Understanding Plant Communication With Commercial Sensors

Thesis/Dissertation by

Ryan P. Lenfant

In Partial Fulfillment of the Requirements

For the Degree of

Masters of Computer Science

University of Virginia

Charlottesville, Virginia

https://orcid.org/0009-0002-8022-1024

Approval Sheet

This

Thesis

is submitted in partial fulfillment requirements

for the degree of

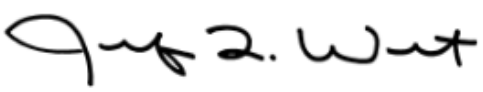
Master of Computer Science

Author: Ryan Philippe Lenfant

This Thesis has been read and approved by the examing committee:

Advisor:	Brad Campbell
Committee Member:	Arsalan Heydarian
Committee Member:	Michael Timko

Accepted for the School of Engineering and Applied Science:



Jennifer L. West, School of Engineering and Applied Science

May 2023

ABSTRACT

Understanding Plant Communication With Commercial Sensors
Ryan P. Lenfant

Plants communicate using terpenes, a form of volatile organic compounds (VOCs). These signaling chemicals are released due to abiotic and biotic factors, such as pathogens present in the air, defense against predators, or attraction of pollinators. Recent advances have shown that plants can be genetically modified to release terpenes in the presence of COVID-19. By combining genetically modified plants and VOC sensors, we can create energy efficient sensors that be genetically modified to sense different pathogens in the air. While some research exists on using plants as sensors, there are little to none that use commercially available cheap sensors to interpret the plant. Our study aims to integrate plant emissions into existing digital systems to interpret plant terpenes. We used 16 terpenes and the air quality sensors to determine which terpenes can be detected by these sensors. Monoterpenes such as Alpha-Terpinene, Citral, D-Limonene, and Cis-Beta Ocimene were picked up by the sensors. The detectable terpenes were used to explore how each of the terpenes spread in a room in an ideal world to see if they were identifiable. A simple mass based equation for VOC emissions from a concentration was used demonstrate physical world equations could not be used to classify the terpene in the room. Using the VOC data from the detectable terpenes and basil plants, we use machine learning to classify the chemical in the room. This paper provides the groundwork for how machine learning, VOC sensors, and plants can be combined to create new innovative sensors.

ACKNOWLEDGEMENTS

I would like to thank Viswajith Govinda Rajan for helping me learn about the sensors and how they are set up within a space. Viswajith provided helpful feedback which helped navigate the layout for this study.

I would like to thank Hamidreza for assisting me in gathering traces of the data for the dataset created in this study. He provided different resources which would be useful to understand plant communication and insight which improved the methodology of the study.

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Volatile organic compounds (VOCs) are key to plant survival. They indicate when plant is stressed due to abiotic or biotic factors. These factors include, but are not limited to, pathogen attacks [1], plant reproduction [2], or oxidative stress [3, 4]. Plants utilize VOCs to communicate and encrypt information to other plants or predators [5]. They do this by releasing terpenes, unsaturated hydrocarbons responsible for how plants smell. Plants can be genetically modified to release specific terpenes when they detect chosen pathogens. By modifying these plants and monitoring the VOCs released, they can act as sensors of the environment around them.

Air quality sensors determine the status of a location by measuring compounds not viewable to the naked eye, such as the amount of PM2.5, VOCs, and CO_2 in a room. Although VOC sensors are a useful resource, the best ones can often be expensive or require large setups. E-Noses are one of the most common expensive VOC sensors. They are an array of VOC sensors. Many studies have been done with highly sensitive equipment however, these are not readily available or scalable due to the high cost, need for skilled technicians, and time consuming nature [6]. As such, commercial sensors provide scalable and cheaper option for VOC detection. The AWAIR Omni sensor (shown in Figure 1.1) has the ability to determine the VOC particles in the room.

Using the ability to determine the VOC particle in the room, we can build a pipeline for using plants as sensors. A genetically modified plant would release a particular VOC when it detects a pathogen in the air. This VOC would be detected by the AWAIR sensor and have a curve associated with it. This curve

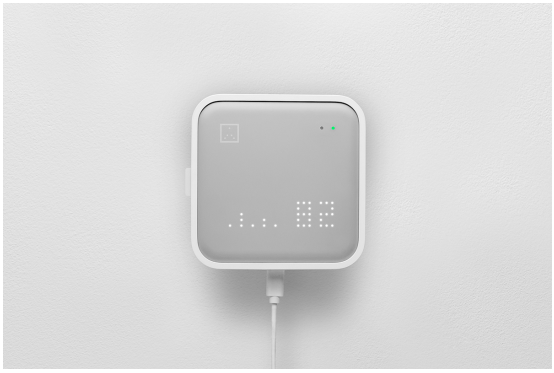


Figure 1.1: Image of Awair Sensor

would be analyzed to determine the terpene is in the room. This terpene would be associated with a pathogen, thus we know the pathogen is present in the room. This pipeline can be seen in Figure 1.2. This study lays the initial groundwork of using commercial VOC sensors in conjunction with plants to indirectly sense pathogens in the air. We aim to answer the questions of which chemicals are sensitive to the AWAIR, are physics-based models usable to determine the chemical in the room, and if we can use machine learning to identify the chemical in the room. The overall goal of this study is to use commercially available VOC sensors to gather data which we can use to discern if a plant has release a terpene or not.

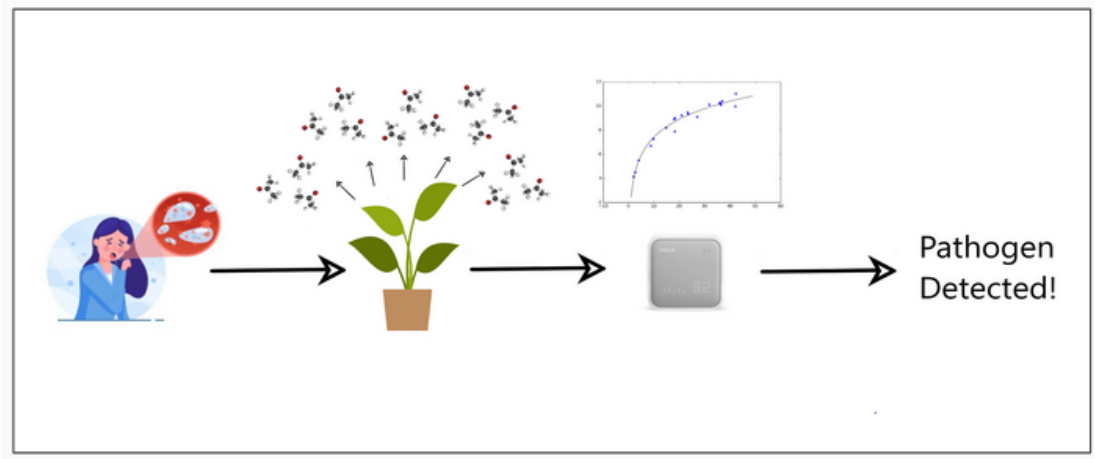


Figure 1.2: Pipeline of Using Commercial Sensors to Detect Pathogens

1.1 Background

1.1.1 VOCs in Plants – How Plants Utilize VOCs

Plants are constantly interacting with the world around them. A major way they do this, is by emitting volatile organic compounds (VOCs) called terpenes. Plants emit unique volatiles for different purposes [7]. One of these purposes is defense. Plants can emit VOCs in defense to a pathogen attack [1]. Each plant has it's own form of defense for different scenarios. One study showed that when tomatoes are attacked by whiteflies, they release volatiles that make neighboring tomatoes better hosts for new generations of whiteflies [8]. When a plant is under stress, more carbon is used to produce VOCs, which increases the amount in the atmosphere [9]. Each blend of volatiles encrypts a lot of information to predators and plants [5]. Plants may also emit VOCs to attract pollinators which is vital for plant reproduction [2]. Plants do not just use VOCs to interact with other organisms, they use it to alleviate stress caused by the environment [10]. Previous studies have shown that plants use isoprene and other volatiles to improve recovery from exposure to high temperatures [11, 12, 3]. Other studies have shown that plants which release isoprenes in response to oxidative stress have less cell damage, and higher photosynthetic rates than control plants without these compounds [3, 4]. All parts of the plant are able to release VOCs [9], however, different plant organs emit different groups of VOCs [13].

Different factors can influence the amount of volatiles emitted from a plant. One such factor is the nitrogen in the soil. A study with maize showed that nitrogen deficiency in the soil caused an increase in the emissions of the plant [14]. A lack of nitrogen does not always result in increased emissions. Another study showed that although some plants increase VOC emissions when there is more nitrogen in the soil, other plants emissions are not affected by the amount of nitrogen [15]. Competition can also impact the emissions of plants as intra-species competition can increase emissions while different species can lower emissions [16].

Other factors such as water, humidity, temperature, light have also impacted the emission rates of plants [17].

1.1.2 VOC Sensing – Covers E-Noses and Commercial Sensors

There are an abundance of different VOC sensors. Low cost sensors are useful as they provide a cheap option to determine the VOCs in the room. The main downside of these sensors is the lack of sensitivity [18]. On the other hand, high cost sensors are able to distinguish the differences between different volatile profiles [19]. One such high cost sensor is the electronic nose (e-nose), an array of various VOC sensors. e-noses have high accuracy due to their ability to grab a unique signal for each VOC [20]. Although useful, expensive sensors are not easily scalable for large spaces.

An important aspect of this study is how to measure VOCs released by plants. The two most common ways to analyze plant VOCs is through gas chromatography and mass spectrometry [13]. Research with plant volatiles can be extremely difficult due to environmental stressors and measure concentrations being low (1-100 ppb), the plant needing to be enclosed in a chamber, and VOCs being very reactive to their surroundings and technology [13]. As mentioned in the previous paragraph, e-noses are valuable resources in determining a VOC. One study has shown that e-noses are able to accurately determine different aromas from herbs and plants when combined with machine learning [21]. The traditional methods of detecting plant VOCs are expensive and bulky, meaning new methods for measuring VOCs need to be developed [22].

1.2 Related Work

1.2.1 Using Plants As Sensors

The idea of using plants as sensors is not a novel concept. Prior studies show plants being used to monitor detection of tobacco [23], indoor environment sta-

tus [24, 25], and air quality [26]. Although the idea of using plants is shared, the methodology for how to get the plant data varies between different studies. For example, one study uses wearable nanobiotechnology to monitor the stress of a plant [27]. Other studies have taken leaves from the plants and crushed them to see the concentration of chemicals in the leaf [23]. This thesis builds on using plants as sensors by using commercial sensors to interpret plant stress.

1.2.2 Identifying VOCs

Identifying VOCs based on sensor readings is an important aspect to this study. This can be done looking at features in the VOC curve. One prior study using e-noses found that they were accurately able to identify which VOC was being detected by using the normalized sensor response, reaction time, and half of the time the VOC takes to clear [28]. The VOC curves in this study were all found to be logarithmic curves. When analyzing VOC curves, another good feature to observe is the median [29]. Although identifying VOCs through their curves can be done, consistent dosages and chemicals do not always produce the same curve. Yeoman et al. found that consistent times and dosages do not always have the same peak concentration [30]. A common method for diagnosing a VOC is machine learning. One prior study uses machine learning to determine if something is bacteria or fungi based on its VOC signature [31].

Chapter 2

Measuring VOCs from Specific Terpenes

2.1 Litmus Test

2.1.1 Motivation

As plants can release a variety of different terpenes, it is important to understand which terpenes are detected by the AWAIR sensor. AWAIR does not have a list of all chemicals it can detect due to the various number of VOCs and inability for one sensor to detect all VOCs. As a result, the sensor aims to measure VOCs as a group. It looks for Alcohols, Ketones, Aromatic Hydrocarbons, Aliphatic Hydrocarbons, Aldehydes, Human Occupant VOCs, and others as shown in Figure 2.1.1. AWAIR does provide numerous examples of detectable chemicals in these groups on their website. Using a group of 16 terpenes, we show which plant chemicals are detectable by our sensor. The results of this study show the best terpenes that be implemented into genetically modified plants used for sensing.

AIR POLLUTANT CLASS	EXAMPLE COMPOUND	TYPICAL SOURCES
Alcohols	Ethanol	Paints, drinks
Ketones	Acetone	Cosmetics, solvents
Aromatic hydrocarbons	Toluene	Solvents, consumer products
Aliphatic hydrocarbons	Cyclohenaxe	Solvents, cleanign agents
Aldehydes	Formaldehyde	Building materials, home furnishings
Human occupants	Hydrogen	Human breath
Others	Carbon Monoxide	Incomplete combustion

Figure 2.1: Types of VOCs Detected By AWAIR

2.1.2 Methodology

This experiment was conducted in an office space of size $29.65m^3$. An AWAIR Omni sensor was set up on the desk and was left in the room for two days in the room to calibrate. The device requires 24 to 48 hours to calibrate (more information here). Once two days passed, testing could be conducted. The 16 terpenes used in this study and their chemical formula can be seen in Table 2.1. In this list, Cis-Beta Ocimene 70% does not list what else makes up the terpene other than cis-beta ocimene, therefore we assume the only compound is Cis-Beta Ocimene. For each individual terpene, 100 μL were applied to a piece of filter paper outside of the room. The pipette, filter paper, and example terpenes can be seen in Figure 2.2.

Figure 2.2: Image of the chemical vials, pipettes, and filter paper

Immediately after applying the terpene to the filter paper, the sample was placed directly next to sensor on the desk for 5 minutes. The VOC levels for the sensor were monitored during that period and the initial reading, max reading, and final reading were recorded for the 5 minute period. The delta between the initial and final reading ($\Delta_{final-initial}$) are used to determine the longevity of the terpene in the room. The delta between the initial and max reading ($\Delta_{max-initial}$) are used to determine the AWAIR's sensitivity to the terpene.

Table 2.1: Table of all 16 Chemicals - Description

Chemical	Formula
α-Bisabolol	$C_{15}H_{26}O$
α-Caryophyllene	$C_{15}H_{24}$
α-Phellandrene	$C_{10}H_{16}$
α-Pinene	$C_{10}H_{16}$
α-Terpineol	$C_{10}H_{18}O$
α-Terpinine	$C_{10}H_{16}$
β-Caryophyllene	$C_{15}H_{24}$
β-Pinene	$C_{10}H_{16}$
Cedrene	$C_{15}H_{24}$
Cis Beta-Ocimene 70%	$C_{10}H_{16}$
Citral	$C_{10}H_{16}O$
Citronellol	$C_{10}H_{20}O$
D-Limonene	$C_{10}H_{16}$
Delta-3-Carene	$C_{10}H_{16}$
Farnesene	$C_{15}H_{24}$
Geranoil, Natural	$C_{10}H_{18}O$
Linalool	$C_{10}H_{18}O$

2.1.3 Results

The results of the litmus test study can be seen in Table 2.2. All of the chemicals were picked up by the sensor at different magnitudes. Of the 16 terpenes, Cis-Beta Ocimene was the most sensitive with a $\Delta_{max-initial}$ of 38820 ppb. Following this were D-Limonene, Alpha-Terpinene, and Citral with changes over 1000 ppb. The terpene with the best longevity was also Cis-Beta Ocimene. Following this were also D-Limonene, Alpha-Terpinene, and Citral with changes over 1000 ppb. Overall, the terpenes with the best longevity and sensitivity were Cis-Beta Ocimene, D-Limonene, Citral, and Alpha-Terpinene. The bar chart for sensitivity can be seen in Figure 2.3 and the chart for longevity can be seen in Figure 2.4.

These four terpenes share the same chemical makeup other than Citral which has an extra oxygen atom and are all classified as monoterpenes. All terpenes were able to be sensed by the AWAIR, however, the sensitivity and longevity of each was different. Of note, we were able to sense the two compounds from the

Table 2.2: Table of all 16 Chemicals - Results

Chemical	$\Delta_{final-initial}$ (ppb)	$\Delta_{max-initial}$ (ppb)
α-Bisabolol	235	257
α-Caryophyllene	2	36
α-Phellandrene	530	857
α-Pinene	59	59
α-Terpinine	1266	1811
β-Caryophyllene	42	44
β-Pinene	15	29
Cedrene	1	16
Cis Beta-Ocimene 70%	38774	38774
Citral	1106	1106
Citronellol	112	119
D-Limonene	4356	4356
Delta-3-Carene	142	142
Farnesene	45	54
Geranoil, Natural	45	290
Linalool	23	27

AWAIR website, D-Limonene and α-pinene.

2.1.4 Discussion

As mentioned in the results, the AWAIR sensor was able to detect Cis-Beta Ocimene, D-Limonene, Alpha-Terpinene, and Citral the most and for the longest amount of time. Although they have the same chemical formualas (other than Citral), the structure of the molecules are not the same. Different structures have an impact on the normalized sensor response and reaction time of a sensor [28]. Furthermore, chemicals with different formulas have different sensor responses [32]. With regard to plants, basil plants are are large emitters of Cis-Beta Ocimene, making them a useful plant to genetically modfiy for the AWAIR Sensor. This chapter emphasizes the need to understand the terpene properties when determining if it is detectable by a specific sensor.

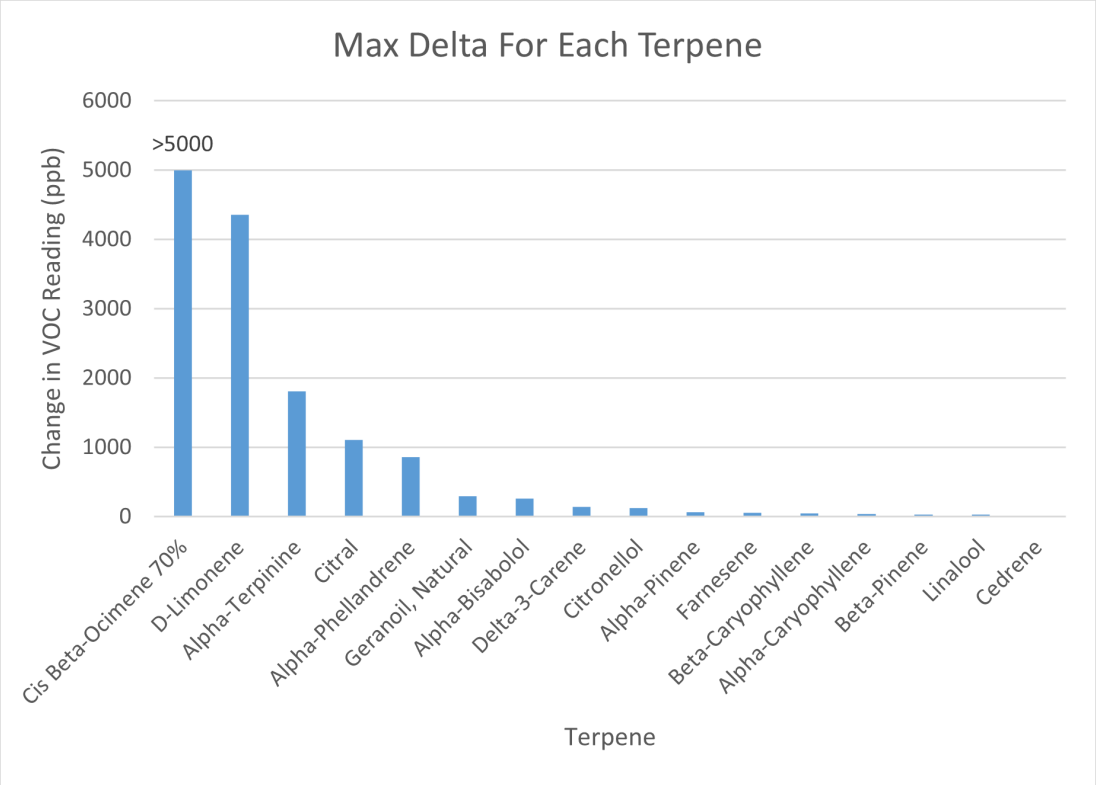


Figure 2.3: Image of max delta for each terpene

2.2 Emission/Machine Learning Data Collection

2.2.1 Motivation

In order to understand the different chemicals VOC readings, we gathered data for the best plant terpenes from Section 2.1. Prior studies have shown that machine learning can be used to distinguish between different VOC signatures [31] and expensive sensors can determine which chemical is in a room based on it's curve [28]. By using the VOC reading of the AWAIR Omni, we create a repository of VOC curves for different plant based terpenes, nothing, and basil plants to determine if they are distinguishable from each other. It will give us an idea of the sensitivity and accuracy of the AWAIR sensor.

2.2.2 Methodology

The experiment was set up in an enclosed office space of size $29.65m^3$, the same one used in Section 2.1. To monitor the the environment, 4 AWAIR Omni sensors

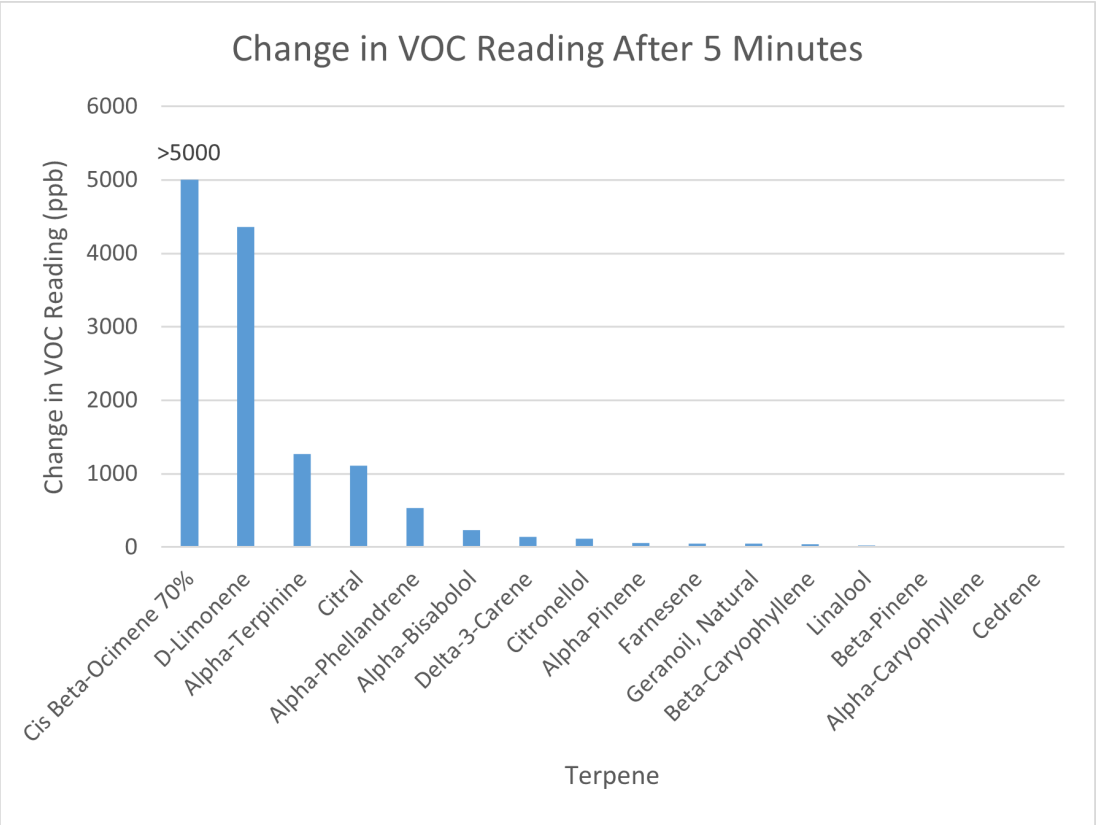


Figure 2.4: Image of final delta for each terpene

were set up on a desk. Four sensors were set up with two being 75 cms away from the terpene and two being 125 cm away from the terpene. A diagram of the setup is shown in Figure 2.5, while the actual room setup can be seen in Figure 2.6. Each sensor was labeled, with AWAIR Sensor 2 and 3 being the closest to the terpene and Sensor 1 and 4 being farther away from the terpene. The sensors were left in the room for 1-2 days prior to any testing to calibrate to the room. A binder clip held 30 cms above the desk by a string was used hold the terpene in the air.

For each test, a strict methodology was followed with all tests occurring when the HVAC system was on. Before any sample was placed in the room, each AWAIR sensor had to be at or below baseline (¡150 ppb). If the sensors do not read at baseline, the window and door were opened the clear the VOCs in the room and given 5 minutes to settle. To ensure validity of tests, testers should enter and exit the room slowly to make sure no VOCs are blown into the room

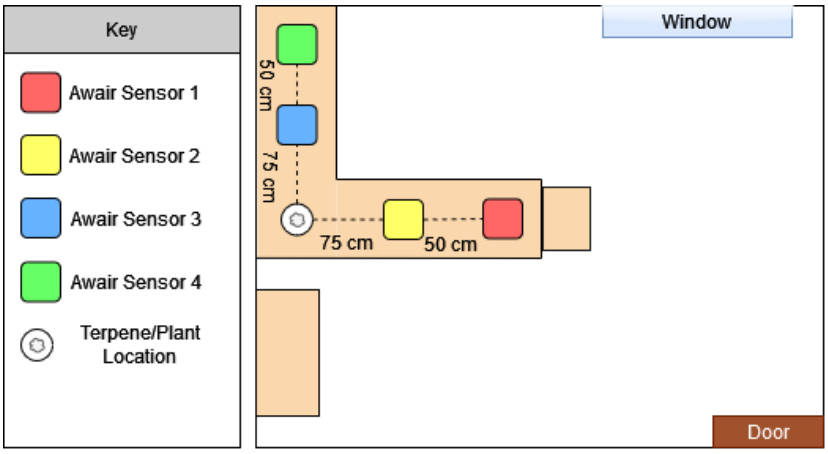


Figure 2.5: Image of Room (Diagram)

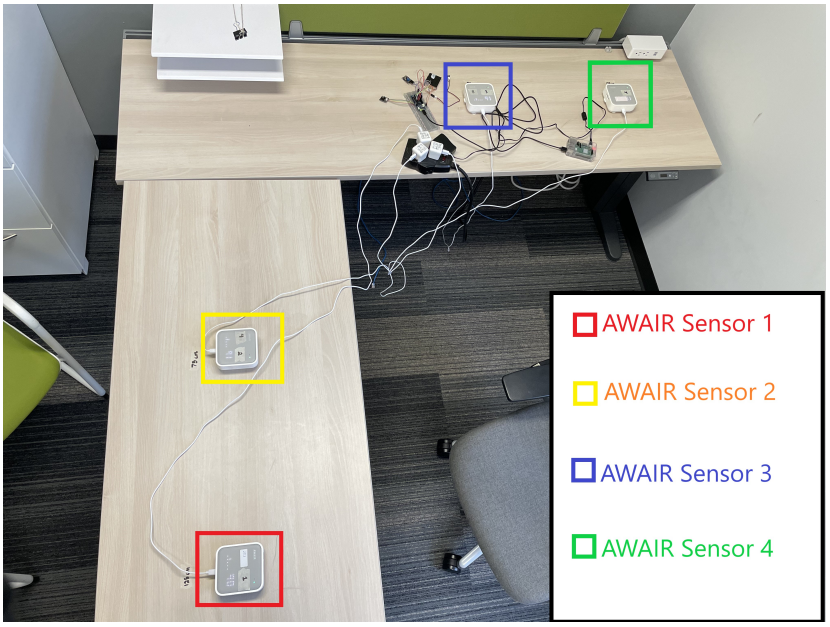


Figure 2.6: Image of room and setup

from the sliding door. Testers should also make sure to not wear any sort of cologne/perfume to ensure no sensors give inaccurate readings. Once each of the sensors measured at or below 150 ppb and 5 minutes have passed, the tester should enter the room and pour the desired dosage of the terpene onto the filter paper and hang it on the binder clip. Leave the sample in the room for 15 minutes. Once 15 minutes have elapsed, remove the used filter paper from the room and let the room sit for 5 minutes. After 5 minutes have elapsed, open the door and window so the room can clear. Repeat these steps for each individual test. The chemicals and dosages used for this study were Cis-Beta Ocimene (70%),

D-Limonene, Citral, and Alpha-Terpinene at measures of 200 μL or 100 μL as well as "Nothing" in the room to get the control.

If a plant was being tested, the plant would be in the room in the same location as the terpenes for at least 1-2 days so the sensor would be calibrated to the plant in the room. When running tests, the tester would still have to wait for the baseline (150 ppb) to be reached and wait 5 minutes for the room to settle. Once the 5 minutes have passed, enter the room and crush three leaves of the plant with their hand. This stresses the plant to make it release volatile compounds. After 15 minutes remove the crushed leafs from the plant and let the room settle for 5 minutes. Once the 5 minutes have passed, open the door and window so the VOCs can clear from the room. The only plant used for this study is basil. Basil was used due to it releasing Cis-Beta Ocimene, the best performer in Section 2.1.

2.2.3 Dataset Description

In total, we ran 81 tests which gave us 324 lines (One line for each AWAIR Sensor). Each line is considered a datapoint in our study. There were 52 data points for Citral, 92 data points for D-Limonene, 40 data points for Cis-Beta Ocimene 70%, 40 data points for Alpha-Terpinine, 28 data points for Basil Plant, and 72 points for Nothing. For Citral, 12 data points were gathered using 100 μL and 40 using 200 μL. For D-Limonene, 40 data points were gathered using 100 μL and 52 using 200 μL. For Cis-Beta Ocimene 70%, 12 data points were gathered using 100 μL and 28 using 200 μL. All Alpha-Terpinene data points were collected using 200 μL. A summary of all data can be seen in Table 2.3

2.2.4 Note on Airflow

All tests were conducted when the HVAC system was running. Although the airflow is relatively consistent given the HVAC system, it is important to note how air flow can impact the results. An example of a complete test for cis-beta

Table 2.3: Dataset Summary

Chemical	Total Traces	200μL Traces	100μL Traces
Alpha-Terpinene	40	40	0
Cis-Beta Ocimene	40	28	12
Citral	52	40	12
D-Limonene	92	52	40
Nothing	72	N/A	N/A
Basil	28	N/A	N/A

ocimene at 200μL can be seen in Figure 2.7. In this test there is a small spike at 15:55 when the terpene is removed from the room. This shows that when entering and exiting the room with the terpene, the airflow and source of the VOC change which can lead to spikes in the data. This further illustrates the need to understand how airflow and VOC location impacts sensor readings.

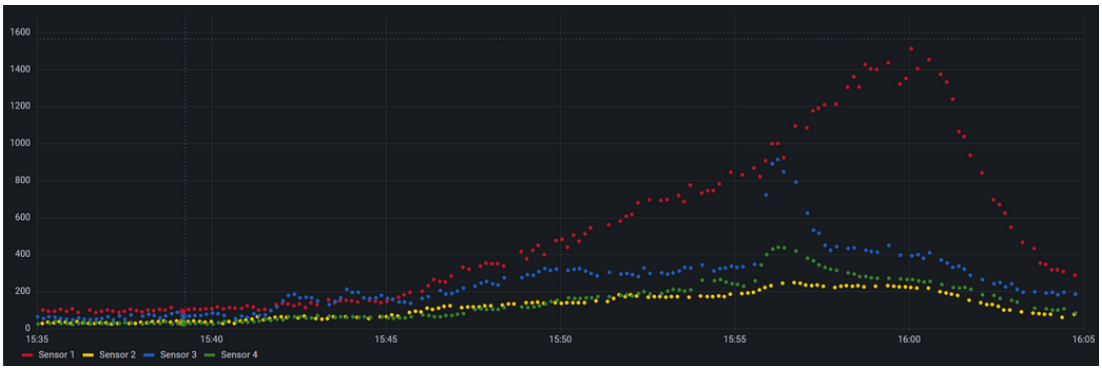


Figure 2.7: Example of Full Trace

2.2.5 Results

To get an idea of what the data looked like for each terpene, the average trace and traces for each terpene-dosage pair and sensor were gathered for the 15 minutes the terpene was in the room. This includes the average trace and actual traces for nothing (Figure 2.8), alpha-terpinene 200μL (Figure 2.9), cis-beta ocimene 200μL (Figure 2.11), cis-beta ocimene 100μL (Figure 2.10), citral 200μL (Figure 2.13), citral 100μL (Figure 2.12), d-limonene 200μL (Figure 2.15), d-limonene 100μL (Figure 2.14), and the basil stress tests (Figure 2.16). In these figures, the

white line with the black outline is the average while the other colored lines are the individual traces. Note that the VOC axes do not have the same values and the curves are not similar across the different sensors. For each chemical dosage pair, the curves are relatively different for each sensor with sensor 3 and sensor 1 having the highest readings. When nothing is in the room, there is an increase in the VOC reading of the room. This is due to the room going back up to baseline after opening the window and door.

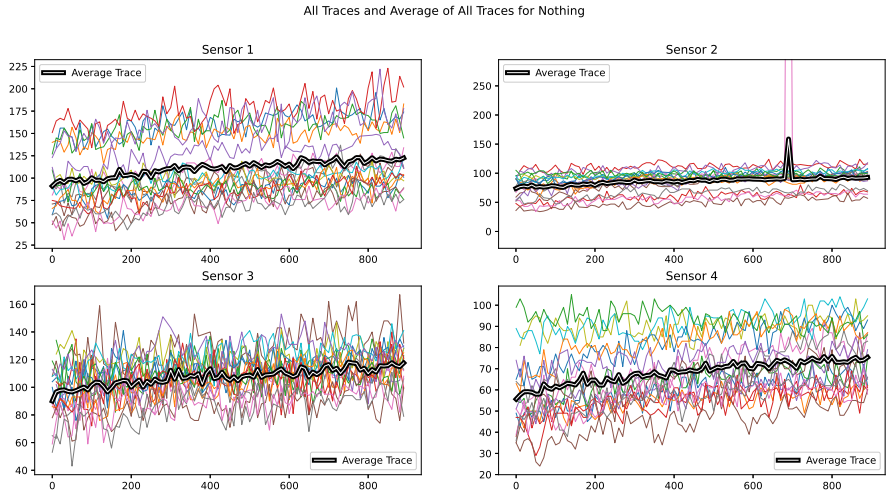


Figure 2.8: Average Curve For Nothing At Each Sensor

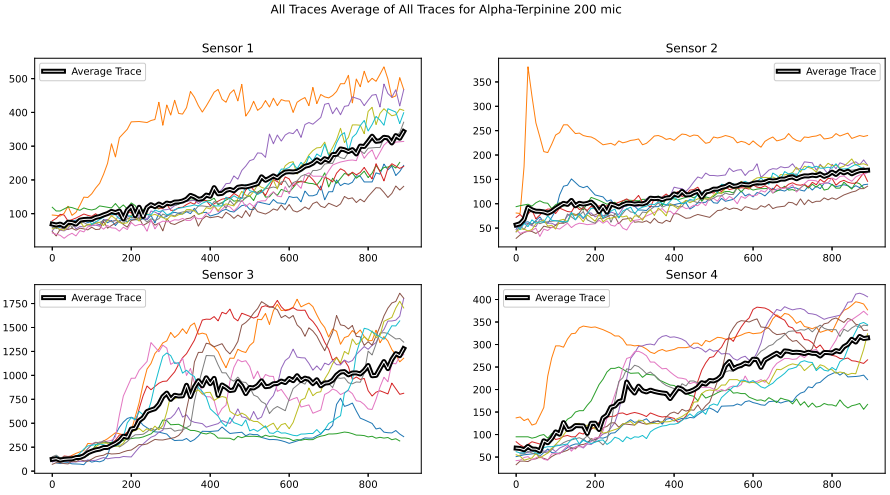


Figure 2.9: All Traces and Average Trace For Alpha-Terpinene 200μL At Each Sensor

The traces for each plant terpene at each sensor were also observed. All

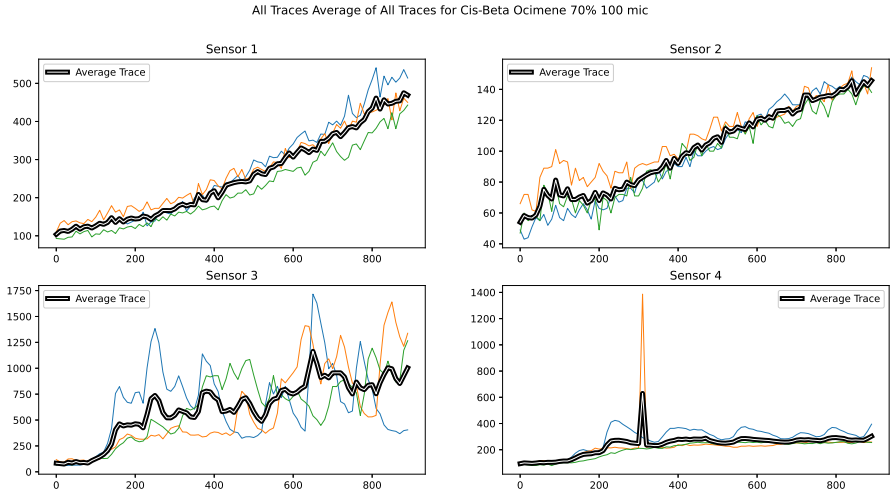


Figure 2.10: All Traces and Average Trace For Cis-Beta Ocimene 100μL At Each Sensor

terpene traces for sensor 1 can be seen in Figure 2.17, sensor 2 can be seen in Figure 2.18, sensor 3 can be seen in Figure 2.19, and sensor 4 can be seen in Figure 2.20. The main takeaway from each of these figures is that the sensor response depends on the terpene, not just distance. The different terpene traces (different colored lines) are unique for each terpene.

2.3 Takeaways

The first study observes the sensitivity of a commercially available sensor (the AWAIR Omni) to pure plant terpene compounds. This study showed that commercial sensors can be used to detect all of the plant VOCs, with the best ones being citrus based terpenes. Although the molecular formula of these terpenes are the same, the longevity and sensor sensitivity are unique to each chemical. Using these data we gathered sensor data to the strongest chemicals, D-Limonene, Citral, Cis-Beta Ocimene, and Alpha-Terpinene.

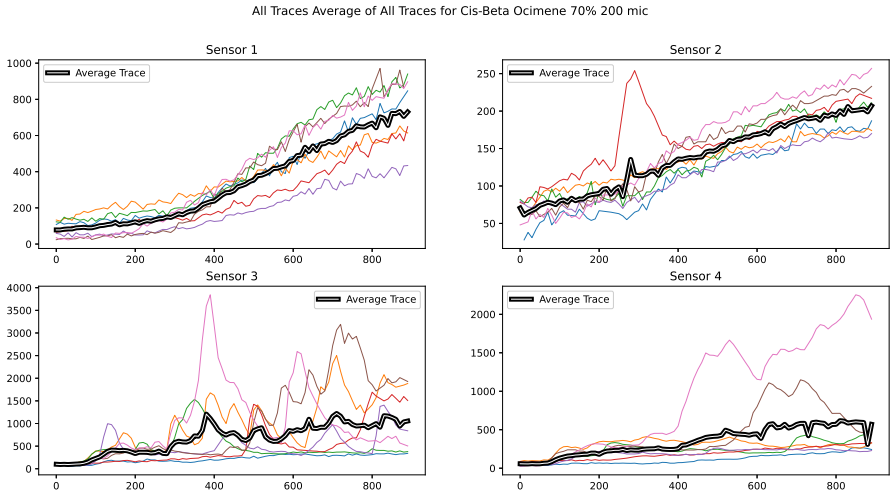


Figure 2.11: All Traces and Average Trace For Cis-Beta Ocimene 200μL At Each Sensor

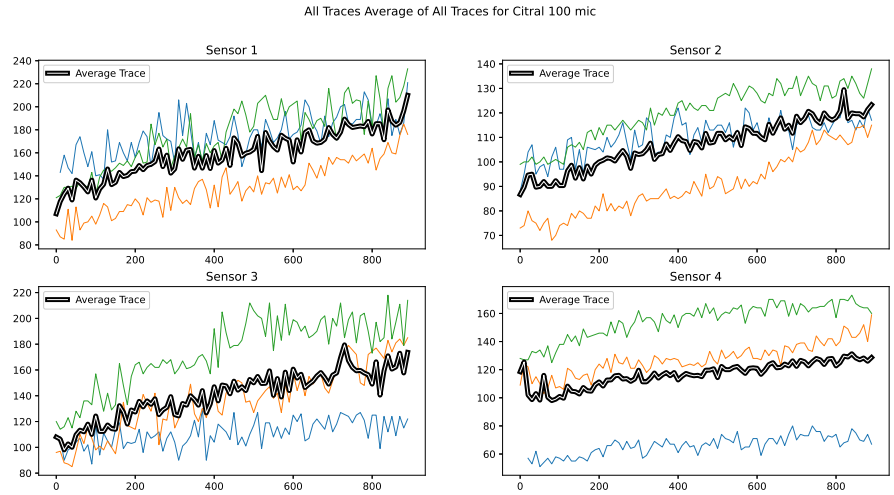


Figure 2.12: All Traces and Average Trace For Citral 100μL At Each Sensor

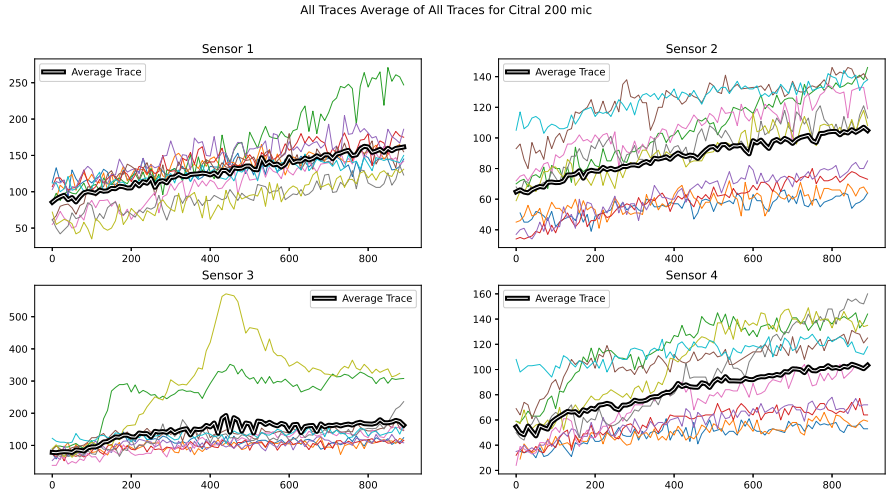


Figure 2.13: All Traces and Average Trace For Citral 200μL At Each Sensor

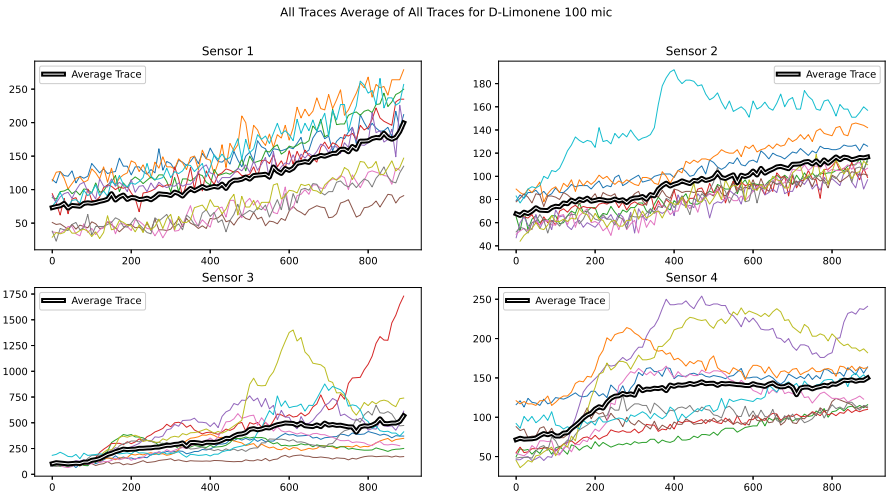


Figure 2.14: All Traces and Average Trace For D-Limonene 100μL At Each Sensor

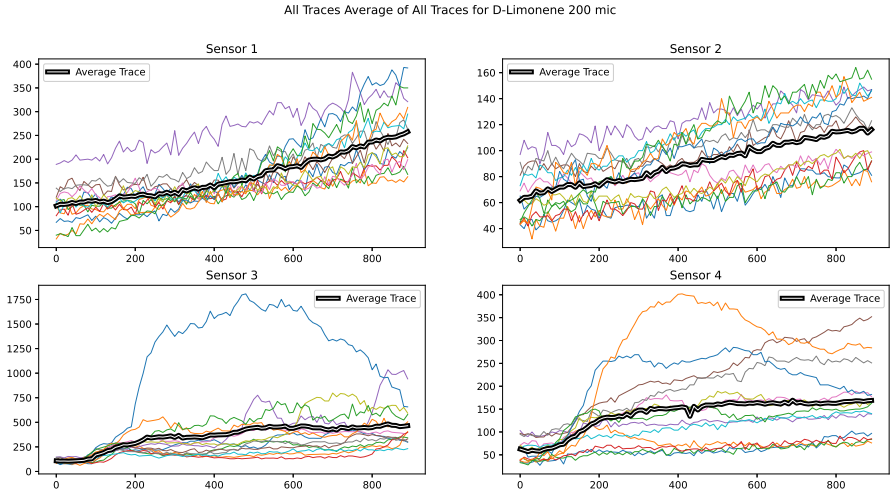


Figure 2.15: All Traces and Average Trace For D-Limonene 200μL At Each Sensor

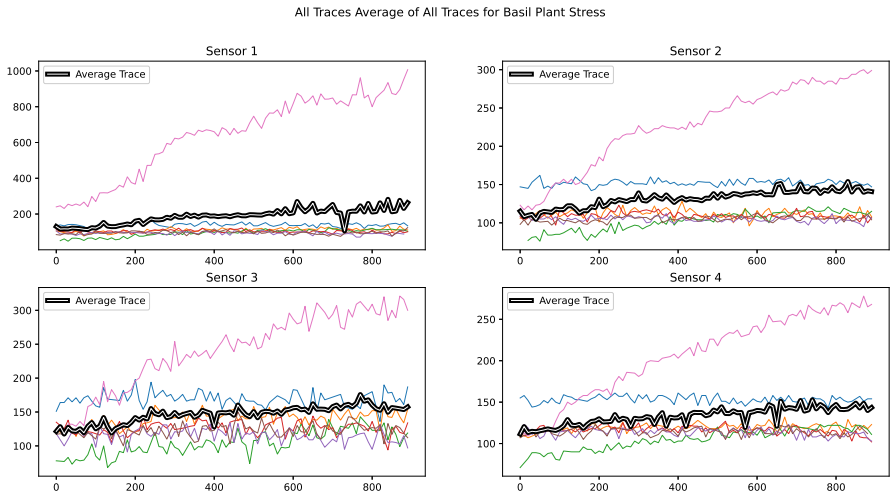


Figure 2.16: All Traces and Average Trace For Basil Stress At Each Sensor

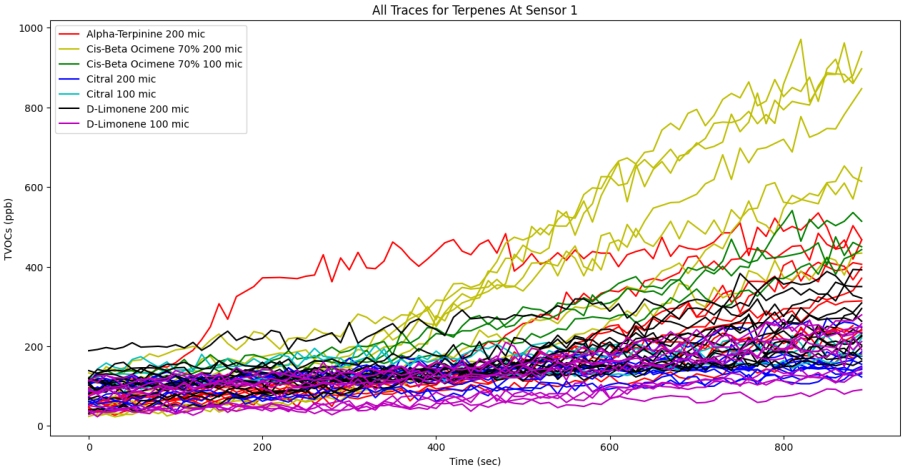


Figure 2.17: All Terpene Traces for Plant Terpenes At Sensor 1

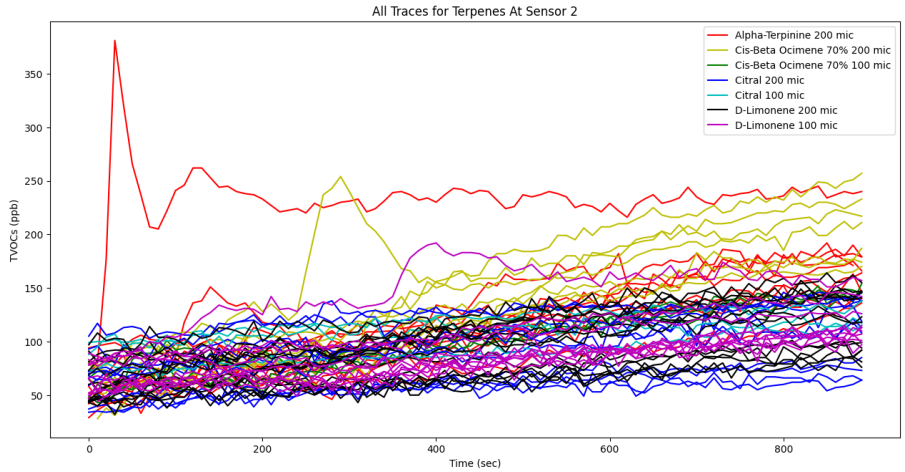


Figure 2.18: All Terpene Traces for Plant Terpenes At Sensor 2

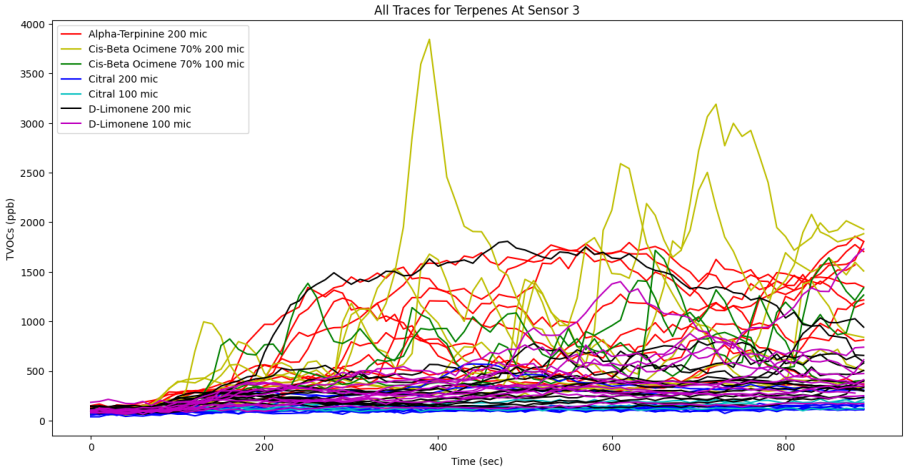


Figure 2.19: All Terpene Traces for Plant Terpenes At Sensor 3

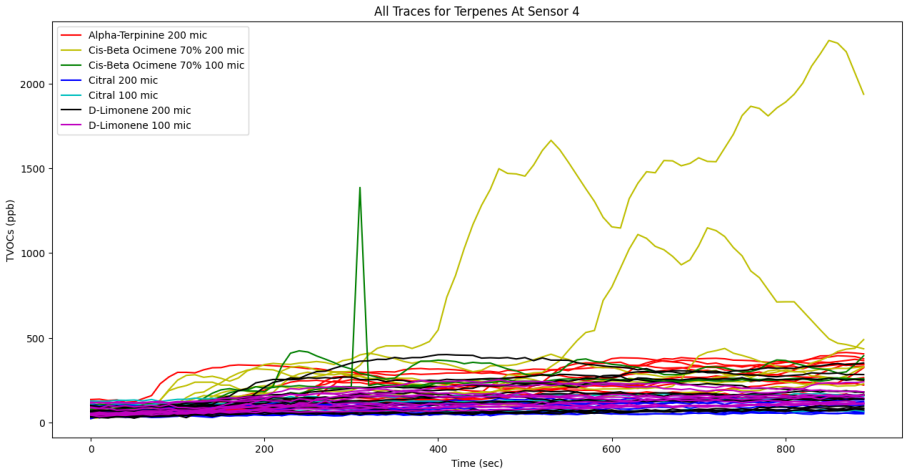


Figure 2.20: All Terpene Traces for Plant Terpenes At Sensor 4

Chapter 3

Modeling and Environment

3.1 Motivation

Different volatile compounds spread in rooms at a different emission rates. To model this difference, this study aims to observe and understand the physics of how VOCs are spread in a space. One prior study has used a simple mass based equation to determine the VOC concentration in the room at a specific time [33]. This equation allows us to demonstrate if the data from all four AWAIR sensors follows this "ideal world" equation. By combining the VOC readings and equation, we could identify the chemical be emission rate in an ideal environment.

3.2 Determining the Emission Rate In A Room

As mentioned in the motivation section, we use a simple mass based equation to determine the emission rate of the substance in the room. The equation used in the study done by Hori et al. [33] can be seen below and assumes the room is a perfect mixing.

$$C = (C_{in} - \frac{F}{Q})e^{(-\frac{Q}{V}t)} + \frac{F}{Q} \tag{3.1}$$

This equation solves for C, the concentration in the room ($\mu g/m^3$). C_{in} is the initial concentration ($\mu g/m^3$), F is the emission rate ($\mu g/h$), Q is the ventilation air flow rate (m^3/h), V is the volume of the room (m^3) and t is time (h). Using this equation, we can solve for the emission rate of the VOC.

$$
\begin{aligned}
C &= (C_{in} - \frac{F}{Q})e^{(-\frac{Q}{V}t)} + \frac{F}{Q} \\
C &= C_{in}e^{(-\frac{Q}{V}t)} - \frac{F}{Q}e^{(-\frac{Q}{V}t)} + \frac{F}{Q} \\
C - C_{in}e^{(-\frac{Q}{V}t)} &= -\frac{F}{Q}e^{(-\frac{Q}{V}t)} + \frac{F}{Q} \\
Q(C - C_{in}e^{(-\frac{Q}{V}t)}) &= -Fe^{(-\frac{Q}{V}t)} + F \\
Q(C - C_{in}e^{(-\frac{Q}{V}t)}) &= F(1 - e^{(-\frac{Q}{V}t)}) \\
\frac{Q(C - C_{in}e^{(-\frac{Q}{V}t)})}{1 - e^{(-\frac{Q}{V}t)}} &= F \\
F &= Q\frac{(C - C_{in}e^{(-\frac{Q}{V}t)})}{1 - e^{(-\frac{Q}{V}t)}}
\end{aligned}
$$

The AWAIR sensor gathers VOC data with the unit particles per billion (ppb), however, the equation takes the units of $\mu g/m^3$. To convert this we make use of the following equation [34].

$$\mu g/m^3 = (ppb) * (12.187) * (M)/(273.15 +^\circ C) \tag{3.2}$$

In this equation M is the molecular weight of the gaseous pollutant being measured, $^\circ C$ is the temperature of the room in degrees Celsius. For our study we consider the M to be the molecular weight of air ($M = 28.96$) as the sensors measure more than just the chemical being used and does not distinguish how many ppb of the chemical there are.

3.3 Methodology

Data were collected from the building and AWAIR sensors. Specifically, the ventilation air flow rate was gathered from building data, and the temperature, concentration, and initial concentration were taken from the AWAIR Omni. All AWAIR data were used from the dataset described in Section 2.2.3. The 15

minutes after the plant leaves were crushed or terpenes placed in the room was the data used. The room had a volume of $29.65m^3$.

The emission rate was calculated for each chemical and dosage pair at each sensor. As the sensors are prone to error, the median VOC reading of the first 5 values (first 50 seconds the sample is in the room) in the line were for the initial concentration in the room. This prevented any outliers from impacting the emission rate. The emission rate was calculated for the last 5 values (last 50 seconds the sample is in the room) and the median of these values was used as the emission rate. After getting the emission rate for each test at each sensor, the emission rates of the same chemical dosage pair were averaged for each sensor.

3.4 Results

The average emission rates for each dosage and chemical pair at each sensor can be seen in Table 3.1 and Figure 3.1. The emission rate of all sensors greatly varies for Alpha-Terpinene ($200\mu L$), Cis-Beta Ocimene ($200\mu L$ and $100\mu L$) and D-Limonene ($200\mu L$ and $100\mu L$). There was a slight variation in the emission rates for all sensors for Citral ($200\mu L$ and $100\mu L$), Basil and Nothing. Using the emission rates, we map out what the predicted curve would be of a Cis-Beta Ocimene (100 μL) trace (Figure 3.5), D-Limonene (100 μL) trace (Figure 3.3), Citral (200 μL) trace (Figure 3.4, and 3.2. The data lines up for D-Limonene, Cis-Beta Ocimene, and Nothing, however, the Citral actual trace does not. This is due to the actual data being an outlier. All chemicals and the basil plant had higher emission rates than nothing in the room.

3.5 Discussion

Using an idealistic equation to determine the emission rate, we assumed the rates would be constant across the different sensors for each chemical and the basil plant. This was not the case, however, as there were different emission rates for many of the chemicals. Although controlled, our test data was done in an office

Table 3.1: Average Emission Rate For All Terpenes At Each Sensor

Chemical and Dosage	Sensor 1	Sensor 2	Sensor 3	Sensor 4
Alpha-Terpinene (200μL)	90,972	44,243	339,482	87,085
Citral (200μL)	43,850	28,494	49,050	28,604
Citral (100μL)	40,610	24,317	33,186	22,701
Cis-Beta Ocimene (200μL)	228,250	60,732	333,420	181,767
Cis-Beta Ocimene (100μL)	91,528	26,163	188,678	52,009
D-Limonene (200μL)	69,458	31,613	132,430	46,614
D-Limonene (100μL)	46,122	28,125	139,622	36,019
Nothing	28,069	21,408	27,130	17,346
Basil	47,412	30,809	34,041	31,572

space where limitations of sensor distance, sensor sensitivity, and the inability to determine the specific compounds in the air can impact the results. Due to the difference in the rates, we believe applying an "ideal world" equation cannot be used to identify the chemical in the room.

3.6 Takeaways

The second study uses a simple mass based equation to identify the chemical in the room. This study showed that physics equations which assume a perfect mixing, ventilation, and pressure can not be used to determine which VOC is in the room. We show that the same chemicals do not have the same emission rate at different sensors. As emission rate should be constant among all sensors, we claim that better equations need to be created to predict the chemical in the room.

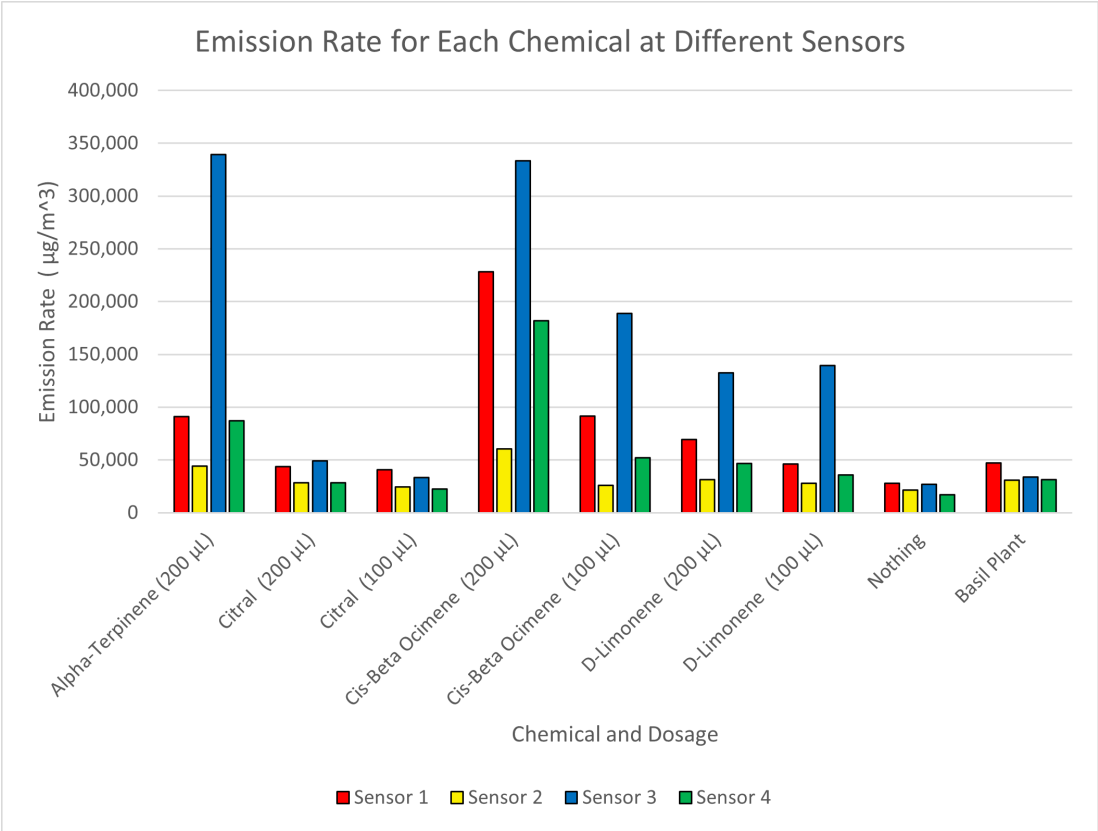


Figure 3.1: Graph showing difference in Emission Equation As Bar Chart

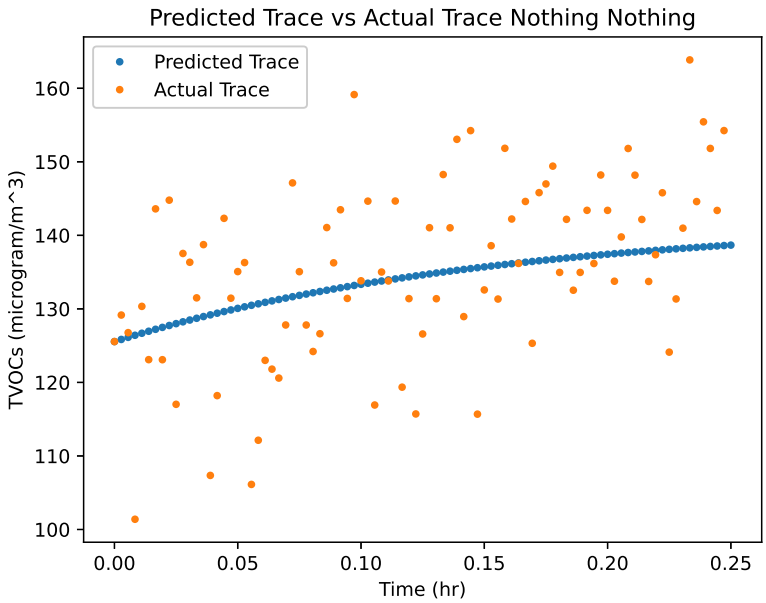


Figure 3.2: Graph of Predicted Trace vs Actual Trace for nothing in the room

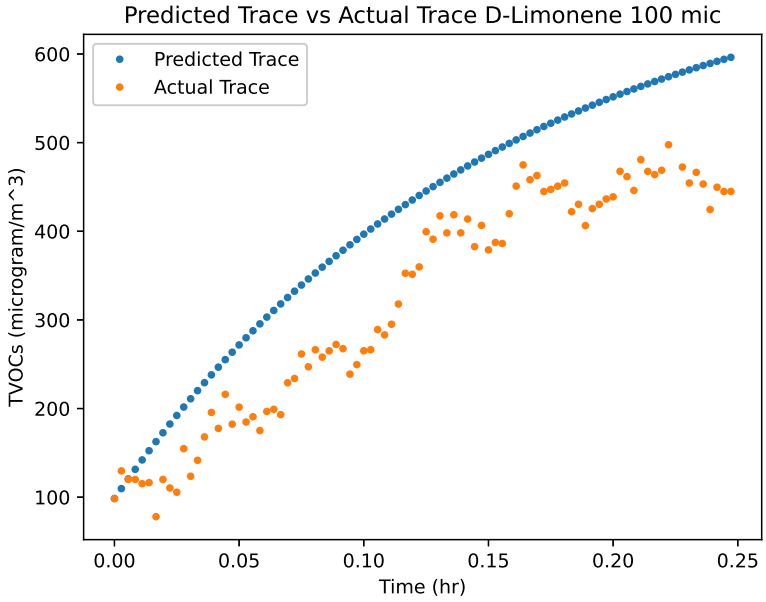


Figure 3.3: Graph of Predicted Trace vs Actual Trace for D-Limonene 100 μL

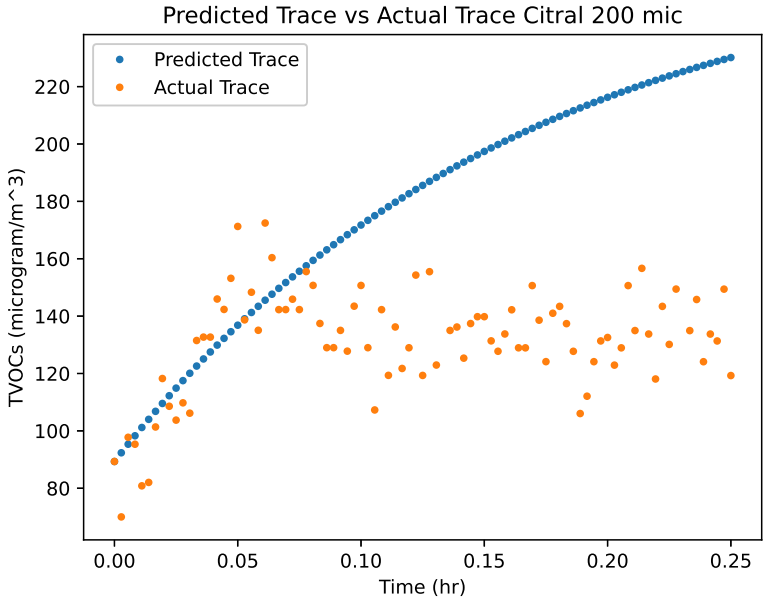


Figure 3.4: Graph of Predicted Trace vs Actual Trace for Citral 200 μL

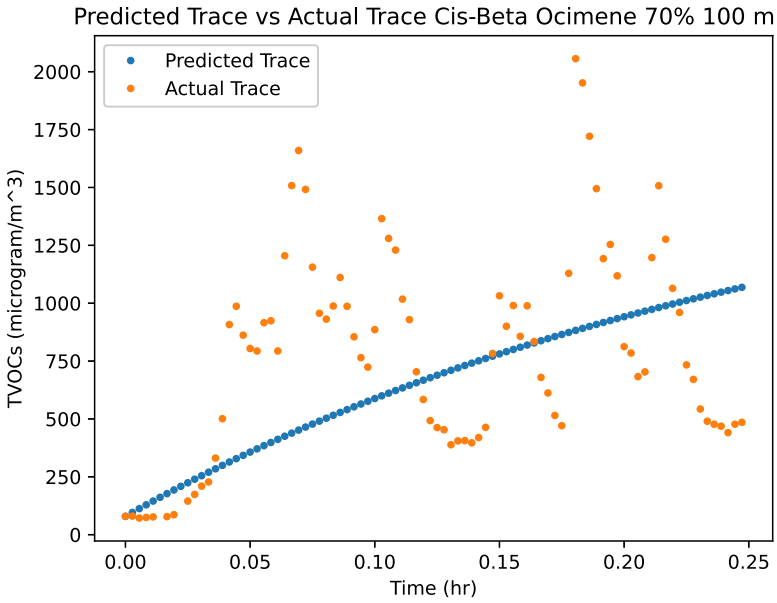


Figure 3.5: Graph of Predicted Trace vs Actual Trace for Cis-Beta Ocimene 100 μL

Chapter 4

Machine Learning and Classification

4.1 Motivation

As determining the chemical with an emission equation was not possible using the AWAIR sensor, we propose using machine learning. Prior studies have shown success in using machine learning to identify which VOC is in the room [31] [21], however, these studies used e-nose sensors. Using commercial sensors and machine learning allows us to interpret which terpene is released so we can create affordable and scalable plant sensors that do not require large setups.

4.2 Methodology

4.2.1 Data Selection and Split

All AWAIR data were used from the dataset described in Section 2.2.3. The dataset has 4 different unique chemicals, nothing, and basil plants. To test the data, we split the data into various groups for the machine learning model to classify. First, we observe the collective of all chemicals vs nothing to see if machine learning can be used to determine if a chemical is in the room. Next, we observe the individual chemicals vs nothing to see if the individual chemicals are distinguishable from nothing. Then, we look at the difference between 100 μL and 200 μL dosages of D-Limonene. Following individual chemicals, we observe if the machine learning models can distinguish between each of the chemicals without nothing. Finally, observe if machine learning can be used to determine the plant stress tests from nothing. For each of the different tests, 20% of the data was used in the testing set while the other 80% were used as the training

set. These data are not enough to create a fully fledged classification system, however, it can serve as a proof of concept for future work.

4.2.2 Model Selection

As the goal of the machine learning model is classification, we used Random Forest, SVM, and XGBoost. These are all good models for classification [35]. We chose to use all three of these to determine if different classifiers worked better than others. Sklearn's Random Forest and SVM were used [1] and xgboost's XGBClassifer was used [2].

4.2.3 Feature Selection

To extract features from the timeseries data, the library *tsfresh* was used. *tsfresh* is a python library used to extract a large number of features from time series data. The list of all features can be found on their website. As this library extracts numerous features, only the relevant features were selected using a function in tsfresh. Even with this functionality, there can be up to 300-400 relevant features. We select the best 15 features from these 300-400 by doing further feature selection to prevent overfitting. The best features between Random Forest, SVM, and XGBoost were analyzed as well.

4.3 Results

4.3.1 Chemicals vs Nothing

The first tests conducted were the chemicals vs nothing tests. Results for the Random Forest, SVM and XGBoost classifiers can be seen in Table 4.1 and Figure 4.1, Table 4.2 and Figure 4.2, and Table 4.3 and Figure 4.3 respectively. All classifiers had an accuracy of 93% with 43 samples in the test set being chemicals and 17 samples of the test set being nothing.

[1] https://scikit-learn.org/stable/
[2] https://xgboost.readthedocs.io/en/stable/install.html

Table 4.1: Random Forest - Chemicals vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Chemical	1.00	0.91	0.95	43
Nothing	0.81	1.00	0.89	17
Accuracy			0.93	60
Macro Avg	0.90	0.95	0.92	60
Weighted Avg	0.95	0.93	0.94	60

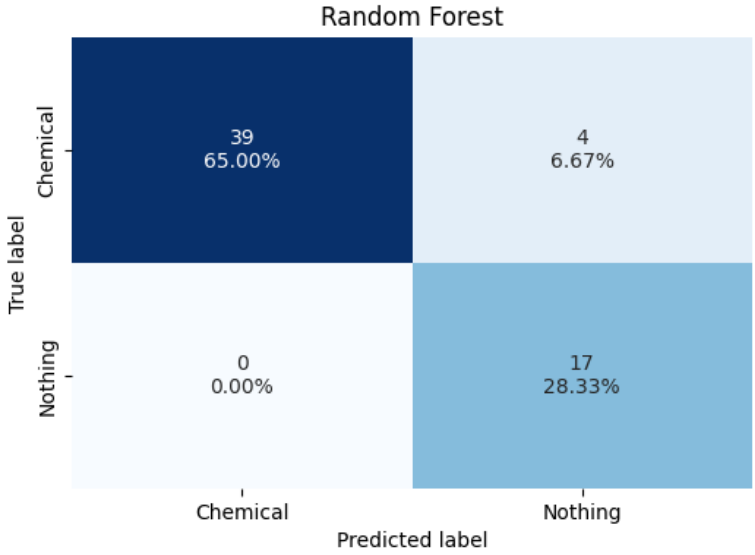


Figure 4.1: Confusion Matrix Heatmap - Random Forest Chemicals vs Nothing

Table 4.2: SVM - Chemicals vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Chemical	0.95	0.95	0.95	43
Nothing	0.88	0.88	0.88	17
Accuracy			0.93	60
Macro Avg	0.92	0.92	0.92	60
Weighted Avg	0.93	0.93	0.93	60

Table 4.3: XGBoost - Chemicals vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Chemical	0.95	0.95	0.95	43
Nothing	0.88	0.88	0.88	17
Accuracy			0.93	60
Macro Avg	0.92	0.92	0.92	60
Weighted Avg	0.93	0.93	0.93	60

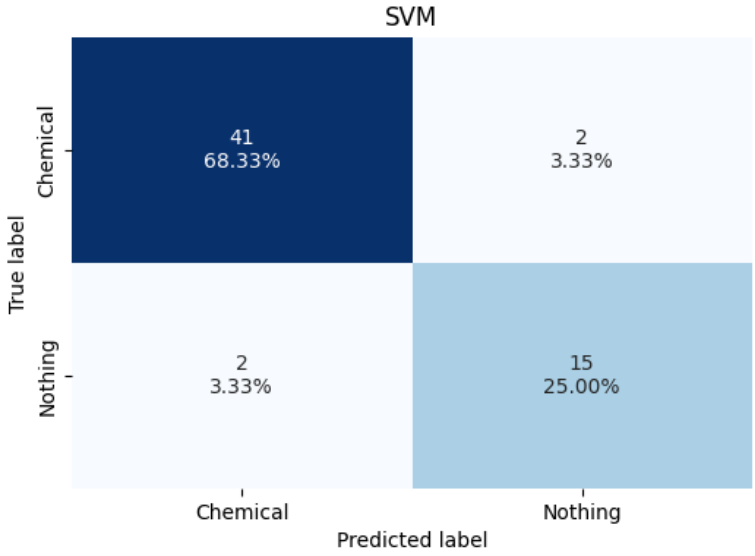


Figure 4.2: Confusion Matrix Heatmap - SVM Chemicals vs Nothing

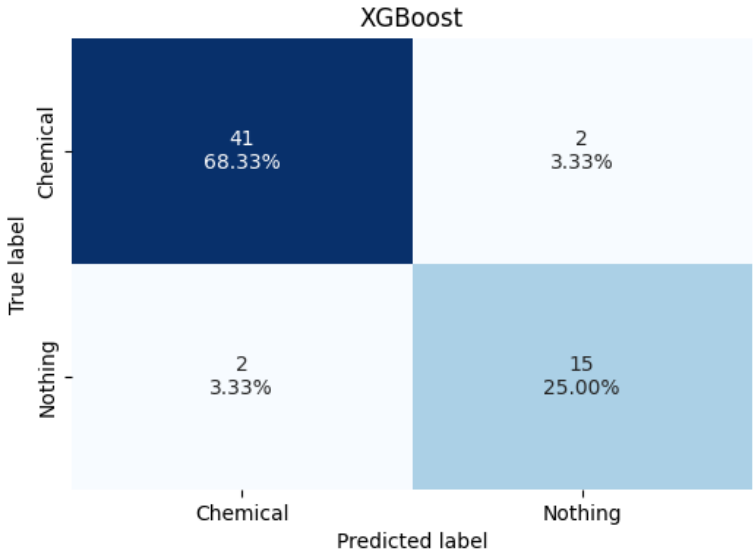


Figure 4.3: Confusion Matrix Heatmap - XGBoost Chemicals vs Nothing

4.3.2 Individual Chemicals vs Nothing

Cis-Beta Ocimene

The results for Cis-Beta Ocimene vs Nothing and confusion matricies for the random forest, SVM, and XGBoost can be seen in Table 4.4 and Figure 4.4, Table 4.5 and Figure 4.5, and Table 4.6 and Figure 4.6. Given a test set of 9 Cis-Beta Ocimene lines and 14 nothing lines, each model had an accuracy of 100%. The models were able to identify the difference between Cis-Beta Ocimene and nothing in the room.

Table 4.4: Random Forest - Cis-Beta Ocimene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

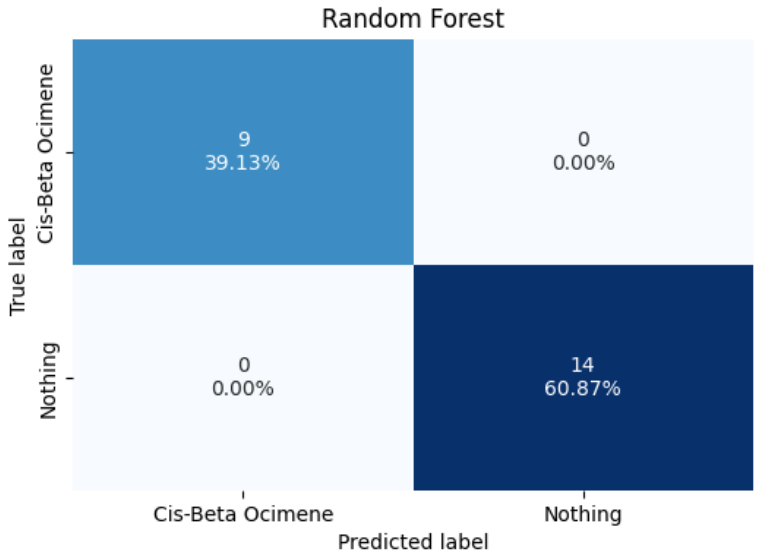


Figure 4.4: Confusion Matrix Heatmap - Random Forest Cis-Beta Ocimene vs Nothing

.

Table 4.5: SVM - Cis-Beta Ocimene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

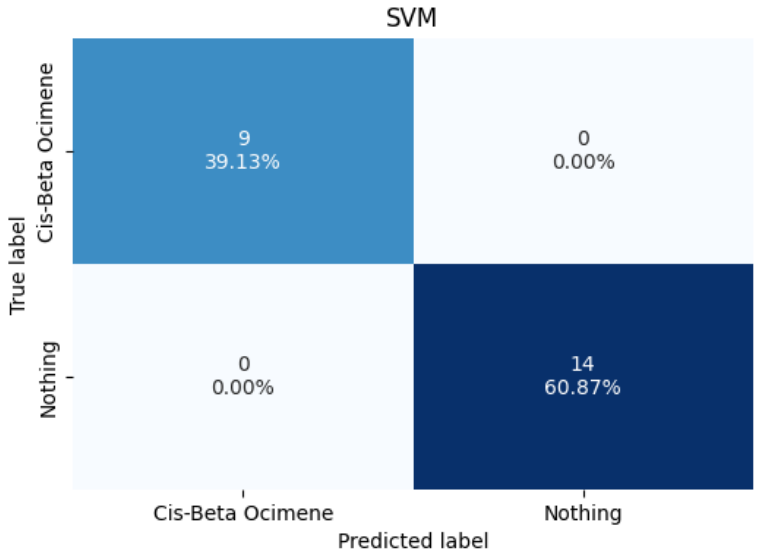


Figure 4.5: Confusion Matrix Heatmap - SVM Cis-Beta Ocimene vs Nothing

Table 4.6: XGBoost - Cis-Beta Ocimene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

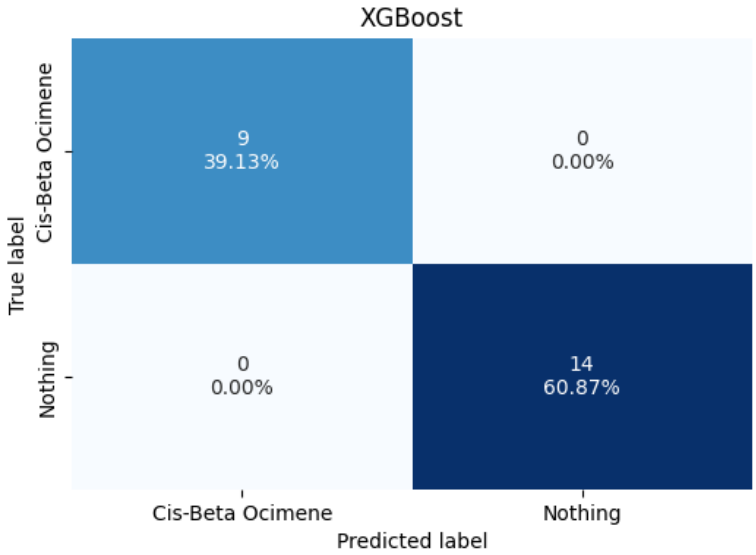


Figure 4.6: Confusion Matrix Heatmap - XGBoost Cis-Beta Ocimene vs Nothing

D-Limonene

The Random Forest, SVM, and XGBoost results and matrices for D-Limonene and nothing tests can be seen in Table 4.7 and Figure 4.7, Table 4.8 and Figure 4.8, and Table 4.9 and Figure 4.9 respectively. Utilizing the D-Limonene tests and nothing tests the machine learning models were able to accurately predict if D-Limonene was in the room. The Random Forest and XGBoost classifiers had the best accuracy of 97%. SVM did not perform as well as the other two models, however it had a high accuracy of 94%.

Table 4.7: Random Forest - D-Limonene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
D-Limonene	0.94	1.00	0.97	16
Nothing	1.00	0.94	0.97	17
Accuracy			0.97	33
Macro Avg	0.97	0.97	0.97	33
Weighted Avg	0.97	0.97	0.97	33

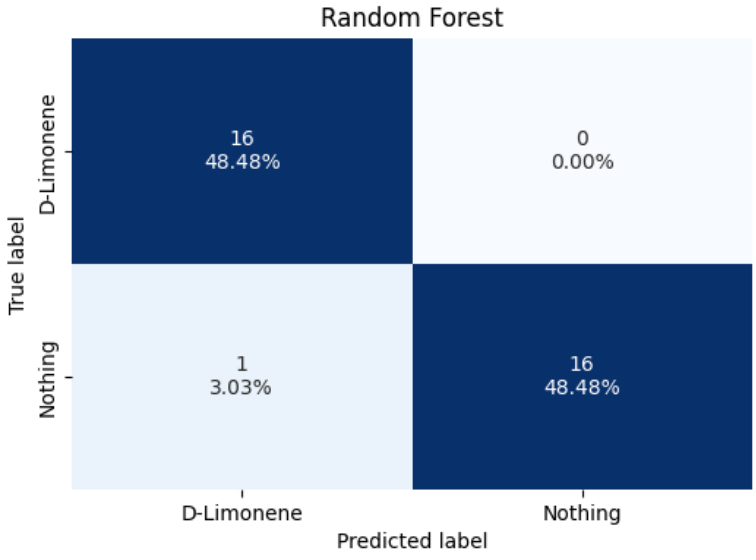


Figure 4.7: Confusion Matrix Heatmap - Random Forest D-Limonene vs Nothing

Table 4.8: SVM - D-Limonene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
D-Limonene	0.94	0.94	0.94	16
Nothing	0.94	0.94	0.94	17
Accuracy			0.94	33
Macro Avg	0.94	0.94	0.94	33
Weighted Avg	0.94	0.94	0.94	33

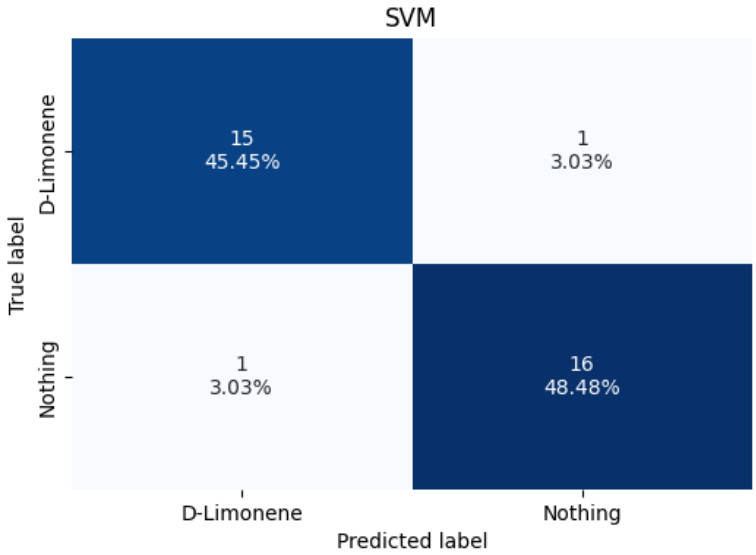


Figure 4.8: Confusion Matrix Heatmap - SVM D-Limonene vs Nothing

Table 4.9: XGBoost - D-Limonene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
D-Limonene	0.94	1.00	0.97	16
Nothing	1.00	0.94	0.97	17
Accuracy			0.97	33
Macro Avg	0.97	0.97	0.97	33
Weighted Avg	0.97	0.97	0.97	33

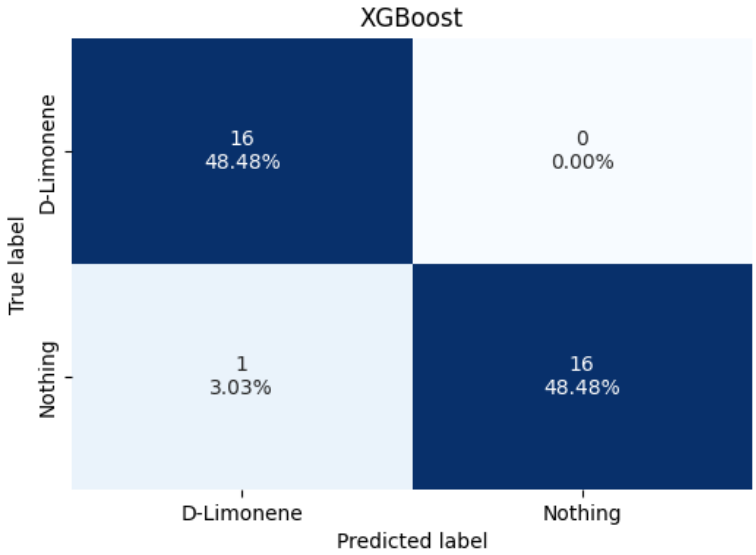


Figure 4.9: Confusion Matrix Heatmap - XGBoost D-Limonene vs Nothing

Citral

The Random Forest, SVM, and XGBoost results and matrices for citral and nothing tests can be seen in Table 4.10 and Figure 4.10, Table 4.11 and Figure 4.11, and Table 4.12 and Figure 4.12 respectively. The SVM had the highest accuracy of 96% while the Random Forest and XGBoost models had a worse accuracy of 80%.

Table 4.10: Random Forest - Citral vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Citral	0.67	1.00	0.80	10
Nothing	1.00	0.67	0.80	15
Accuracy			0.80	25
Macro Avg	0.83	0.83	0.80	25
Weighted Avg	0.87	0.80	0.80	25

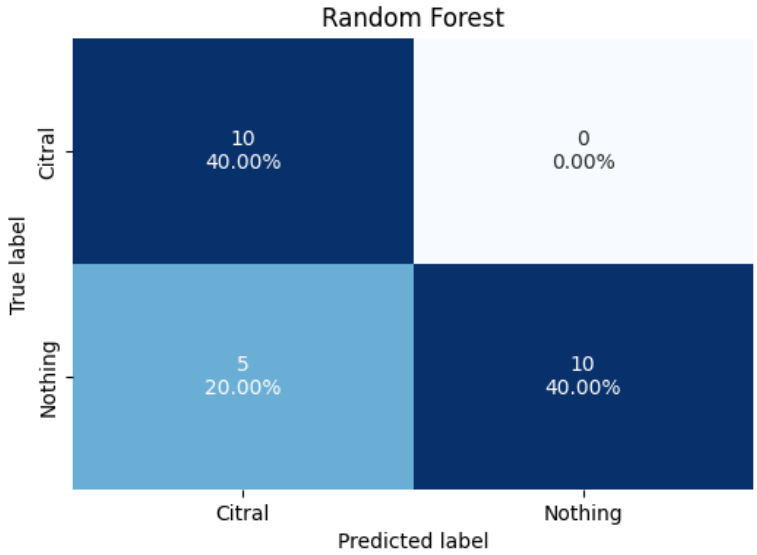


Figure 4.10: Confusion Matrix Heatmap - Random Forest Citral vs Nothing

Table 4.11: SVM - Citral vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Citral	0.91	1.00	0.95	10
Nothing	1.00	0.93	0.97	15
Accuracy			0.96	25
Macro Avg	0.95	0.97	0.96	25
Weighted Avg	0.96	0.96	0.96	25

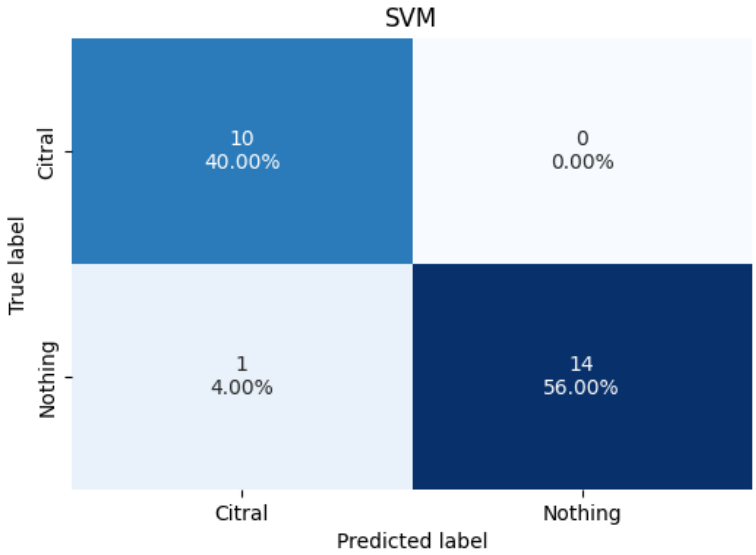


Figure 4.11: Confusion Matrix Heatmap - SVM Citral vs Nothing

Table 4.12: XGBoost - Citral vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Citral	0.67	1.00	0.80	10
Nothing	1.00	0.67	0.80	15
Accuracy			0.80	25
Macro Avg	0.83	0.83	0.80	25
Weighted Avg	0.87	0.80	0.80	25

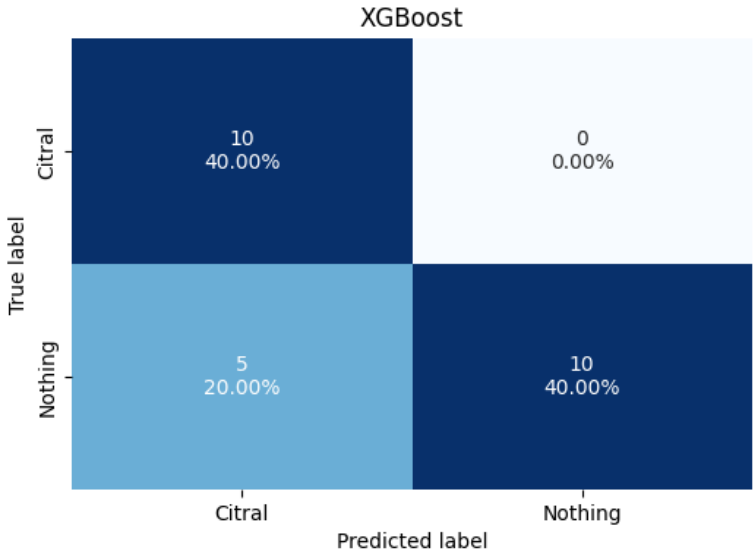


Figure 4.12: Confusion Matrix Heatmap - XGBoost Citral vs Nothing

Alpha-Terpinene

The Random Forest, SVM, and XGBoost results and matrices for alpha-terpinene and nothing tests can be seen in Table 4.13 and Figure 4.13, Table 4.14 and Figure 4.14, and Table 4.15 and Figure 4.15 respectively. Each model had an accuracy of 100% and they were able to identify the difference between alpha-terpinene and nothing in the room.

Table 4.13: Random Forest - Alpha-Terpinene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

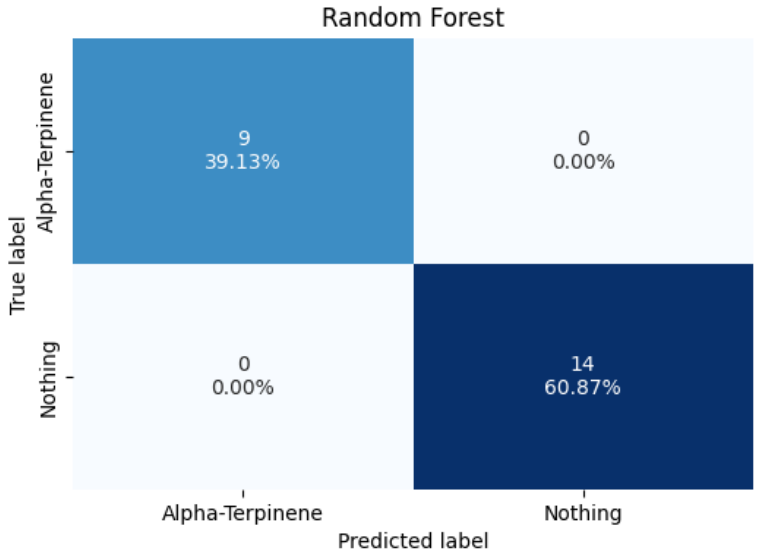


Figure 4.13: Confusion Matrix Heatmap - Random Forest Alpha-Terpinene vs Nothing

Table 4.14: SVM - Alpha-Terpinene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

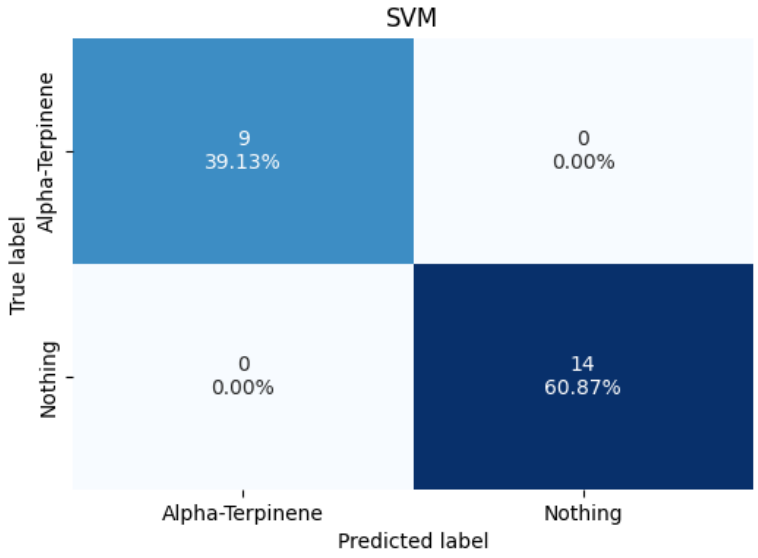


Figure 4.14: Confusion Matrix Heatmap - SVM Alpha-Terpinene vs Nothing

Table 4.15: XGBoost - Alpha-Terpinene vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	1.00	1.00	1.00	9
Nothing	1.00	1.00	1.00	14
Accuracy			1.00	23
Macro Avg	1.00	1.00	1.00	23
Weighted Avg	1.00	1.00	1.00	23

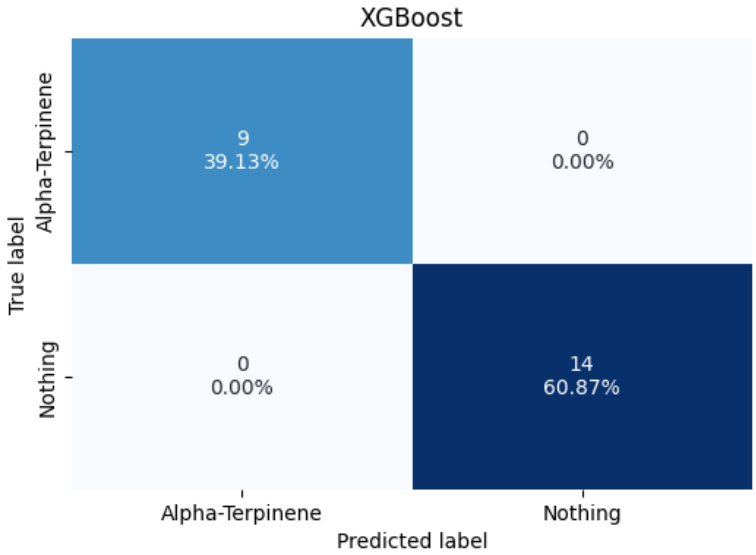


Figure 4.15: Confusion Matrix Heatmap - XGBoost Alpha-Terpinene vs Nothing

4.3.3 D-Limonene Dosages

The different dosages of D-Limonene were used to see if machine learning could tell the difference between $100\mu L$ and $200\mu L$. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.16 and Figure 4.16, Table 4.17 and Figure 4.17, and Table 4.18 and Figure 4.18 respectively. The accuracy was poor for all models with the worst accuracy being the SVM with 37%. Random Forest and XGBoost had slightly higher accuracy with 42%. The models were not accurately able to determine the difference between $100\mu L$ and $200\mu L$.

Table 4.16: Random Forest - D-Lim 200 vs D-Lim 100

Chemical	Precision	Recall	F1-Score	Support
D-Limonene $200\mu L$	0.33	0.38	0.35	8
D-Limonene $100\mu L$	0.50	0.45	0.48	11
Accuracy			0.42	19
Macro Avg	0.42	0.41	0.41	19
Weighted Avg	0.43	0.42	0.42	19

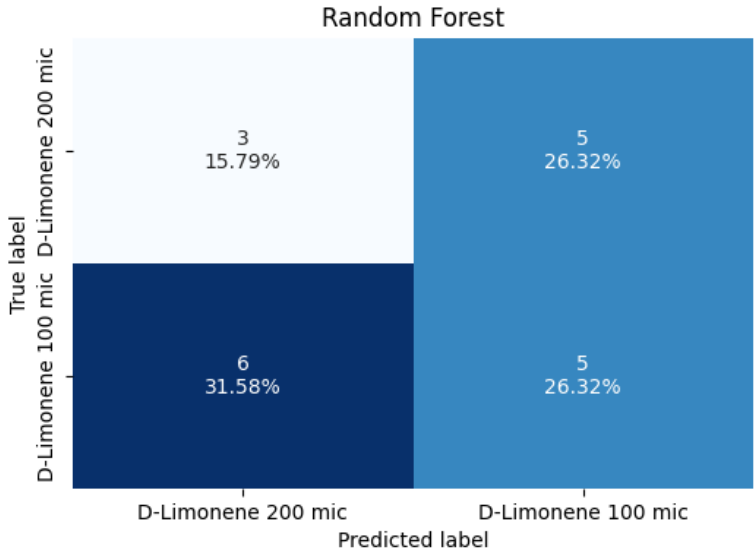


Figure 4.16: Confusion Matrix Heatmap - Random Forest D-Lim 200 vs D-Lim 100

Table 4.17: SVM - D-Lim 200 vs D-Lim 100

Chemical	Precision	Recall	F1-Score	Support
D-Limonene 200μL	0.25	0.25	0.25	8
D-Limonene 100μL	0.45	0.45	0.45	11
Accuracy			0.37	19
Macro Avg	0.35	0.35	0.35	19
Weighted Avg	0.37	0.37	0.37	19

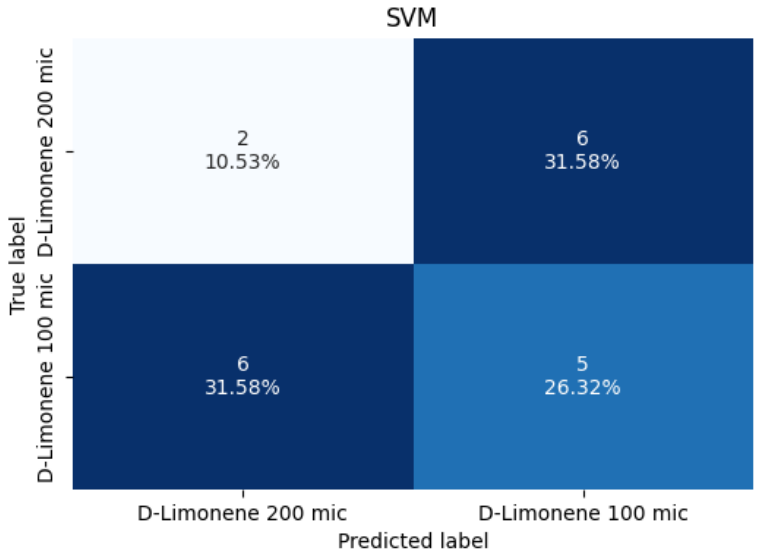


Figure 4.17: Confusion Matrix Heatmap - D-Lim 200 vs D-Lim 100

Table 4.18: XGBoost - D-Lim 200 vs D-Lim 100

Chemical	Precision	Recall	F1-Score	Support
D-Limonene 200μL	0.33	0.38	0.35	8
D-Limonene 100μL	0.50	0.45	0.48	11
Accuracy			0.42	19
Macro Avg	0.42	0.41	0.41	19
Weighted Avg	0.43	0.42	0.42	19

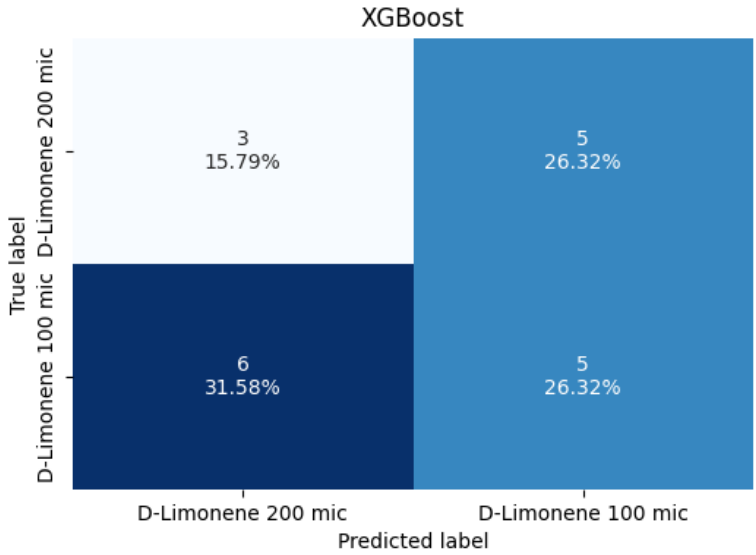


Figure 4.18: Confusion Matrix Heatmap - XGBoost D-Lim 200 vs D-Lim 100

4.3.4 Chemical Comparison

All Chemicals

All chemicals were used to see if the chemicals were distinguishable from each other. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.19 and Figure 4.19, Table 4.20 and Figure 4.20, and Table 4.21 and Figure 4.21 respectively. SVM was the most accurate model with an accuracy of 62% while Random Forest and XGBoost had a slightly lower accuracy of 60%. For each of the models, D-Limonene and Cis-Beta Ocimene had the highest accuracy.

Table 4.19: Random Forest - All Chemicals

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.57	0.40	0.47	10
Cis-Beta Ocimene	0.54	0.78	0.64	9
Citral	0.67	0.44	0.53	9
D-Limonene	0.63	0.71	0.67	17
Accuracy			0.60	45
Macro Avg	0.60	0.58	0.58	45
Weighted Avg	0.61	0.60	0.59	45

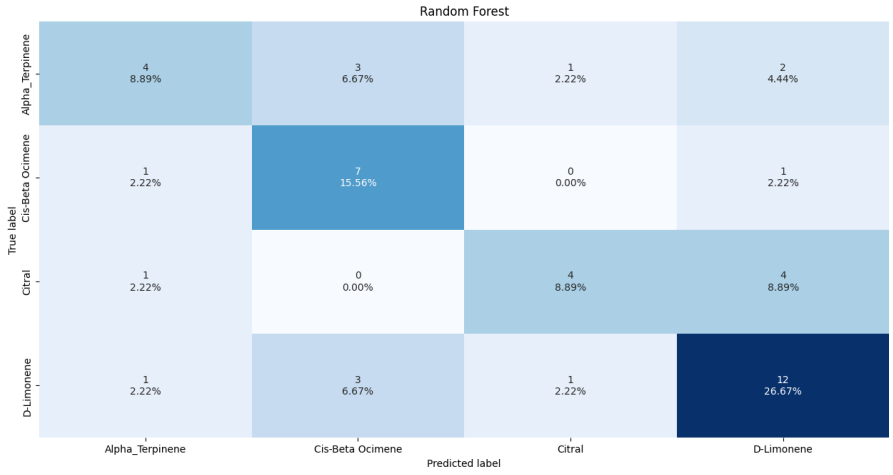


Figure 4.19: Confusion Matrix Heatmap - Random Forest All Chemicals

Table 4.20: SVM - All Chemicals

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.62	0.50	0.56	10
Cis-Beta Ocimene	0.58	0.78	0.67	9
Citral	0.57	0.44	0.50	9
D-Limonene	0.67	0.71	0.69	17
Accuracy			0.62	45
Macro Avg	0.61	0.61	0.60	45
Weighted Avg	0.62	0.62	0.62	45

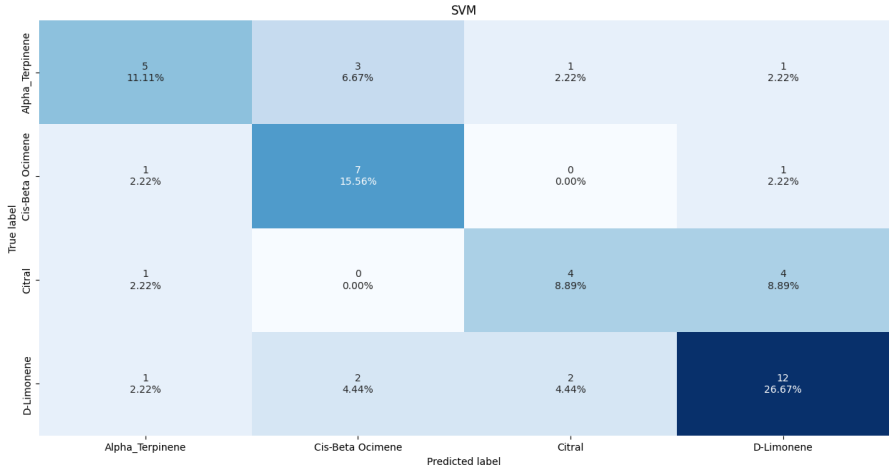


Figure 4.20: Confusion Matrix Heatmap - SVM All Chemicals

Table 4.21: XGBoost - All Chemicals

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.57	0.40	0.47	10
Cis-Beta Ocimene	0.53	0.89	0.67	9
Citral	0.56	0.56	0.56	9
D-Limonene	0.71	0.59	0.65	17
Accuracy			0.60	45
Macro Avg	0.59	0.61	0.58	45
Weighted Avg	0.61	0.60	0.59	45

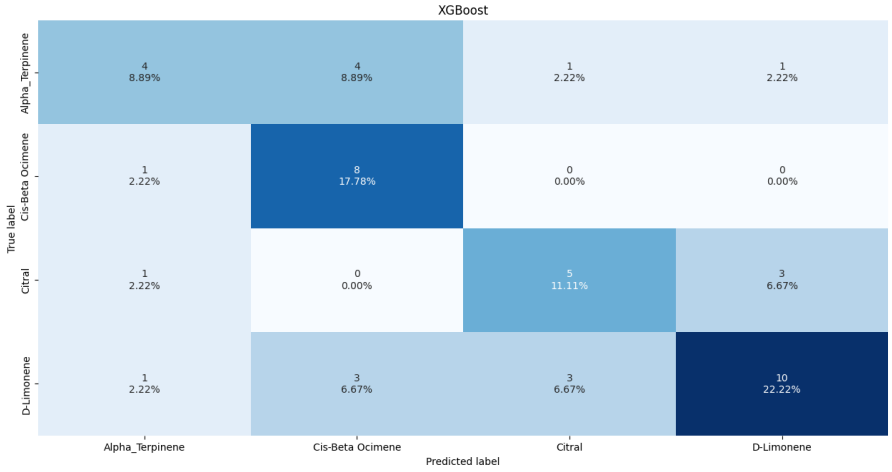


Figure 4.21: Confusion Matrix Heatmap - XGBoost All Chemicals

D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

As citral and alpha-terpinene each had a low accuracy, we decided to remove each one of them individually from the tests. For this test we observe D-Limonene, citral, and alpha-terpinene. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.22 and Figure 4.22, Table 4.23 and Figure 4.23, and Table 4.24 and Figure 4.24 respectively. Decreasing the amount of chemicals by removing Citral increased the overall accuracy for all models. The Random Forest had the largest increase and was 69% accurate as compared to 60% with all chemicals. Meanwhile, the SVM model was 66% accurate as compared to 62% with all chemicals and XGBoost model was 63% accurate as compared to 60% with all chemicals.

Table 4.22: Random Forest - D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.67	0.40	0.50	10
Cis-Beta Ocimene	0.71	0.62	0.67	8
D-Limonene	0.68	0.88	0.77	17
Accuracy			0.69	35
Macro Avg	0.69	0.64	0.65	35
Weighted Avg	0.68	0.69	0.67	35

Table 4.23: SVM - D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.46	0.60	0.52	10
Cis-Beta Ocimene	1.00	0.38	0.55	8
D-Limonene	0.74	0.82	0.77	17
Accuracy			0.66	35
Macro Avg	0.73	0.60	0.61	35
Weighted Avg	0.72	0.66	0.65	35

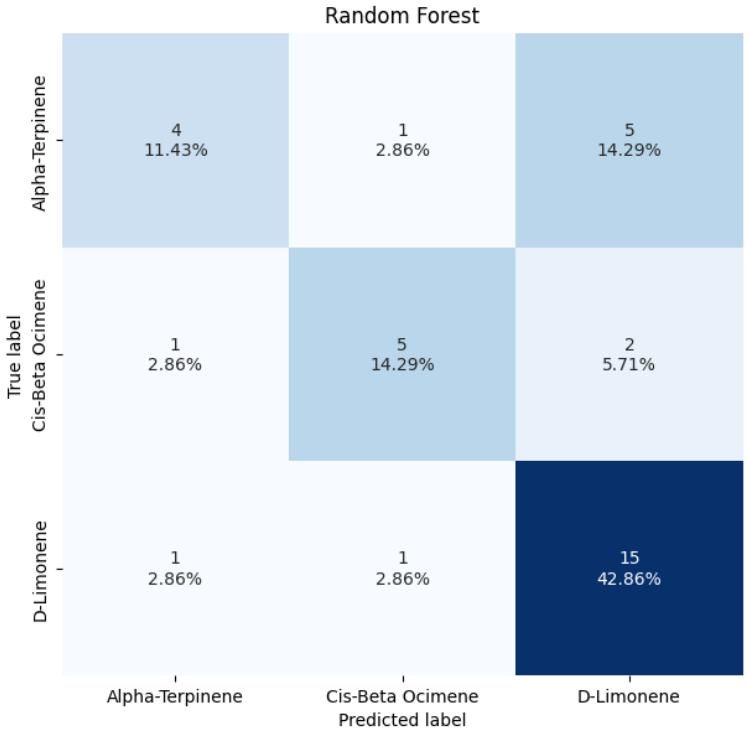


Figure 4.22: Confusion Matrix Heatmap - Random Forest D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

Table 4.24: XGBoost - D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

Chemical	Precision	Recall	F1-Score	Support
Alpha-Terpinene	0.71	0.50	0.59	10
Cis-Beta Ocimene	0.40	0.50	0.44	8
D-Limonene	0.72	0.76	0.74	17
Accuracy			0.63	35
Macro Avg	0.61	0.59	0.59	35
Weighted Avg	0.65	0.63	0.63	35

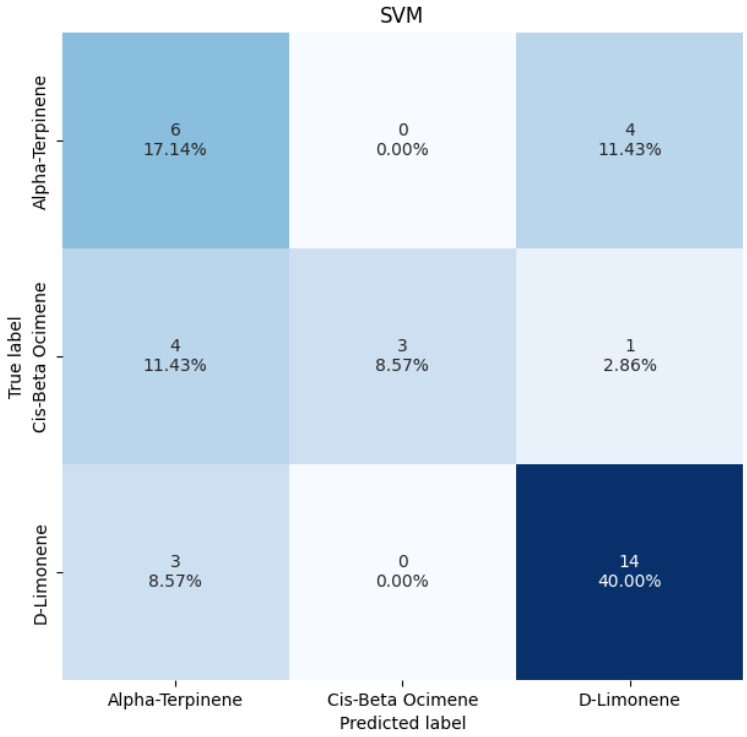


Figure 4.23: Confusion Matrix Heatmap - SVM D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

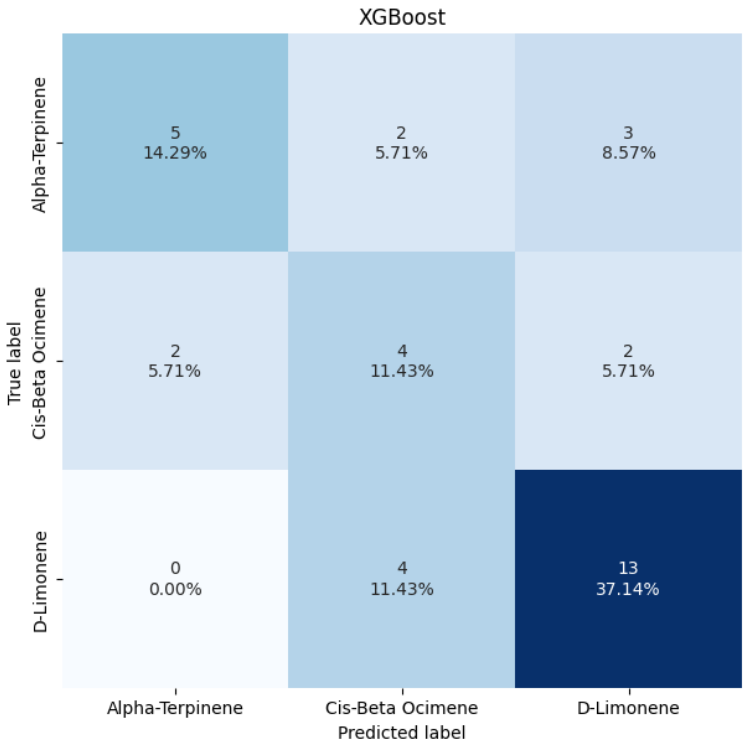


Figure 4.24: Confusion Matrix Heatmap - XGBoost D-Limonene vs Cis-Beta Ocimene vs Alpha-Terpinene

D-Limonene vs Cis-Beta Ocimene vs Citral

In this subsection we observe the chemicals without alpha-terpinene. We observe D-Limonene, cis-Beta ocimene, and citral. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.25 and Figure 4.25, Table 4.26 and Figure 4.26, and Table 4.27 and Figure 4.27 respectively. The accuracy for the Random Forest and XGBoost are higher than those for all chemicals. The Random Forest model had an accuracy of 73% (compared to 60%) and the XGBoost had an accuracy of 70% (compared to 60%). The SVM model had a lower accuracy than all chemicals with a 43% accuracy (compared to 62%).

Table 4.25: Random Forest - D-Limonene vs Cis-Beta Ocimene vs Citral

Chemical	Precision	Recall	F1-Score	Support
Citral	0.83	0.83	0.83	6
Cis-Beta Ocimene	0.60	0.60	0.60	10
D-Limonene	0.76	0.76	0.76	21
Accuracy			0.73	37
Macro Avg	0.73	0.73	0.73	37
Weighted Avg	0.73	0.73	0.73	37

Table 4.26: SVM - D-Limonene vs Cis-Beta Ocimene vs Citral

Chemical	Precision	Recall	F1-Score	Support
Citral	0.14	0.33	0.20	6
Cis-Beta Ocimene	0.71	0.50	0.59	10
D-Limonene	0.56	0.43	0.49	17
Accuracy			0.43	37
Macro Avg	0.47	0.42	0.42	37
Weighted Avg	0.54	0.43	0.47	37

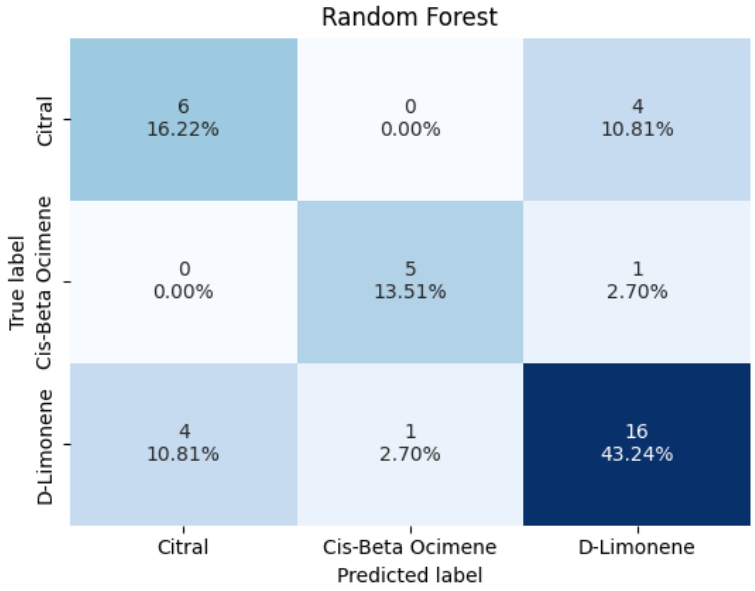


Figure 4.25: Confusion Matrix Heatmap - Random Forest D-Limonene vs Cis-Beta Ocimene vs Citral

Table 4.27: XGBoost - D-Limonene vs Cis-Beta Ocimene vs Citral

Chemical	Precision	Recall	F1-Score	Support
Citral	0.62	0.83	0.71	6
Cis-Beta Ocimene	0.62	0.50	0.56	10
D-Limonene	0.76	0.76	0.76	21
Accuracy			0.70	37
Macro Avg	0.67	0.70	0.68	37
Weighted Avg	0.70	0.70	0.70	37

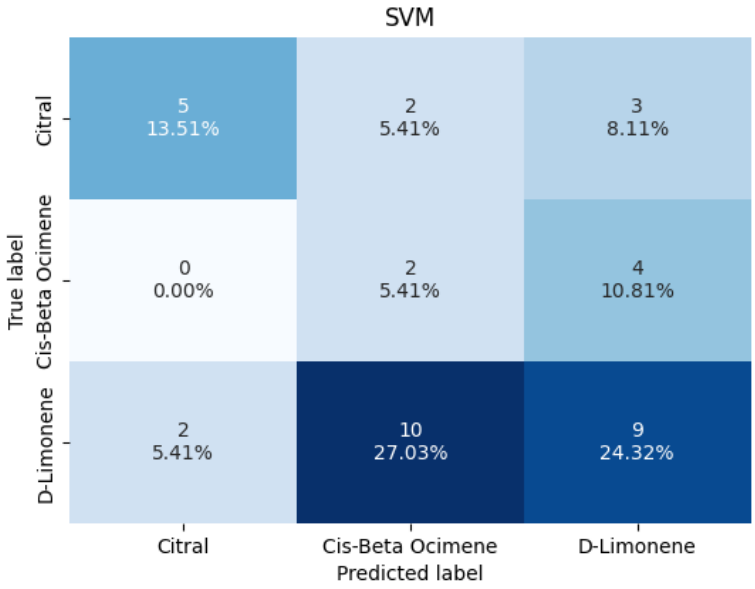


Figure 4.26: Confusion Matrix Heatmap - SVM D-Limonene vs Cis-Beta Ocimene vs Citral

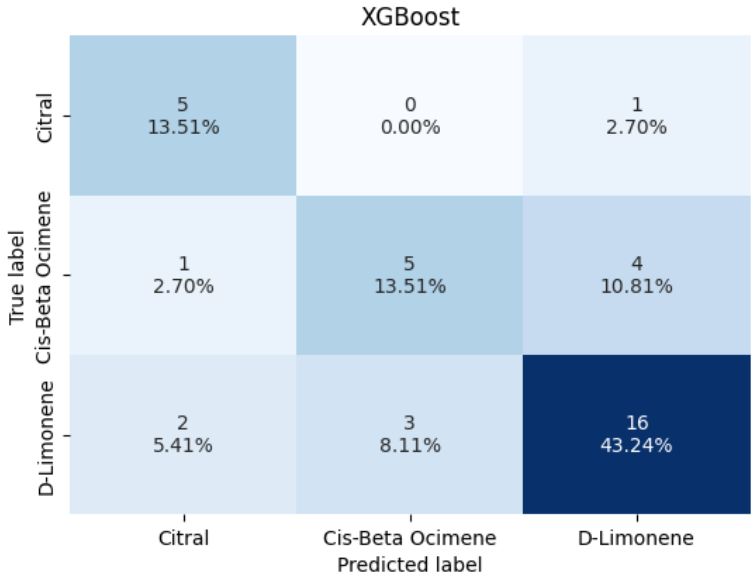


Figure 4.27: Confusion Matrix Heatmap - XGBoost D-Limonene vs Cis-Beta Ocimene vs Citral

D-Limonene Vs Cis-Beta Ocimene

Due to the high accuracy of D-Limonene and Cis-Beta Ocimene in the all chemical test, we used the models to determine the difference between D-Limonene and Cis-Beta Ocimene. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.28 and Figure 4.28, Table 4.29 and Figure 4.29, and Table 4.30 and Figure 4.30 respectively. Using only D-Limoneone and Cis-Beta Ocimene, the models had a lot higher accuracy with the random forest having an accuracy of 89%. XGBoost and SVM performed decently well with an accuracy of 78%.

Table 4.28: Random Forest - D-Limonene vs Cis-Beta Ocimene

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	0.88	0.78	0.82	9
D-Limonene	0.89	0.94	0.92	18
Accuracy			0.89	27
Macro Avg	0.88	0.86	0.87	27
Weighted Avg	0.89	0.89	0.89	27

Table 4.29: SVM - D-Limonene vs Cis-Beta Ocimene

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	1.00	0.33	0.50	9
D-Limonene	0.75	1.00	0.86	18
Accuracy			0.78	27
Macro Avg	0.88	0.67	0.68	27
Weighted Avg	0.83	0.78	0.74	27

Table 4.30: XGBoost - D-Limonene vs Cis-Beta Ocimene

Chemical	Precision	Recall	F1-Score	Support
Cis-Beta Ocimene	0.71	0.56	0.63	9
D-Limonene	0.80	0.89	0.84	18
Accuracy			0.78	27
Macro Avg	0.76	0.72	0.73	27
Weighted Avg	0.77	0.78	0.77	27

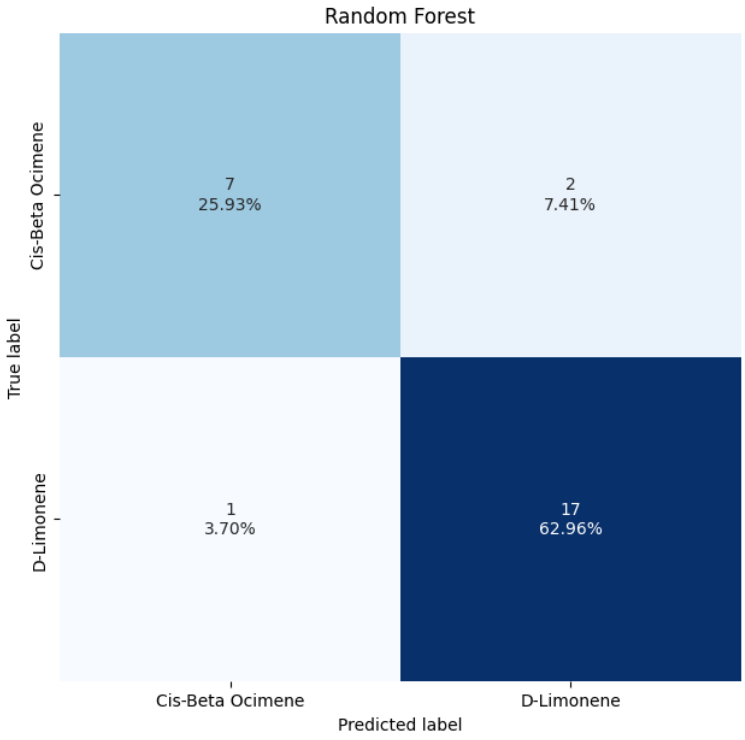


Figure 4.28: Confusion Matrix Heatmap - Random Forest D-Limonene vs Cis-Beta Ocimene

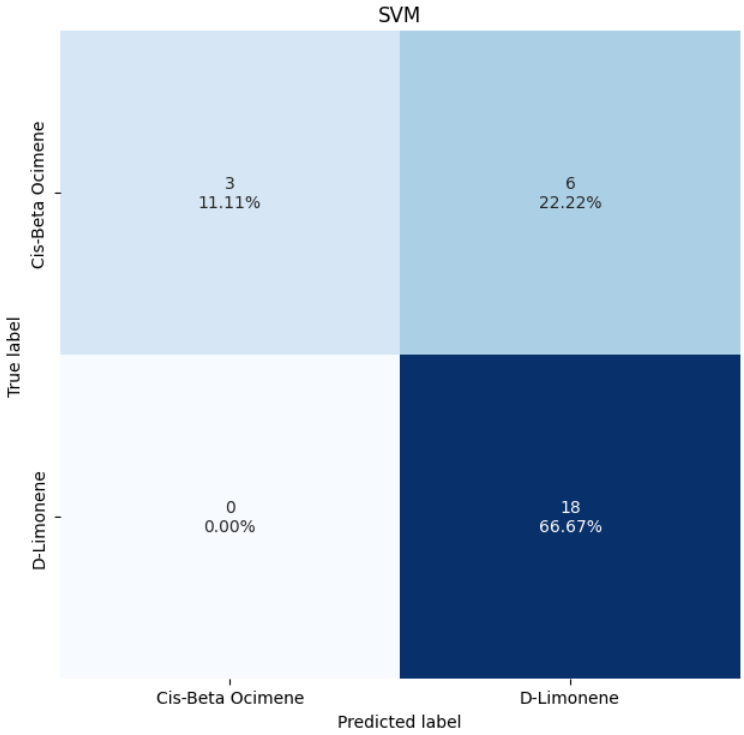


Figure 4.29: Confusion Matrix Heatmap -SVM D-Limonene vs Cis-Beta Ocimene

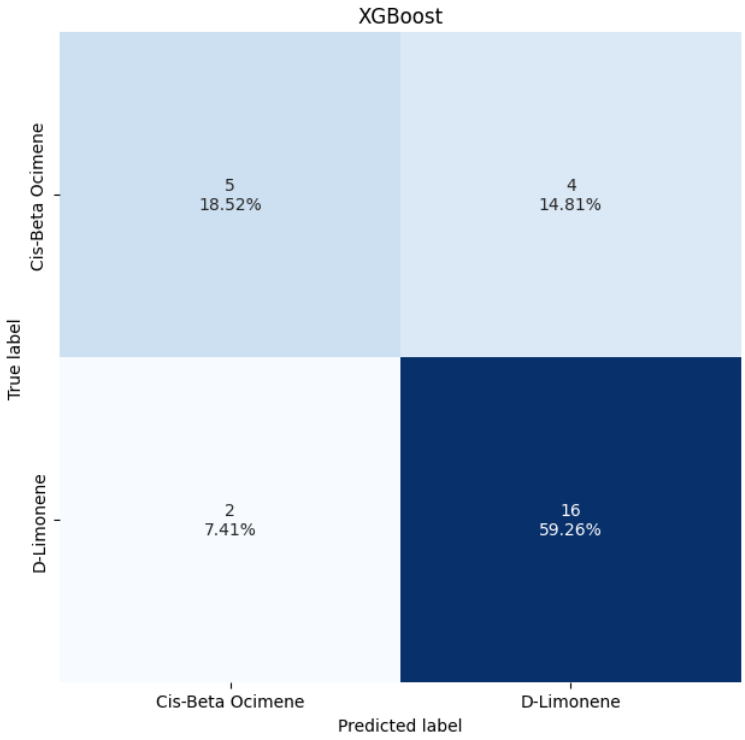


Figure 4.30: Confusion Matrix Heatmap - XGBoost D-Limonene vs Cis-Beta Ocimene

4.3.5 Plants vs Nothing

The final part of the results was to observe if the models can tell the difference between a plant being stressed and nothing in the room. The Random Forest, SVM, and XGBoost results and confusion matrices can be seen in Table 4.31 and Figure 4.31, Table 4.32 and Figure 4.32, and Table 4.33 and Figure 4.33 respectively. The XGBoost model had the best accuracy with 95%, followed by the random forest with 90%, and SVM with 75%. Although the accuracy is high, it is important to note that only 6 basil stress tests were in the testing set.

Table 4.31: Random Forest - Plant vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Basil Plant	0.83	0.83	0.83	6
Nothing	0.93	0.93	0.93	14
Accuracy			0.90	20
Macro Avg	0.88	0.88	0.88	20
Weighted Avg	0.90	0.90	0.90	20

Table 4.32: SVM - Plant vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Basil Plant	0.56	0.83	0.67	6
Nothing	0.91	0.71	0.80	14
Accuracy			0.75	20
Macro Avg	0.73	0.77	0.73	20
Weighted Avg	0.80	0.75	0.76	20

Table 4.33: XGBoost - Plant vs Nothing

Chemical	Precision	Recall	F1-Score	Support
Basil Plant	0.86	1.00	0.92	6
Nothing	1.00	0.93	0.96	14
Accuracy			0.95	20
Macro Avg	0.93	0.96	0.94	20
Weighted Avg	0.96	0.95	0.95	20

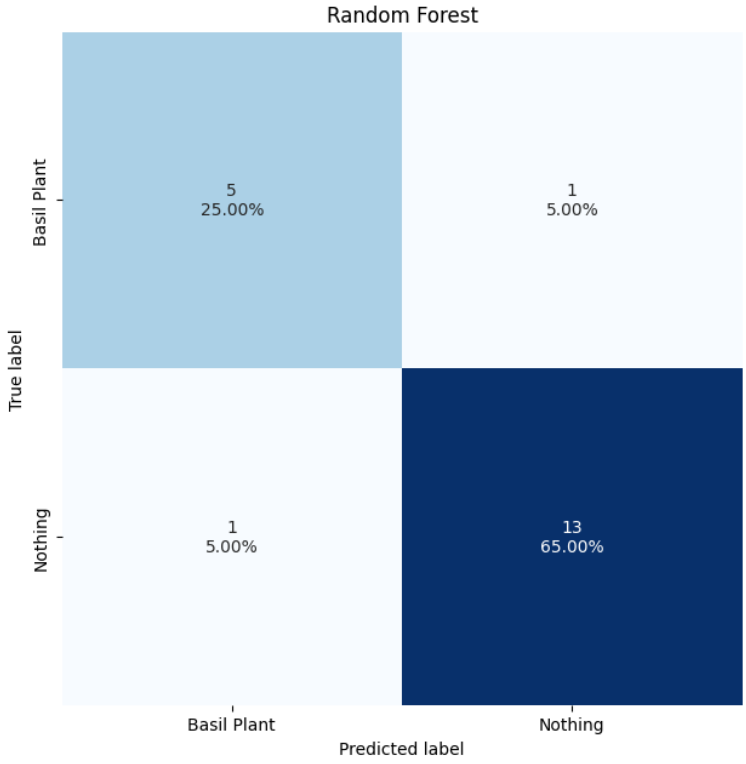


Figure 4.31: Confusion Matrix Heatmap - Random Forest - Plant vs Nothing

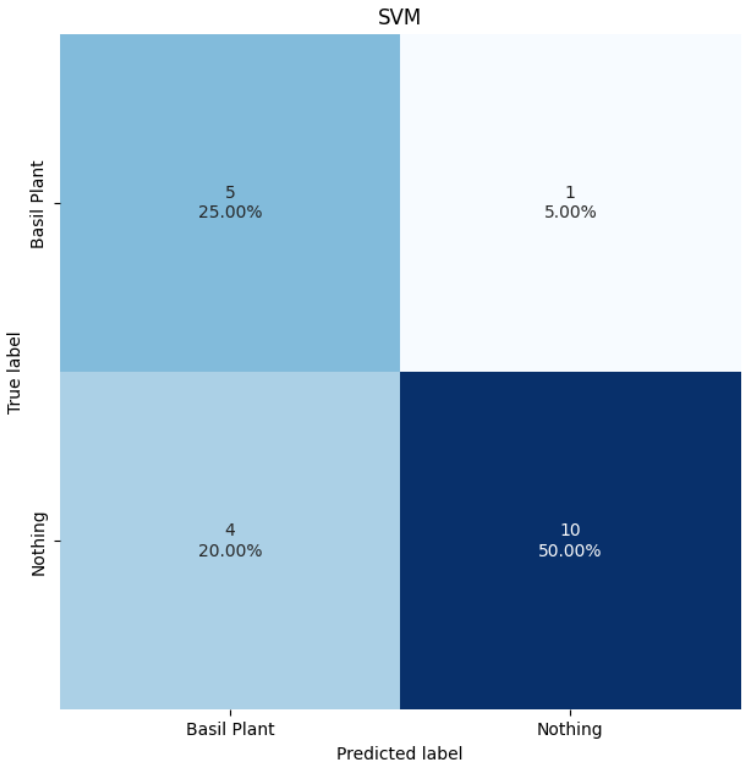


Figure 4.32: Confusion Matrix Heatmap - SVM - Plant vs Nothing

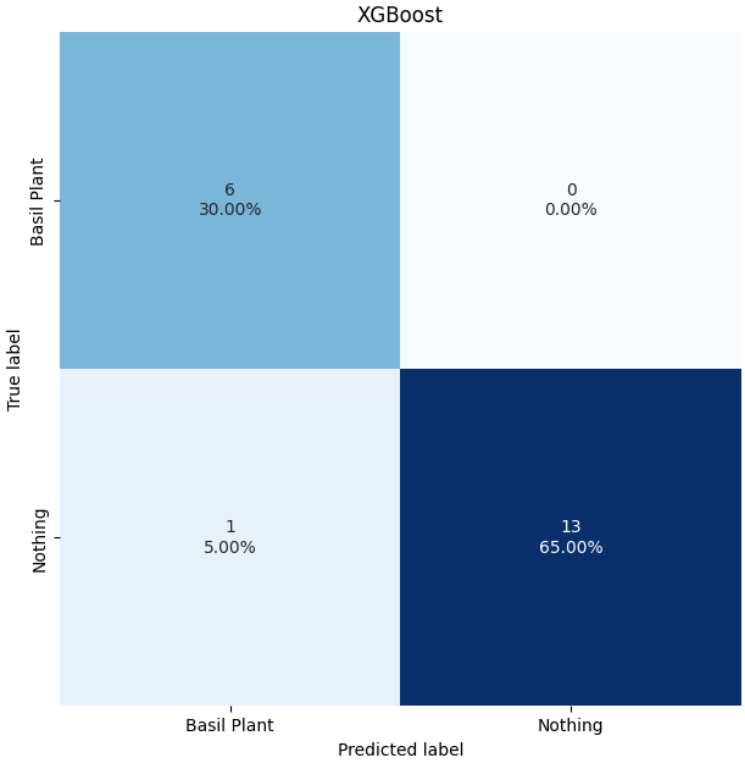


Figure 4.33: Confusion Matrix Heatmap - XGBoost - Plant vs Nothing

4.3.6 Best Common Features

After running all of the tests, we observe the best features for classification. As best 15 features are selected by the program for each test, they are not always the same. Although they are different, three show up in at least 7 out of the 9 machine learning tests. These features are the autocorrelation lag, permutation entropy, and approximate entropy. Autocorrelation lag finds the correlation between values that are a certain timestamp apart. All models use lag values between 1 and 7 (10 seconds to 70 seconds). Permutation entropy captures the complexity of they system by capturing order relations between values of time series data. Approximate entropy measures how unpredictable the fluctuations in the data are. All of these features look at multiple sections of the time series data. These are different features than the normalized sensor response [28], reaction time [28], and median [29] suggested by previous studies.

4.4 Discussion

Overall, the use of machine learning allowed us to accurately determine if there is any chemical in the room, a specific chemical in the room, which chemical is in the room between 2 chemicals, and if a plant is stressed in the room. The tests against nothing in the room versus something in the room can be explained with the change in the VOC readings of the room. Naturally, the composition of VOCs in a room changes naturally throughout the day when nothing is in it [36]. Although there is a change, it is not as drastic as when a terpene is in the room. This makes it easier to identify when using features such as the autocorrelation lag, permutation entropy, and approximate entropy. With regard determining which of the two chemicals is in the room, expensive sensors can detect unique responses to different VOCs [20]. This demonstrates that sensors can have unique responses to the chemical, even with the lack of sensitivity provided by lower costing sensors [18]. If the responses are unique enough and there are limited chemicals to choose from, the machine learning model has the ability to determine

the difference.

We were unable to determine which chemical was in the room between 3 or 4 chemicals and could not tell the dosage of the chemical released. Due to the lack of sensitivity of lower cost sensors [18], commercial sensors are not as accurate as more expensive e-nose sensors, especially when not in a testbed designed to get the best VOC readings. E-nose sensors are able to distinguish the differences between different volatile profiles [19]. The lack of an expensive sensor combined with the nature of VOC curves to be different with consistent dosages and times [30] can make it difficult to collect similar enough data each time. This can cause accuracy to go down when using terpenes with similar "ideal" VOC curves, as the sensor insensitivity and environment can cause variations in the readings.

4.5 Takeaways

The final study combines machine learning with VOC curves to classify what is in the room. The machine learning models were able to accurately determine if any chemical was in the room, if a specific chemical was in the room, if a basil plant was stressed, and could distinguish between two of them chemicals. This means that we can determine if genetically modified plants released a volatile due to a pathogen. Furthermore, we could distinguish between two pathogens by having the plant release one terpene for one pathogen and another for a different pathogen. We were unable to get machine learning to be accurate for the different dosages and when there are 3 or 4 chemicals to chose from. This is due to individual sensor sensitivity and lack of an environment where good VOC data can be collected. The enviornment can impact the spread of VOCs and therefore impact the readings for each individual sensor.

Chapter 5

Concluding Remarks

5.1 Future directions

Many different directions can be taken from the results found in this study. Firstly, more data can be gathered for different spaces and different senor locations to see how the room size impacts the AWAIR sensor. Based on the results of the emission study, the sensor location can have an impact on the readings. By knowing this impact, we can modify the simple mass based equation better predict the emission rate and terpene in the room.

Secondly, different sensors can be used to gather the data. The 4 AWAIR sensor used for this study could have their own sensitivity meaning that other AWAIR sensors may be better. Furthermore, expensive sensors can detect unique responses to different VOCs [20]. Our machine learning models could get better accuracy if there is less noise in the data. If purchasing an e-nose sensors is too expensive, we propose creating a new e-nose by combining multiple low powered sensors.

Finally, gathering plant data can also be improved. As plant VOCs can be impacted by various factors and can sometimes be inconsistent, it is important to have a stable environment where the plant conditions are maintained and VOCs from the plant are captured. This is vital to getting plants to work as sensors as they need to emit enough VOCs for there to be a signal. To do this, we suggest an eco-chamber where plant conditions such as light, water, soil, and humidty are maintained. This eco-chamber would take in air from the outside environment so the genetically modified plant can release the VOC if pathogens are detected.

After a set amount of time, the VOCs built up in the eco-chamber would be released into a a container holding the AWAIR sensor so it can capture the VOC signal.

5.2 Summary

Understanding and utilizing raw plant terpene data from commercial sensors is complicated given the issue of sensor sensitivity, air flow, sensor location, and difference between terpenes. This paper lays the groundwork for using commercial sensors to interpret genetically modified plant VOCs. Plant terpines can be detected by commercial. Although they can be detected, they data provided does not work with real world equations. We show a proof of concept that machine learning can be used to determine if a chemical is in the air, a plant has been stressed, and differentiate between two chemicals. Using this knowledge we can genetically modify plants to release volatile only when a pathogen is in the air. Future studies can be conducted to create better test beds which get more accurate plant data.

REFERENCES

[1] M. Riedlmeier, A. Ghirardo, M. Wenig, C. Knappe, K. Koch, E. Georgii, S. Dey, J. E. Parker, J.-P. Schnitzler, and A. C. Vlot, "Monoterpenes support systemic acquired resistance within and between plants," *The Plant Cell*, vol. 29, no. 6, pp. 1440–1459, 2017.

[2] R. A. Raguso, "Wake up and smell the roses: the ecology and evolution of floral scent," *Annual review of ecology, evolution, and systematics*, vol. 39, pp. 549–569, 2008.

[3] C. E. Vickers, M. Possell, C. I. Cojocariu, V. B. Velikova, J. Laothawornkitkul, A. Ryan, P. M. Mullineaux, and C. Nicholas Hewitt, "Isoprene synthesis protects transgenic tobacco plants from oxidative stress," *Plant, Cell & Environment*, vol. 32, no. 5, pp. 520–531, 2009.

[4] F. Loreto, M. Mannozzi, C. Maris, P. Nascetti, F. Ferranti, and S. Pasqualini, "Ozone quenching properties of isoprene and its antioxidant role in leaves," *Plant Physiology*, vol. 126, no. 3, pp. 993–1000, 2001.

[5] I. T. Baldwin, "Plant volatiles," *Current Biology*, vol. 20, no. 9, pp. R392–R397, 2010.

[6] A. Pathak and C. Viphavakit, "A review on all-optical fiber-based voc sensors: Heading towards the development of promising technology," *Sensors and Actuators A: Physical*, p. 113455, 2022.

[7] H. Bouwmeester, R. C. Schuurink, P. M. Bleeker, and F. Schiestl, "The role of volatiles in plant communication," *The Plant Journal*, vol. 100, no. 5, pp. 892–907, 2019.

[8] P.-J. Zhang, J.-N. Wei, C. Zhao, Y.-F. Zhang, C.-Y. Li, S.-S. Liu, M. Dicke, X.-P. Yu, and T. C. Turlings, "Airborne host–plant manipulation by whiteflies via an inducible blend of plant volatiles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 15, pp. 7387–7396, 2019.

[9] F. Loreto, M. Dicke, J.-P. SCHNITZLER, and T. C. Turlings, "Plant volatiles and the environment," *Plant, cell & environment*, vol. 37, no. 8, pp. 1905–1908, 2014.

[10] J. K. Holopainen and J. Gershenzon, "Multiple stress factors and the emission of plant vocs," *Trends in plant science*, vol. 15, no. 3, pp. 176–184, 2010.

[11] T. D. Sharkey and E. L. Singsaas, "Why plants emit isoprene," *Nature*, vol. 374, no. 6525, pp. 769–769, 1995.

[12] L. O. Copolovici, I. Filella, J. Llusià, U. Niinemets, and J. Peñuelas, "The capacity for thermal protection of photosynthetic electron transport varies for different monoterpenes in quercus ilex," *Plant Physiology*, vol. 139, no. 1, pp. 485–496, 2005.

[13] D. Materić, D. Bruhn, C. Turner, G. Morgan, N. Mason, and V. Gauci, "Methods in plant foliar volatile organic compounds research," *Applications in Plant Sciences*, vol. 3, no. 12, p. 1500044, 2015.

[14] E. A. Schmelz, H. T. Alborn, J. Engelberth, and J. H. Tumlinson, "Nitrogen deficiency increases volicitin-induced volatile emission, jasmonic acid accumulation, and ethylene sensitivity in maize," *Plant Physiology*, vol. 133, no. 1, pp. 295–306, 2003.

[15] E. Ormeño, V. Baldy, C. Ballini, and C. Fernandez, "Production and diversity of volatile terpenes from plants on calcareous and siliceous soils: effect of soil nutrients," *Journal of chemical ecology*, vol. 34, pp. 1219–1229, 2008.

[16] E. Ormeño, A. Bousquet-Mélou, J.-P. Mévy, S. Greff, C. Robles, G. Bonin, and C. Fernandez, "Effect of intraspecific competition and substrate type on terpene emissions from some mediterranean plant species," *Journal of chemical ecology*, vol. 33, pp. 277–286, 2007.

[17] S. P. Gouinguené and T. C. Turlings, "The effects of abiotic factors on induced volatile emissions in corn plants," *Plant physiology*, vol. 129, no. 3, pp. 1296–1307, 2002.

[18] L. Spinelle, M. Gerboles, G. Kok, S. Persijn, and T. Sauerwald, "Review of portable and low-cost sensors for the ambient air monitoring of benzene and other volatile organic compounds," *Sensors*, vol. 17, no. 7, p. 1520, 2017.

[19] M. Delgado-Rodríguez, M. Ruiz-Montoya, I. Giraldez, R. López, E. Madejón, and M. Díaz, "Use of electronic nose and gc-ms in detection and monitoring some voc," *Atmospheric Environment*, vol. 51, pp. 278–285, 2012.

[20] S. Dragonieri, G. Pennazza, P. Carratu, and O. Resta, "Electronic nose technology in respiratory diseases," *Lung*, vol. 195, pp. 157–165, 2017.

[21] M. Rasekh, H. Karami, A. D. Wilson, and M. Gancarz, "Performance analysis of mau-9 electronic-nose mos sensor array components and ann classification methods for discrimination of herb and fruit essential oils," *Chemosensors*, vol. 9, no. 9, p. 243, 2021.

[22] S. MacDougall, F. Bayansal, and A. Ahmadi, "Emerging methods of monitoring volatile organic compounds for detection of plant pests and disease," *Biosensors*, vol. 12, no. 4, p. 239, 2022.

[23] W. E. O. Ghoma, H. Sevik, and K. Isinkaralar, "Using indoor plants as biomonitors for detection of toxic metals by tobacco smoke," *Air Quality, Atmosphere & Health*, vol. 15, no. 3, pp. 415–424, 2022.

[24] A. Zuniga, N. H. Motlagh, H. Flores, and P. Nurmi, "Smart plants: Low-cost solution for monitoring indoor environments," *IEEE internet of things journal*, vol. 9, no. 22, pp. 23 252–23 259, 2022.

[25] H. Sareen, "Cyborg botany: augmented plants as sensors, displays and actuators," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[26] A. Mulgrew and P. Williams, "Biomonitoring of air quality using plants," 2000.

[27] J. P. Giraldo, H. Wu, G. M. Newkirk, and S. Kruss, "Nanobiotechnology approaches for engineering smart plant sensors," *Nature nanotechnology*, vol. 14, no. 6, pp. 541–553, 2019.

[28] M. Gancarz, A. Nawrocka, and R. Rusinek, "Identification of volatile organic compounds and their concentrations using a novel method analysis of mos sensors signal," *Journal of food science*, vol. 84, no. 8, pp. 2077–2085, 2019.

[29] H. Koç, J. King, G. Teschl, K. Unterkofler, S. Teschl, P. Mochalski, H. Hinterhuber, and A. Amann, "The role of mathematical modeling in voc analysis using isoprene as a prototypic example," *Journal of breath research*, vol. 5, no. 3, p. 037102, 2011.

[30] A. M. Yeoman, M. Shaw, and A. C. Lewis, "Estimating person-to-person variability in voc emissions from personal care products used during showering," *Indoor air*, vol. 31, no. 4, pp. 1281–1291, 2021.

[31] M. Arora, S. C. Zambrzycki, J. M. Levy, A. Esper, J. K. Frediani, C. L. Quave, F. M. Fernández, and R. Kamaleswaran, "Machine learning approaches to identify discriminative signatures of volatile organic compounds (vocs) from bacteria and fungi using spme-dart-ms," *Metabolites*, vol. 12, no. 3, p. 232, 2022.

[32] A. Ponzoni, A. Depari, M. Falasconi, E. Comini, A. Flammini, D. Marioli, A. Taroni, and G. Sberveglieri, "Bread baking aromas detection by low-cost electronic nose," *Sensors and Actuators B: Chemical*, vol. 130, no. 1, pp. 100–104, 2008.

[33] H. Hori, S. Ishimatsu, Y. Fueta, and T. Ishidao, "Evaluation of a real-time method for monitoring volatile organic compounds in indoor air in a japanese university," *Environmental health and preventive medicine*, vol. 18, no. 4, pp. 285–292, 2013.

[34] "PPB conversion," https://www2.dmu.dk/atmosphericenvironment/expost/database/docs/ppm_conversion.pdf, accessed: 2023-04-02.

[35] R. Schapire, "Machine learning algorithms for classification," *Princeton University*, vol. 10, 2015.

[36] M. J. Hewitt, I. Belluomo, S. Zuffa, P. R. Boshier, and A. Myridakis, "Variation of volatile organic compound levels within ambient room air and its impact upon the standardisation of breath sampling," *Scientific Reports*, vol. 12, no. 1, p. 15887, 2022.

A Appendix A

B Appendix B

C Papers Submitted and Under Preparation

• Author 1 Name, Author 2 Name, and Author 3 Name, "Article Title", *Submitted to Conference/Journal Name*, further attributes.

• Author 1 Name, Author 2 Name, and Author 3 Name, "Article Title", *Submitted to Conference/Journal Name*, Mon. Year.

• Author 1 Name, Author 2 Name, and Author 3 Name, "Article Title", *Submitted to Conference/Journal Name*, Mon. Year.