically converting the displayed graphs.

The measuring station consists of one sensor for each category of measured gaseous pollutant with ESP32 as their connection point with the corresponding software. The ESP32 microcontroller showed stable, reliable operation at all test intervals and with the support of communication protocols (I2C, Serial, Analog) made it possible to connect all components without problems. The total cost of procurement of the devices was relatively low for the capabilities it offers. The the existence of a GPS sensor so that location determination is done automatically and more accurately.

It is interesting to calculate the IAQ using machine learning as, after training and testing the model, the process becomes much simpler with the help of the appropriate NodeRed library and the layout created. The calculation weight is transferred from the end of the system to a central point, so that there is a clear separation of functions, giving the advantage of improving one (IAQ ML model) without having to upgrade the other system (measuring station).

Based on what was mentioned in this thesis, it can be said that the field of air quality measurements offers several possibilities for development and integration of new tools. The plethora of sensor options gives many options for low-cost testing. The lack of standardized measurement methods and the corresponding delayed adoption of similar initiatives by states, gives possibilities for further proposals to agree on common standards among societies.

References

[1] M. Lazaridis, Dzumbova, L and Kopanakis, I., "PM10 and PM2.5 Levels in the Eastern Mediterranean (Akrotiri Research Station, Crete, Greece).," *Water Air Soil Pollut 189, 85–101,* 2008.

[2] Z. Peng and J. L. Jimenez, "Radical chemistry in oxidation flow reactors for atmospheric chemistry research," *Chem. Soc. Rev,* pp. 2570-2616, 2020.

[3] T. Grant, Brigham, E. P and McCormack, M. C., "Childhood Origins of Adult Lung Disease as Opportunities for Prevention.," *The journal of allergy and clinical immunology. In practice, 8(3),* p. 849–858, 2020.

[4] D. Brdarić, Kovač-Andrić, E. and Šapina, M., "Indoor air pollution with benzene, formaldehyde, and nitrogen dioxide in schools in Osijek, Croatia.," *Air Qual Atmos Health 12,* p. 963–968, 2019.

[5] Balakrishnan and Kalpana, "The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017".*The Lancet Planetary Health, Volume 3, Issue 1, e26 - e39.*

[6] Frederickson LB, Lim S, Russell HS, Kwiatkowski S, Bonomaully J, Schmidt JA, Hertel O, Mudway I, Barratt B and Johnson MS, "Monitoring Excess Exposure to Air Pollution for Professional Drivers in London Using Low-Cost Sensors," *Atmosphere*, p. 11(7):749, 2020.

[7] "World Air Quality Report," IQAir, 2019.

[8] S. Grossberndt (NILU), A. Bartonova (NILU) and A. G. Ortiz (EEA), "Public awareness and efforts to improve air quality in Europe," EEA, 2021.

[9] H. Liu and Salvo, A., "Severe air pollution and child absences when schools and parents respond.," *Journal of Environmental Economics and Management,* pp. 92, 300-330, 2018.

[10] H. Chojer, Branco, P. T. B. S, Martins, F. G., Alvim-Ferraz, M. C. M and Sousa, S. I. V., "Development of low-cost indoor air quality monitoring devices: Recent advancements," *Science of The Total Environment,* pp. 727, 138385, 2020.

[11] F. Franceschini, M. Galetto and D. Maisano, "A short survey on air quality indicators: properties, use, and (mis)use," *Management of Environmental Quality,* 2005.

[12] "Overview of India's National Air Quality Index," 2015. [Online]. Available: https://aqicn.org/faq/2015-05-15/india-national-air-quality-index/.

[13] N. R. Dalezios, A. Blanta, A. Loukas, M. Spiliotopoulos, . I. N. Faraslis and N. Dercas, "Satellite methodologies for rationalising crop water requirements in vulnerable agroecosystems," *International Journal of Sustainable Agricultural Management and Informatics,* pp. 37-58, 2019.

[14] Demanega, Mujan, Singer, Anđelković, Babich and Licina, "Performance assessment of low-cost environmental monitors and single sensors under variable indoor," *Building and Environment,* 2020.

[15] A. Martinez, E. Hernandez-Rodríguez, L. Hernandez, O. Schalm, R. A. González-Rivero and D. Alejo-Sánchez, "Design of a low-cost portable system for the measurement of variables associated with air quality," in *XII Congreso Argentino de Sistemas Embebidos (CASE 2021),* Argentina, 2021.

[16] M. LOPES, J. REIS, A. FERNANDES, L. DIOGO, R. LOURENÇO, T. NUNES, C. FARIA, C. BORREGO and A. CESAM, "INDOOR AIR QUALITY STUDY USING LOW-COST SENSORS," *WIT Transactions on Ecology and the Environment,* no. 244, 2020.

[17] M. Narayana, D. Jalihal and S. Nagendra, "Establishing A Sustainable Low-Cost Air Quality Monitoring Setup: A Survey of the State-of-the-Art," *Sensors,* 2022.

[18] T. Baur, Schultealbert, C, Schütze, A. and Sauerwald, T., "Novel method for the detection of short trace gas pulses with metal oxide semiconductor gas sensors," *Journal of Sensors and Sensor Systems,* pp. 411-419, 2018.

[19] E. Bagkis, Kassandros, T, Karteris, M., Karteris, A and Karatzas, K, "Analyzing and improving the performance of a particulate matter low cost air quality monitoring device," *Atmosphere,* p. 251, 2021.

[20] N. Brilis, C. Foukaraki, E. Bourithis, D. Tsamakis, A. Giannoudakos, . M. Kompitsas, T. Xenidou and A. Boudouvis, "Development of NiO-based thin film structures as efficient H2 gas sensors operating at room temperatures," *Thin Solid Films,* pp. 8484-8489, 2007.

[21] D. Theodoritsis, *Comparing Iot Systems,* 2019.

[22] L. A, S. E and P. R., "Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications," *World Meteorological Organization,* 2018.

[23] P. Simon, Too Big to Ignore: The Business Case for Big Data, Wiley.

[24] "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/.

[25] S. Raschka and V. Mirjalili, Python Machine Learning, 2017.

[26] "MQTT," [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html.

[27] "InfluxDB," [Online]. Available: https://www.influxdata.com/products/.

[28] "Grafana," [Online]. Available: https://grafana.com/grafana/.

[29] "NodeRed," [Online]. Available: https://nodered.org.

[30] S. M. Saad, A. Y. M. Shakaff, A. R. M. Saad, A. M. Yusof, A. M. Andrew, A. Zakaria and A. H. Adom, "Development of indoor environmental index: Air quality index and thermal comfort index," in *AIP Conf. Proc.*, 2017.

[31] *Nova SDS011 Laser PM2.5 Sensor Specifications,* Nova Fitness Co., Ltd, 2015.

[32] *PMS7003 series data manual,* Plantower, 2016.

[33] *SPS30 datasheet,* Sensirion, 2020.

[34] J. Tryner, Mehaffy, J., Miller-Lionberg, D. and Volckens, J., "Effects of aerosol type and simulated aging on performance of low-cost PM sensors," *Journal of Aerosol Science,* September 2020.

[35] *HPM Series Particulate Matter Sensors datasheet,* Honeywell, 2019.

[36] *DHT22 Specifications,* Aosong Electronics Co.,Ltd.

[37] S. Afzali, "Measurements of indoor and outdoor concentrations of particulate matter in the Athens area," 2012.

[38] *SGP30 Sensirion Gas Platform Datasheet,* Sensirion, 2017.

[39] *CCS811 Datasheet,* AMS, 2019.

[40] *BME280 Datasheet,* Bosch, 2018.

[41] *HDC1080 datasheet,* Texas Instruments, 2016.

[42] *ZE25-O3 user manual,* Winsen, 2017.

[43] *MiCS Application Note,* SGX Sensortech.

[44] *MiCS 4514 datasheet,* SGX Sensortech.

[45] *MiCS-6814 datasheet,* SGX Sensortech.

[46] W. Poser, "Indoor Air Quality Dataset Germany," Aachen, 2023.

[47] Spiliotopoulos, *IoT Low Cost System for Air Quality Monitoring,* 2019.

[48] Chee Hoo Kok, Mohd Azrul, Mohd Azlan and Soon Ee Ong, "IoT based Low Cost Distributed Air Quality Monitoring System for Big Data Collection," in *10th IEEE*

International Conference on Control System, Computing and Engineering (ICCSCE2020), Penang, Malaysia, 2020.

Appendix A - Software used during the development of this thesis and for the operation phase of the system

Software	Version	Field of use
Arduino IDE	1.8.X	Environment in C for the development of • sensors' libraries • sensors' testing software • sensor node software
Jupyter	7.0.6	Develop and test the ML model for IAQ prediction, using different ML algorithms.
Oracle VirtualBox	6.1.26	Create a VM, as the server of the system, that hosts every other data management application
Linux Ubuntu	5.15.0-43 x64	The operating system of the server
Mosquitto	2.0.14	Communication protocol that is used for transfer raw data from the sensor nodes to the database sending the predicted IAQ to the sensor nodes
InfluxDB	1.8.10	Timeseries database that stores raw data and air quality indecies
Node-Red	3.1.0	Flow-based programming tool that uses the pre-trained ML model to predict IAQ, based on raw data from the InfluxDB, for each sensor node that is in use
Grafana	9.0.6	Interactive data-visualization platform used to display the metrics, the position and the IAQ of each sensor node