ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

SENSOR PLATFORM BASED ON ENVIRONMENTAL SENSING AND DATA FUSION

M.Sc. THESIS

Muhammad Ali NAGARIA

Department of Mechatronics Engineering
Mechatronics Engineering Graduate Program

DECEMBER 2018

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

SENSOR PLATFORM BASED ON ENVIRONMENTAL SENSING AND DATA FUSION

M.Sc. THESIS

Muhammad Ali NAGARIA
518161039

Department of Mechatronics Engineering

Mechatronics Engineering Graduate Program

Thesis Advisor: Assoc. Prof. Dr. Hülya YALÇIN

DECEMBER 2018

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ÇEVRESESEL ETMENLERİ ALGILAYAN SENSÖR PLATFORMU VE VERİ ANALİZİ

YÜKSEK LİSANS TEZİ

Muhammad Ali NAGARIA
518161039

Mekatronik Mühendisliği Anabilim Dalı

Mekatronik Mühendisliği Programı

Tez Danışmanı: Dr. Öğr. Üyesi Hülya YALÇIN

ARALIK 2018

Muhammad Ali Nagaria, a M.Sc. student of İTU Graduate School of Science Engineering and Technology student, successfully defended the thesis entitled "SENSOR PLATFORM BASED ON ENVIRONMENTAL SENSING AND DATA FUSION", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Dr. Öğr. Üyesi Hülya YALÇIN** İstanbul Technical University	.....
Jury Members :	**Dr. Öğr. Üyesi Hatice MERCAN** Yıldız Technical University	.....
	Doç. Dr. Gülay ÖKE GÜNEL İstanbul Teknik Üniversitesi	.....

Date of Submission : 16 November 2018
Date of Defense : 11 December 2018

To my mentors,

"In God we trust, all others must bring data."

William Edwards Deming

FOREWORD

As I approach the end of my master's degree at İstanbul Technical University, I would like to express my gratitude to all the instructors that have contributed to my development through this multi-disciplinary program. In particular, I would thank the advisor of the thesis Dr. Hülya Yalçın whose interactive style of teaching has enabled me to grasp fundamental concepts of Machine Learning throughout her courses.

To my team leaders at ARÇELIK Sensor Technologies R&D, who assisted me to set aside to time to continue my education despite demanding professional activities. In addition, they provided the necessary infrastructure to perform the extensive tests required for this thesis.

To Mr. Ramin Fadaei Fouladi for offering a priceless critique of my work, and for pushing me to do better at each stage.

Finally, but most importantly my parents who have enabled me to reach this stage and to my spouse Aisha Shamim whose mental support got me through the most challenging phases of this degree.

This thesis was completed with the assistance of ARÇELIK A.Ş. Central R&D Directorate and the Sensor Technologies Team as an R&D Project.

November 2018

Muhammad Ali Nagaria
(Mechatronics Engineer)

TABLE OF CONTENTS

Page

ABBREVIATIONS

VOC:	Volatile Organic Compounds
PM:	Particulate matter
MQTT:	Message Queuing Telemetry Transport
PCA:	Principal Component Analysis
LDA:	Linear Discriminant Analysis
MLP:	Multi-layer perceptron
I^2C:	Inter-integrated circuit communication protocol
SQL:	Structured Query Language
PCB:	Printed Circuit Board

LIST OF TABLES

Page

LIST OF FIGURES

Page

SENSOR PLATFORM BASED ON ENVIRONMENTAL SENSING AND DATA FUSION

SUMMARY

Since a large portion of population spends >90% of their time indoors, it is important to measure and report indoor air quality to the users. These measurements of fine dust, carbon monoxide, volatile organic compounds, carbon monoxide are usually reduced to an air quality index to make reporting easier. The problem with such a metric is that the user does not know how much daily activities such as cooking, cleaning and smoking are influencing the air quality, so that they can improve their behaviour, or the smart-home can trigger signals to other devices to maintain a healthy environment when these activities are performed. As a step to develop such a system, this study focuses on developing a system that measures 9 different air quality parameters and is able to classify activities into one of the four classes (Background, Cleaning, Cooking, Smoking). The study contains a literature review of Machine learning techniques employed in similiar studies, and drawing influence from works in electronic noses and artificial olfaction for applications of food inspection. A few works also detected human activity and count. The selection of literature was done to include studies that use similar sensors, namely gas sensors of metal oxide semiconductor type, and optical solutions to measure CO_2 and dust constrations. An alternative CFD simulation based approach is also seen in literature, but the focus was kept on sensor data analaysis. Rolling window feature extraction and normalization techniques are popular in this regard, before putting them through a dimentionality reduction algorithm such as PCA to finally classify it using popular classifiers. Some studies have attempted to eliminate the need of feature extraction using Deep Neural Networks but the works are scarce. Each study attempts to create its own dataset, with the unique selection of sensors, so techniques are highly application specific. To asses the feasibilibility in the context of Indoor Air Quality activity detection, an IoT platform is prepared for data acquisition

complete with a sensor hub PCB and a mySQL server to collect important IAQ parameters at regular intervals. A dataset of 80 hours of air quality data is prepared and labelled in a test room. In the last section, the dataset is set as a supervised machine learning problem, time series feature extraction (Vector Array Normalization, Hurst Exponent) and classification techniques (Decision Trees, LDA, MLP) studied in literature are evaluated. The problem is evaluated in two scenarios, one is of binary classification where the algorithms are trained to distinguish between two classes, ‘Background’ and ‘Event’,. The second is the multiclass case where there are four classes ‘Background’, ‘Cleaning’, ‘Cooking’, and ‘Smoking’. The best accuracy at the end of this study is obtained as 81% for the multi-class case, and 94% for the binary one . The recommendation for future work includes extension of the methodology to more uncontrolled scenarios and eventually an actual living space such as a home or office.

ÇEVRESESEL ETMENLERİ ALGILAYAN SENSÖR PLATFORMU VE VERİ ANALİZİ

ÖZET

Günümüzde artış göstermekte olan endüstriyel üretim ve dünya popülasyonunun genişlemesi beraberinde çevre koşullarındaki değişimler insan türünün yaşamının sürdürülebilirliği için çeşitli ihtiyaçlar doğurmuş bulunmaktadır. Bu durum üretim bazında olduğu gibi tüketici tarafında da önemli etkenlere tabidir; bunlardan en önemlilerinden biri de iç mekanlardaki hava koşullarıdır. Şehir nüfusunun zamanının yüzde 90'dan fazlasının kapalı alanlarda geçirdiğini göz önünde bulundurursak kullanıcılar açısından bu alanlardaki hava kalitesinin ölçümlenmesi ve paporlarnması önem kazanmaktadır. Bu ölçüm ve raporların hazırlanmasını kolaylaştırmak için ince toz (PM), karbon monoksit (CO), uçucu organik bileşenler (VOC), karbon dioksit (CO2) gibi maddelerin ortamdaki mevcudiyetini değişkenler haline getirerek, bir arada değerlendirilmesini sağlayan ortak bir "hava kalitesi endeksi" kullanılmaktadır. İç mekanlarda çoğunlukla karşılaştığımız faaliyetler uçucu kimyasal maddelerle yapılan hijyen ve temizlik uygulamaları, besin ihtiyacını gidermeyek yönelik pişirme faaliyleri ve yine yaygın olarak karşılaştığımız tütün mamülleri tüketimidir. Bu duruma bağlı olarak endeks içerisinde öncelikli olarak bu değişkenler önem kazanmaktadır. Aynı zamanda bu 3 faaliyet birbirinden farklı iz bölgesi yaratmaktadır. Ancak ortada olan sorun şudur ki, bu tür cihazların kullanıcıları günlük hayatta sıklıkla gerçekleştirdikleri yemek yapma, temizlik ve sigara içmek gibi faaliyetlerin bulundukları ortamın hava kalitesine etkisinden haberdar olamamaktadırlar. Bu bilgi hem kullanıcıların bulundukları alanın hava kalitesini iyileştirmek için önlemler almalarına imkan tanıyacağı gibi, akıllı ev teknolojilerinin bulunduğu ortamda da çeşitli aygıtların birbiriyle iletişime geçmesini sağlayarak ortamın solunum açısından daha sağlıklı bir hal almasına imkan tanıyacaktır. Bu tür bir sistemin geliştirilmesi için ön ayak olmayı amaçlayan bu araştırma 9 farklı hava kalitesi parametresini ölçümleyen bir mekanizma geliştirmek ve bu 9 değişkenin 4

farklı sınıfta (Arkaplan, Temizlik Faaliyeti, Yemek Yapmak, Sigara İçmek) tanımlamasını sağlmayı hedeflemektedir.

Çalışma amacına ulaşmak için benzer araştırmalarda kullanılan Makina Öğrenme tekniklerine dair bir literatür bulundurmakta olup, besin incelemeleri için kullanılan elektronik burun ve yapar koku alma uygulamalarından da etkiler barındırmaktadır. Bazı çalışmalar aynı zamanda insan faaliyetleri ve sayımlarını da içermektedir. Literatür araştırması için tercih edilen kaynaklar özellikle ilgili sensörlerin kullanıldığı araştırmaları bulundurmaktadır, bunlar başlıca metal oksit yarı iletkenleri tipi gaz sensörleri ve karbon dioksit ile toz yoğunluğunu gözlemlemek için kullanılan optik solüsyonlardır. Literatürde alternatif CFD temelli yaklaşımlara da rastlanmış olsa da, araştırmanın sensör data analizleri çevresinde geliştirilmiştir.Yuvarlanan pencere öznitelik çıkarım ve normalleştirme teknikleri bu çerçevede yaygındır, bu teknikten sonra en yaygın teknik ise PCA gibi boyut indirgeme algoritması kullanarak yaygın sınıflandırıcılar üzerinden tanımlama metodudur.

Bazı araştırmalar Yapay Sinir Ağları (Deep Neural Networks) kullanımıyla öznitelik çıkarımları ihtiyacını ortadan kaldırma girişiminde bulunmuştur ancak bu tür araştırmalar nadir olarak bulunmaktadır. Neredeyse tüm araştırmalar diğerlerinden farklı sensörler kullanarak kendine has veri çıktıları elde etmektedir; bu duruma bağlı olarak teknikler de uygulamalama özelinde kalmaktadır. İç Mekan Hava Kalitesi ölçümleri çerçevesi fizilibitesinin değerlendirilmesi için Sensor Hub, PCB ve tüm önemli IAQ parametrelerini düzenli aralıklarla toplayacak bir mySQL sunucusu bulunduran IoT playformu hazırlanmıştır. Bu platformda insan yaşamı ve konforunu da göz önünde bulundurarak öncelikli olarak sıcaklık, nem, ucucu organik bileşenler, karbondioksit, karbonmonoksit, amonyak, nitrojendioksit ve 25 mikrometre altı maddelerin ölçümü için 4 farklı sensör kullanılmıştır. 80 Saatlik bir ölçüm sonucu hava kalitesi veri seti hazırlanmış ve test odasında sınıflandırılmıştır. Araştırmanın son bölümünde ise veri seti denetimli bir özdevimli öğrenme sorunu olarak kurulmuş; literatürde bulunan zaman serileri öznitelik çıkarımı teknikleri (Vector Array Normalization, Hurst Eksponenti) ve sınıflandırma teknikleri (Decision Tree, LDA, MLP) değerlendirilmiştir. Problem araştırma esnasında 2 senaryo dahilinde ele alınmıştır. Birinci senaryo iki terimli sınıflandırma çerçevesidir. Bu çerçevede algoritmalar 2 sınıf bazında ayrım yapmaya yönelik eğitilmiştir “Arkaplan” ve

“Olay”. İkinci senaryoda ise çok sınıflı bir yapı oluşturulmuştur, bunlar “Arkaplan”, “Temizlik”, “Yemek Yapma” ve “Sigara İçme” şeklinde 4 kategoride tanımlanmıştır. Tüm kategori testleri 22’şer kere değerlendirilmiş olup arkplan tanımlaması için 10 ek test gerçekleştirilerek ilgili sınıfların koşullarla bağlantıları sağlanılmıştır. Bu araştırmanın sonunda çok sınıflı senaryoda yüzde 81 oranında tutarlılık elde edilirken ikili sınıflandırma modelinde yüzde 94’lük bir tutarlılık elde edilmiştir.

Yapılan araştırma çevresel faktörleri göz önünde bulundurarak artan şehirleşme süreçlerinde insanlar için iç mekanlardaki hava koşullarının düzenlenmesi ve iyileştirilmesi adına öncül bir basamak oluşturmaktır. Gelecekte bu alanda yapılması planlanacak araştırmalar için metod çerçevesinin genişletilerek daha kontrolsüz senaryolar üzerinden testler yapılması ve ofis veya ev gibi gerçek yaşam alanlarında testler yapılmasını tavsiye edilir. Alternatif bir rota olarak yapay sinir ağları mimarileri değerlendirmeleri için zaman serileri modellemelerinde başarılı olan RNN, LSTM; verinin boyutsallığının düzenlenmesi için ise CNN kullanılabilir.

Sensör verilere aktivite bazında sonuçları elde ederek sadece çalışıldığı aktiviteler değil, daha farklı uygulama alanlarada yön açılabilir. Nesnelerin internet ortamlarında evde mevcüt bütün cihazlara sadece hava kalite sensörleri kullanarak senaryolar çalışılabilir. Hava kalite ölçümler hem ortamın sağlıklığa dair bilgi veriyor, hem bu çalışmada gösterildiği gibi kullancının kişisel bilgi koruyarak, aktivite takibi senoryalar gerçekleştirilebilir.

1. INTRODUCTION

Air is a medium that humans are continuously submerged in and its composition is necessary for our survival. Though Oxygen (O_2) is the key gas required for our respiration, we are required to breath and separate all other contents from air which for the most part is Nitrogen (N_2) and Carbon-dioxide (CO_2); and also, harmful substances like Particulate Matter ($PM_{2.5}$), Carbon Monoxide (CO), Volatile Organic Compounds (VOCs), Nitrogen Dioxide (NO_2) and many others. As modern cities get more and more crowded, Air Quality is becoming an important issue for 21st century. The main sources of these pollutants are outdoors (combustion, industrial pollution, airborne chemicals), but in indoor living spaces these pollutants can be trapped if the space is poorly ventilated. Also, indoor activities such as smoking, use of cleaning chemicals, and cooking are known to worsen the conditions further.

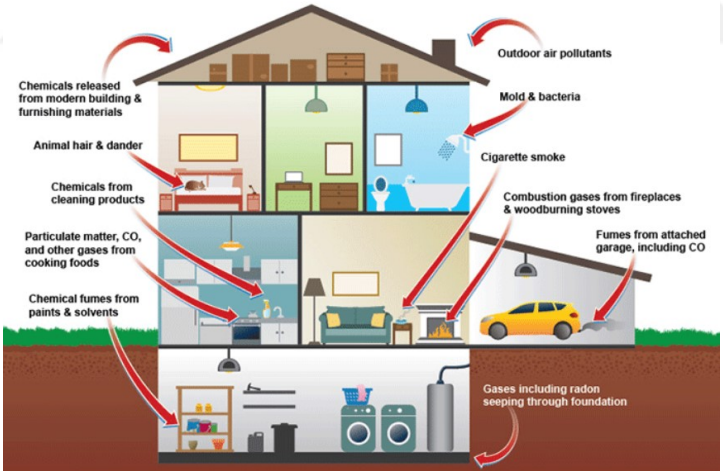


Figure 1.1: Sources of Indoor Air Pollutants [1]

Users do not realize the harm their activities are doing to the air they breathe unless measurements are provided to them [2]. For this purpose, sensing technologies such as Metal Oxide Semiconductor (for Volatile Air Compounds) and Non-Dispersive Infrared have been popular in academic and commercial systems. These sensors do measure the concentration of the pollutants, but very few give insights of what activities are causing these problems. The aim of this study is to develop a Sensor

platform that uses data fusion techniques to classify indoor air pollution events so that recommendations can be generated to either improve the user behavior, or the smart-home's behavior to ultimately maintain a healthy living environment.

1.1 Purpose of thesis

The purpose of this thesis is to develop a sensor platform that measures various air quality parameters and to investigate the possibility of detecting indoor air pollution events using the systems. The study should include, the sensors, embedded system design, and data analysis algorithms necessary for identifying 3 activities inside a test setup. The device will be in the form of a wireless hub that pushes data to a local server. The activities are listed as follows:

i. Cleaning using chemical solutions
ii. Cooking with oil
iii. Smoking tobacco

A literature review around ubiquitous sensing, electronic noses, and time series classification is needed to be performed and a data analysis method proposed. The system at the end of the study will be later evaluated for integration inside smart home applications.

Both for this study and future works, a labelled dataset for these activities needs to be constructed, whereby the sensor hub is to be used to perform to collect the required data.

1.2 Hypothesis

Each of the 3 activities listed previously is believed to have a unique fingerprint on the air, that can be detected through a combination of air quality sensors. The concept must be revaluated for a custom dataset, and a data processing scheme proposed.

2. AIR QUALITY SENSING IN LITERATURE

The focus of the study is concerning sensing problems for Air Quality data and its use in ubiquitous applications such as the electric nose and pollutant source identification. The first part of this section briefly mentions the health implications for certain indoor activities that effect the air quality.

The second part discusses in detail the kind of sensing problems exist in literature in the space of gas sensor arrays, and the techniques used to classify high level information.

2.1 Impact Of Selected Indoor Activities On Human Health

Many indoor human activities worsen the air quality and make the environment unhealthy for living. Table 2.1 show a summarized list of pollutants, health effects, and sources.

Exposure at a wide range of concentrations is not noticeable since most of these pollutants are odorless. Therefore sensors to monitor these levels is critical if we want to ensure optimum air quality inside the indoor space.

Table 2.1: Air pollutants their sources in indoor environment [3] [4] [5] [6]

	Sources/activities in indoor air	**Health implications at unhealthy levels**
Temperature	- Heating - Stoves - Lighting - Electronic devices	- Lethargy - Sleep deprivation - Excessive tiredness - Dehydration - Skin allergies
Relative humidity	- Perspiration - Water flow (showering) - Cooking	- Skin allergies - Respiratory diseases
Carbon Dioxide	- Human metabolism, - Indoor combustion flames (gas stoves) - Central heating systems	- Acidification of blood - İncrease in rate and depth of breathing - Prolonged exposure can affect calcium metabolism in bone

Table 2.1 (continued): Air pollutants their sources in indoor environment

Carbon Monoxide	- Tobacco smoke - Gas powered heaters - Defective central heating - automobile exhaust	- Reduces oxygen flow to brain results in: Nausea, unconsciousness, and even death.
Volatile Organic Compounds	- Paints and lacquers - Cleaning supplies - Building materials and furnishing - Office equipment (copiers, printers) - Glues, adhesives - Overheated cooking oils	- Conjunctival irritation - Nose and throat discomfort - Headache - Dyspnea - Fatigue - Epistaxis - dizziness
PM 2.5	- Combustion of solid and liquid fuels - Urban traffic - Construction sites - Cooking	- High risk of cardiovascular diseases - Pulmonary diseases - Eye irritation
PM 10	- Wood burning stoves - Agricultural burning - Dust	- Reduced lung function - Eye irritation - Coughing - Respiratory diseases
Nitrogen Dioxide	- Tobacco smoke - Gas/wood/oil/coal burning	- Depletes tissue antioxidant defenses - Irritant for eyes, nose, throat. - Pulmonary edema (leading to lung injury)
Ammonia	- Household cleaners - Animal/pet waste - Decaying plants/food	- Eye, nose, throat irritation - Headache - Nausea - Can cause permanent damage to eyes and lungs

2.2 Applications Of Indoor Air Quality Data

Due to increased awareness of Air Quality, and advance in affordable IAQ sensors, there are various products in the market that offer IoT Air Quality monitors. These products provide daily reports of pollutants and classify them with respect to healthy levels [7].

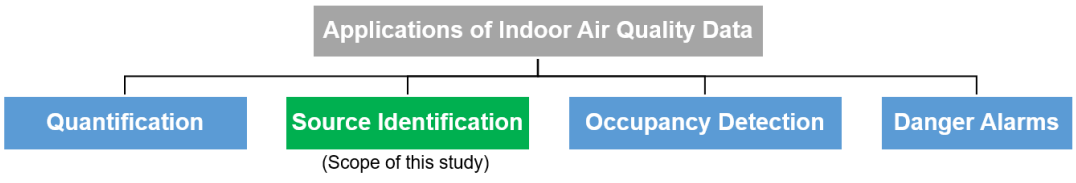


Figure 2.1: Overview of popular IAQ sensing problems in literature

The amount of data generated does point towards potential Machine Learning problems, but such options are rarely implemented products of this class. There are few works in literature that have prototyped these concepts and problems of our interest have been categorized in Figure 2.1. Each of the categories are defined as follows:

Quantification: Air Quality measurement is typically done through multiple sensors. In applications such as HVAC control or smart phone app, the measurements need to be reduced to an Air Quality Index (AQI) or any other single metric scale. A typical reference table is shown in Figure 2.2. Using air quality guidelines [8] the multi-sensor data is quantified with recommended threshold limits. This is the most basic sensing problem, but here calibration and cross-correlation of sensors [9] needs to be studied to provide reliable readings.

Level Category		A		B		C		D	
description		Good		Moderate		Unhealthy		Very Unhealthy	
Values	I_{LO}	0		51		101		251	
	I_{HI}	50		100		250		400	
Concentration		BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}
VOC(µg/m³)		0	200	201	350	351	500	501	757
CO(ppm)		0	4.99	5	9.99	10	14.99	15	2000
PM(µg/m³)		0	30	31	90	91	140	141	750

Figure 2.2: A Comprehensive Real-Time Indoor Air-Quality Level Indicator [10]

Occupancy Detection: For smart ventilation and energy conservation, indoor CO_2 concentration has been utilized in various studies to count the number of occupants in the room [11] [12]

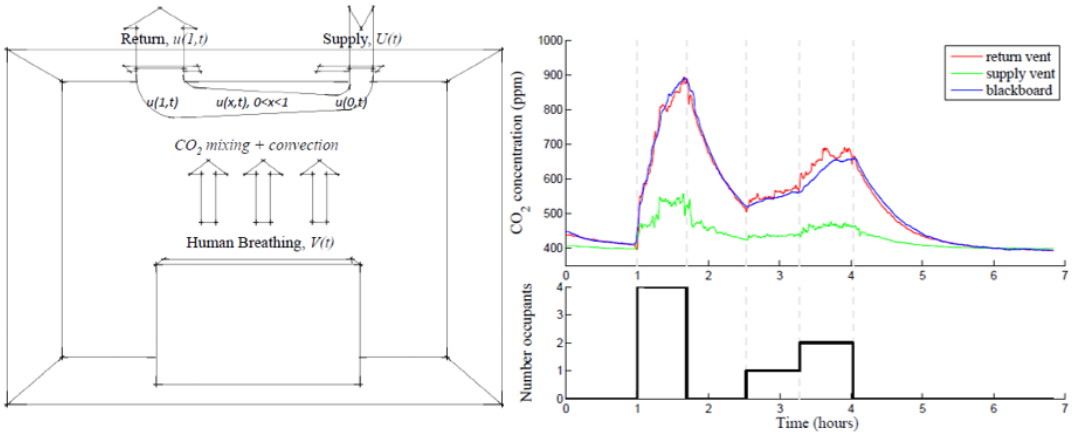


Figure 2.3: Sensing by Proxy: Occupancy Detection Based on Indoor CO2 Concentration **[13]**

Since human being's exhale Carbon-dioxide, crowded rooms can be a nuisance if not properly ventilated, at the same time, this measurement can be harnessed to inversely calculate the number of people. This is an alternative approach to motion - based sensors, especially useful for automatic air conditioning.

Danger Alarms: Fire and smoke detectors have started employing sensor fusion techniques for eliminated false alarms, and detecting type of fire (smoldering, flaming). One such study is [14] where they have compared various gas sensing principles and their performance to detect fires.

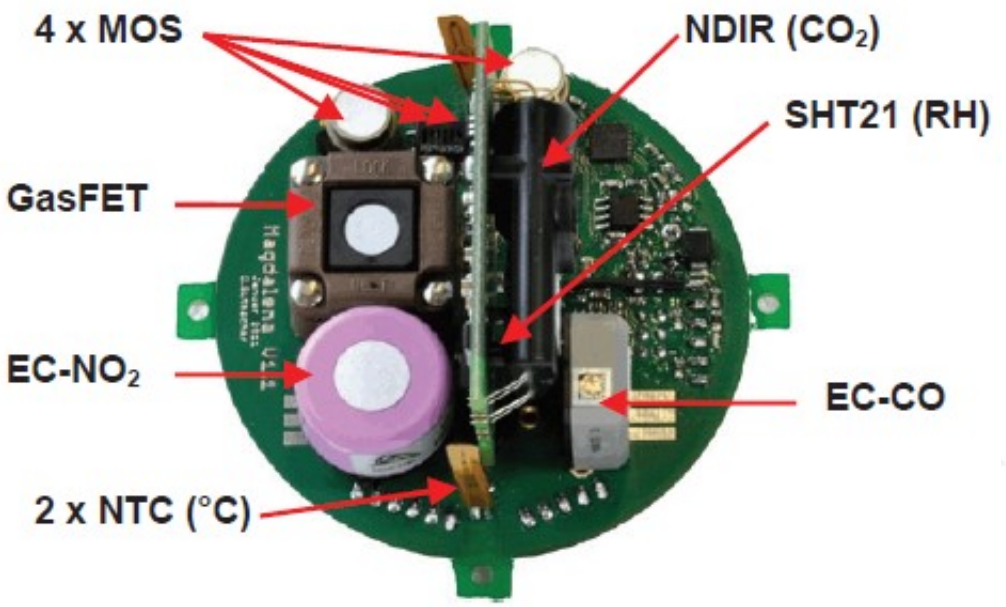


Figure 2.4: Comparison of gas sensor technologies for fire gas detection [14]

Source identification: Discussed in detail in the relevant subsection since it is the focus problem of this study.

2.2.1 Summary of relevant literature studied

The number of works in this area are still limited since most of gas sensing technologies not selective. As a summary, Table 2.2 was compiled to have an overview of how different studies have set their use-case, the sensors they have used to realize their work, and lastly the analysis / machine learning techniques.

Table 2.2: Related works by sensor combination and application

Ref	Year	Category	Summary of Contributions	Sensors										Analysis Techniques
				T	RH	PM2.5	PM10	VOC	CO	NO2	NH3	CO2	Other	
[15]	2015	Portable Context Detection	Observing personal exposure to PM2.5 on a wearable device as one performs daily activities such as walking, in vehicle, rest. Authors have shown raw data for different contexts, but no automatic recognition scheme is proposed.	✔	✔	✔							GPS, Accelerometer	None
[12]	2015	Occupancy Counting	Sensor data was labelled for Occupancy, and evaluated over two test sets, different combinations of variables for 4 different classifier algorithms.	✔	✔							✔	Light	LDA, CART, Random Forest
[13]	2015	Occupancy Counting	Constructed a 'Sensing by Proxy' model that uses known human CO2 emission rates, to construct Partial Differential Equations to predict the unknown number of people. The model is compared and proven better than a few different statistical methods, for controlled experiments.									✔		PDE-ODE
[10]	2016	Quantification	Construction of an IoT device that uses various sensors to quantify values to a standard Air Quality Index	✔	✔		✔	✔	✔					None

Table 2.2 (continued): Related works by sensor combination and application

Ref	Year	Category	Summary of Contributions	T	RH	PM2.5	PM10	VOC	CO	NO2	NH3	CO2	Other	Analysis Techniques
[16]	2012	Quantification	IoT Device, Air Quality Index and visualization toolkit. A rule-based algorithm to control ventilation, HVAC on demand and send alarms for hazardous levels of AQI.	✔	✔	✔	✔	✔	✔					None
[17]	2018	Source Identification	GasNet: Convolutional Neural Networks for Gas Classification. 8 MOS sensors were used to sample various gases in a controlled environment. Better result than SVM and MLP but much larger training time.					✔	✔				Total 8 different Figaro MOS sensors, for air contaminants, combustible gases etc.	Convolutional Neural Networks
[18]	2016	Source Identification	Air Sense - Indoor Air Quality Analytics. Detection of Cooking, Pesticide, and Smoking, and forecast of pollution if no action is taken.	✔	✔	✔		✔						Standard Deviation, Window Statistics,
[19]	2017	Source Identification	Tests conducted by leaving pollutants (Cigarette, Cleaning Agent, Fragrance, Food) on different days in test room. Window feature extraction methods were compared based on PCA variance explained. The best feature, Vector Array Normalization was fed into a Neural Network to classify the source of pollutant.	✔	✔		✔	✔	✔	✔			MOX CO2, O3,	PCA, LDA, KNN

Table 2.2 (continued): Related works by sensor combination and application

Authors	Year	Category	Summary of Contributions	T	RH	PM2.5	PM10	VOC	CO	NO2	NH3	CO2	Other	Analysis Techniques
[20]	2018	Source Identification	'Theodor'. The sensors were placed inside a box, and foodstuff sample was left inside. 24 different items were placed (Milk, Wine, Coffee, Garlic etc.) Data was split into non-overlapping frames, and 3 features extracted: Mean, Variance, Linear Regression. A 10-fold cross validation scheme was used	✔	✔			✔	✔	✔	✔	✔		J48 Decision tree
			Proposed Sensor Combination	✔	✔	✔	✔	✔	✔	✔	✔	✔		

2.2.2 Pollutant Source Identification using IAQ Sensor Arrays

2.2.2.1 The concept of Electronic Nose

The problem of classifying pollutant source from gas sensor data is very similar to the biological sense of olfaction. The mammalian nose contains many odor receptors which are often redundant (Figure 2.5).

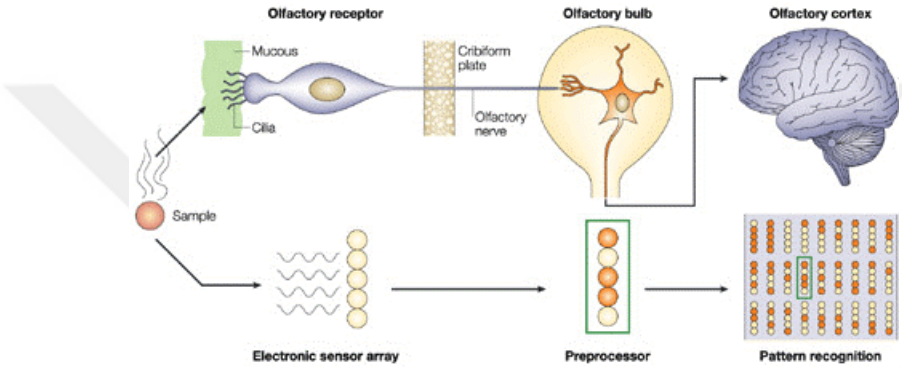


Figure 2.5: The analogy between biological nose and electronic nose [21]

The combination of several hundred such receptors is recognized by the brain to classify an odor as ‘pleasant’ or ‘bad’ or much more complex labels.

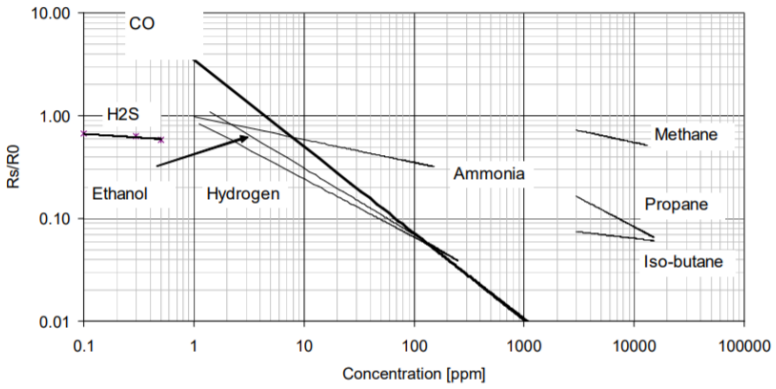


Figure 2.6: Sensor Response of MICS-6814 MOS type sensor [22]

MOS (Metal Oxide Semiconductor) type gas sensors are affordable and long-term stable as compared to electrochemical technology. However, selectivity of these sensors is limited and as shown in Figure 2.6 the response from one of the channels of the MICS-6814 is giving readout to multiple gases in the same readout ranges. For

studies similar to electronic noses, gas sensor arrays with each sensor having a different response is measured. The responses are then trained by exposing the sensors the target sources.

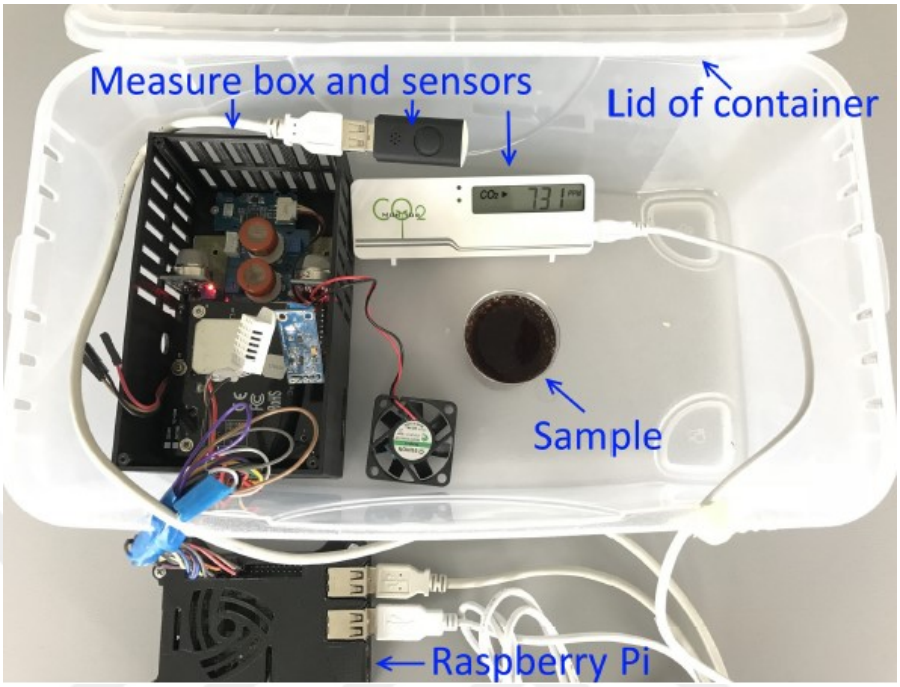


Figure 2.7: Theodore, a foodstuff odor classification study [20]

The study referenced in Figure 2.7 is able to classify 14 classes of different foodstuffs with good accuracy. The sensor combination is similar to ours with the exception of additional 4 MOS sensors for Methane, Alcohol, Propane, Carbon Monoxide respectively. The key difference with our work is that we are concerned in a much larger volume (a living room for example). Still, studying electronic noses is critical to our study since the problem is closely related

2.2.2.2 Pollutant Source Identification in living spaces

Pollutant sources in indoor environments are often linked to human activities which we intend to test within the scope of this study. There is still a dearth of source detection studies when the context is limited to living spaces. One such study is referenced in Figure 2.8 where four pollutant sources common in household air were exposed to a sensor hub measuring 9 different air quality parameters.

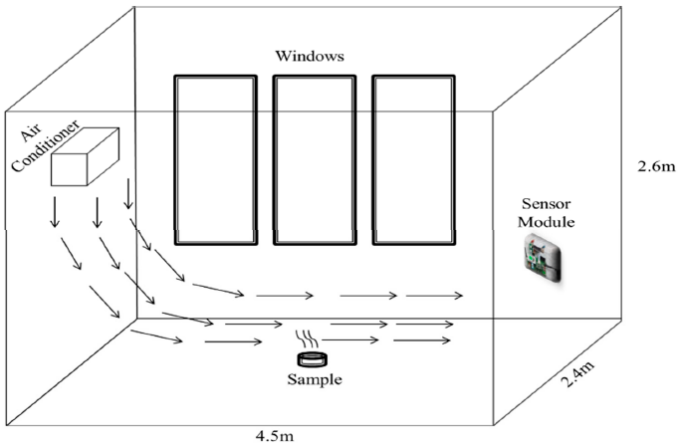


Figure 2.8: Classification of Smoking, Cleaning Chemicals, Air Freshener, Spoilt food in a room [19]

In another study called AirSense [18], only four parameters (Temperature, Humidity, VOC, Particulate matter) were used to classify events labelled as Cooking, Smoking and Pesticide to a high accuracy. The feature extraction and data analysis techniques employed in these studies is discussed in the succeeding section.

2.2.3 Data Analysis Techniques used in related work

2.2.3.1 Feature Engineering based techniques

Feature extraction in the preprocessing phase consists of rolling window operations on the original sensor response. These are done to remove the baseline effects and trends in the data.

Table 2.3: Popular feature extraction methods for MOS electronic noses [23]

Model	Description
Difference	$x_{ij} = V_{ij}^{max} - V_{ij}^{min}$
Relative difference	$x_{ij} = V_{ij}^{max} / V_{ij}^{min}$
Fractional difference	$x_{ij} = \left(V_{ij}^{max} - V_{ij}^{min}\right) / V_{ij}^{min}$
Logarithm difference	$\log\left(V_{ij}^{max} / V_{ij}^{min}\right)$
Sensor normalization	$x_{ij} = x_{ij} / \left(x_i^{max} - x_i^{min}\right)$
Array normalization	$x_{ij} = x_{ij} / \sum_i x_{ij}^2$

The difference methods in [23] are known to eliminate such baseline errors (Table 2.3). Relative and fractional difference methods work on the principal of linearizing the response and log difference to linearize the relationship when the original relationship is highly non-linear. Normalization models limit the sensor output

between 0 and 1. In cases where this method is used for multiple sensors, it ensures that all sensors have equal importance in later cognitive stages. The normalization methods would be really useful when the concentration is not of interest, rather the pattern of the response. In our proposed solution, Array Normalization is used as one of the features. In [19] these features are fed to PCA dimensionality reduction and then to various classifiers.

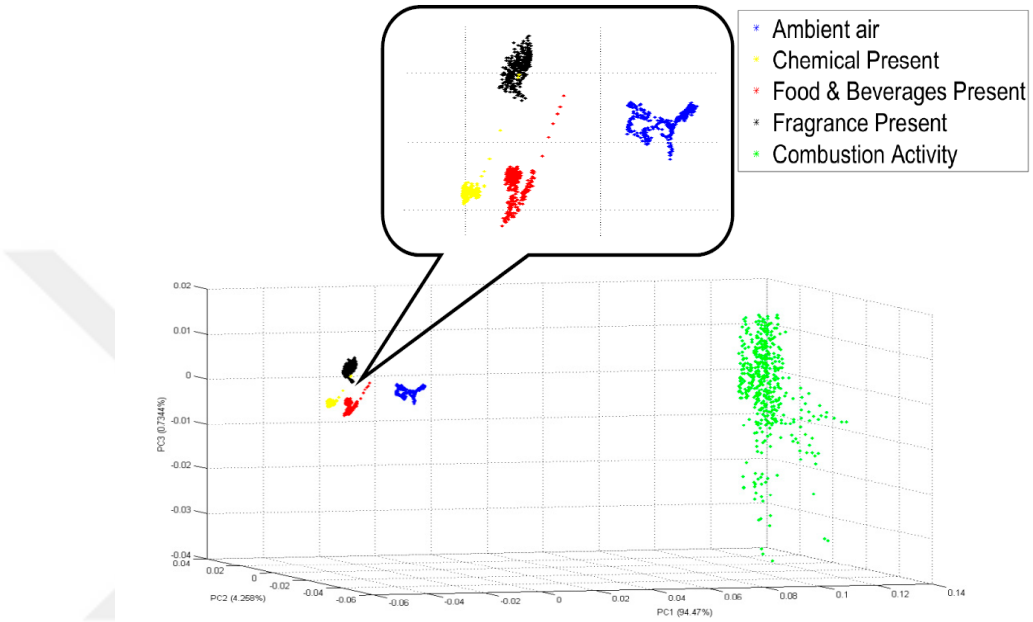


Figure 2.9: PCA Plot of Array Normalization features [19]

One study in particular [18] has devised a Normalized Standard Deviation based technique for binary classification. The classes in this study were smoking, pesticide, and cooking. Due to the similarity of this to our classes, this metric was also evaluated in our study.

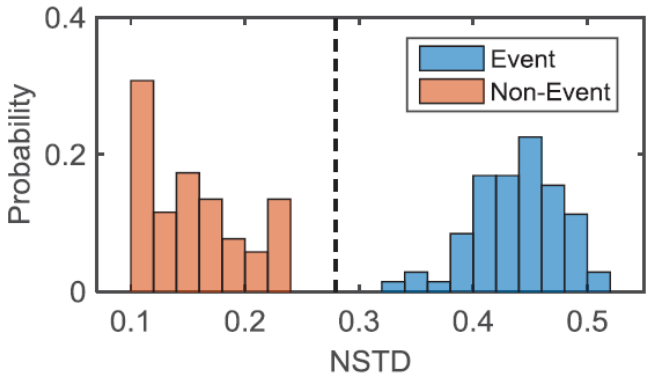


Figure 2.10: Binary classification by threshold of 0.3 on the Normalized Standard Deviation metric [18]

2.2.3.2 Neural Network based techniques

The use of Convolution Neural Networks (CNN) has revolutionized many Machine Learning problems in recent years. In the study GasNet, as similar approach is used. The biggest advantage of CNN is that they are highly skillful at eliminating the need for feature extraction due to the non-linearities modelled by their architecture.

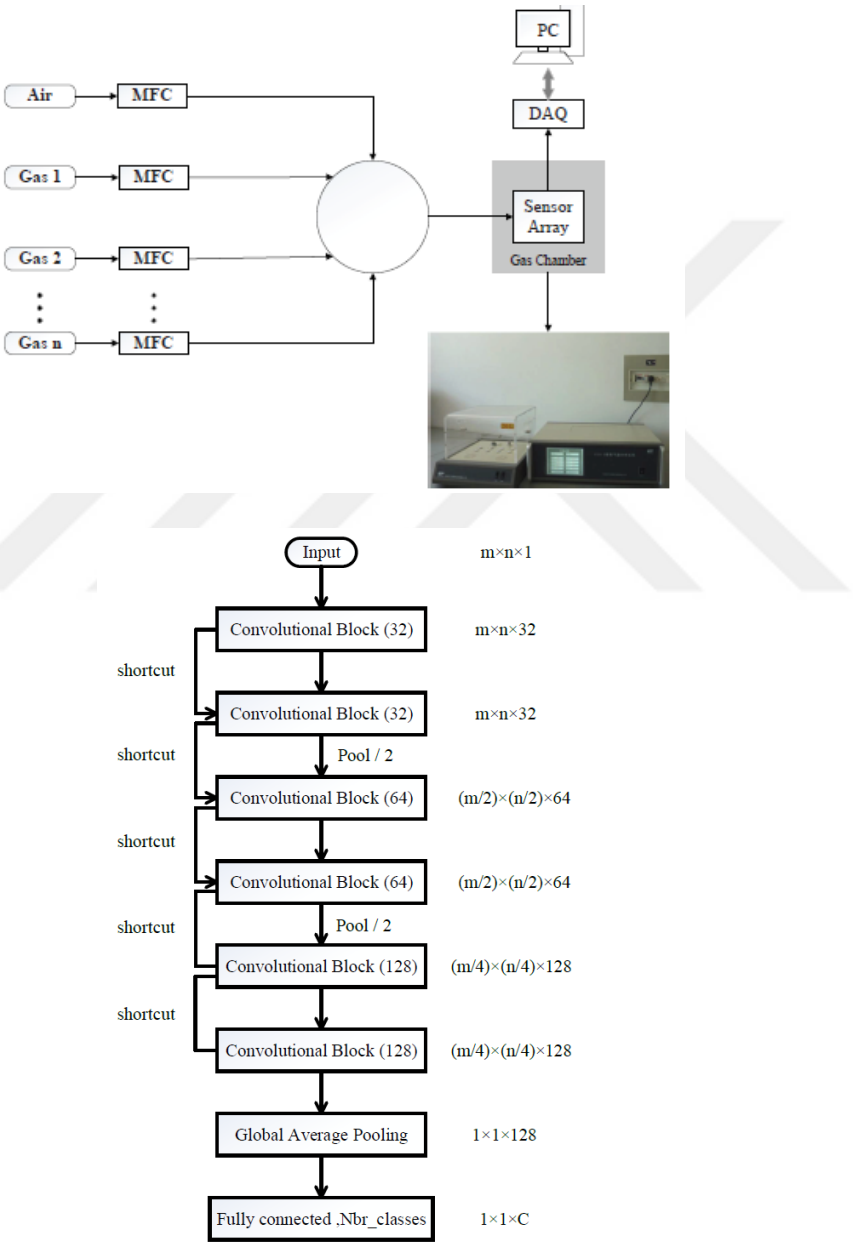


Figure 2.11: GasNet to classify gas samples with MOS array [17]

3. THE SENSOR PLATFORM

Based on literature review, and the important parameters influencing health and comfort. Table 3.1 shows sensors that were a part of our sensor platform.

Table 3.1: Specifications of sensors used for the sensor platform

Parameter	**Unit**	**Component**	**Range**	**Operating Principle**	**Interface**
Temperature	°C	SVM-30	-20 to 85	MEMS	I^2C
Humidity	%	SVM-30	0 to 100	MEMS	I^2C
Volatile Organic Compounds (VOC)	ppb	SVM-30	0 to 60k	Metal Oxide Semiconductor	I^2C
Carbon Dioxide	ppm	CM 1107	0 to 5k	Non-dispersive infrared	UART
Carbon Monoxide (CO)	ppm	MICS-6814	1 to 1000	Metal Oxide Semiconductor	I^2C
Nitrogen Dioxide (NO_2)	ppm	MICS-6814	0.05 to 10	Metal Oxide Semiconductor	I^2C
Ammonia (NH_3)	ppm	MICS-6814	1 to 500	Metal Oxide Semiconductor	I^2C
Particulate Matter <2.5 µm (PM2.5)	$\mu g/m^3$	SM-PWM-01C	0 to 3000	IR LED	PWM
Particulate Matter <10 µm (PM10)	$\mu g/m^3$	SM-PWM-01C	0 to 3000	IR LED	PWM

3.1 Sensor Descriptions

3.1.1 Sensirion SVM 30 (T/RH/VOC)

Gas sensors that measure Volatile Organic Compounds (VOCs) are sensitive to temperature and humidity [24] therefore Sensirion has developed an integrated module that contains a temperature/humidity sensor on the same board (Figure 3.1).

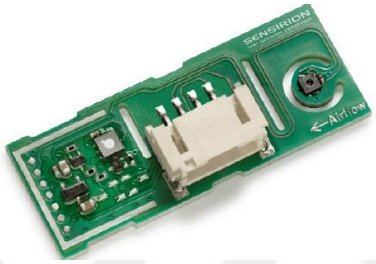


Figure 3.1: Sensirion SVM 30 sensor module with two separate sensing elements

The board also has memory to store calibration information as sensor resistances to calculate output values. By sending a message of 0x201e to the I^2C interface we can set the current conditions as baseline. This procedure was performed only once at the beginning of the project. In normal operation the get_baseline command of 0x2015 is sent to take the stored parameters into effect.

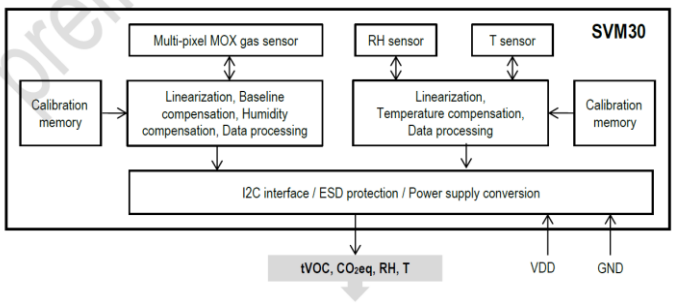


Figure 3.2: Internal block diagram of SVM 30 [25]

Other than for the MOX sensors compensation algorithms, the relative humidity and temperature data is useful for us as our basic parameters of air quality measurement. The I^2C interface uses two pins (one for data and one for clock signal) for communications and data is obtained by sending hexadecimal commands similar to the baseline operations above and the response is then converted to a numeric value according to the datasheet.

CO_2eq output: The sensor also gives a calculated value for the CO_2 concentration and its use for evaluated initially. The output is only reliable when there are no other pollutants present in the environment smells present in the environment greatly effect this reading. Therefore, an NDIR sensor was preferred for reliable CO_2 measurement.

3.1.2 SGX Sensortech MICS 6814

In general, MOS sensors use manipulate the electrical properties of metal oxides in response to a chemical stimulus. The MICS 6814 can detect CO, NO2 and NH3 based on the oxidative or reductive characteristics of these gases.

Figure 3.3: The MICS-6814 3-pixel MOS sensor [26]

The MICS 6814 contains three different sensing elements, each with a zinc oxide layer whose resistance changes in the presence of the target gas [26]. Thus, when CO or other VOCs, with their reductive behavior, remove the oxygen content from the metal oxide layer, its resistance decreases, and this change signals the presence of these gases. Similarly, when oxidizing gases such as NO_2 and NH_3 interact with the metal oxide layer they increase the overall resistance of the layer by adding insulative oxygen species in the system.

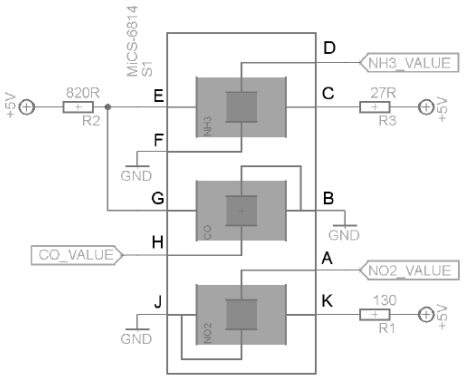


Figure 3.4: Measurement Circuit for MICS-6814 Sensor [26]

The biggest issue with using this sensor is cross-sensitivity, though the main target gases are CO (reduction channel), NO_2 (oxidation channel) and Ammonia (NH_3

channel), the sensor shows response to Ethanol, Hydrogen, H2S and other gases as shown in Figure 3.5. Since we are interested in detecting higher level air pollution events rather than measuring chemical concentrations, the differences in the three elements are still useful for pattern detection [20].

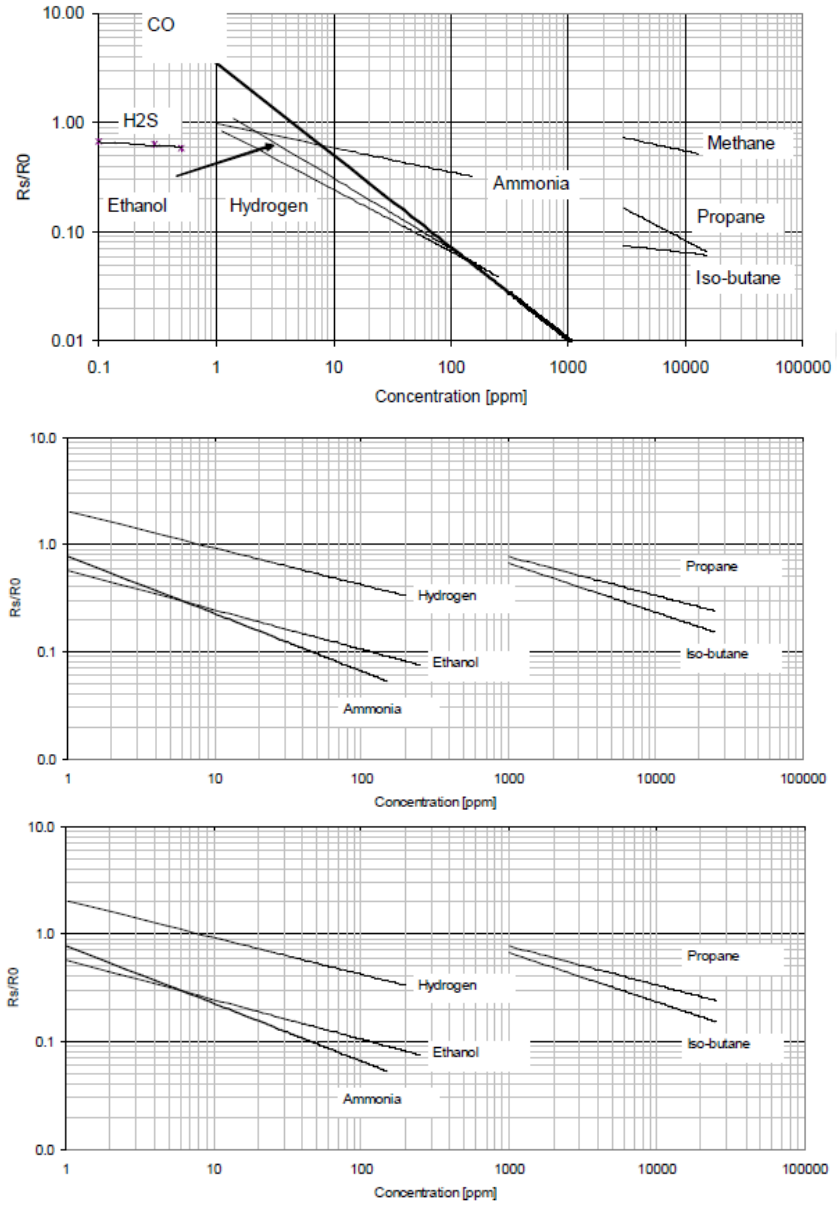


Figure 3.5: Sensor response of each sensing element in MICS 6814. Reduction channel (top), Oxidation channel (middle), NH3 channel (bottom) [26]

The datasheet claims high accuracy for various gases but selectivity is not mentioned in the technical specifications. The limitations of other MOS type sensors are also valid for this sensor. However the MICS-6814 is a popular choice for electronic nose applications as it packs three sensing elements in one integrated circuit.

3.1.3 Wuhan Cubic CM1106 (CO_2)

The CM1107 is a gas sensor module which can uses Non-Dispersive Infrared (NDIR) technology to detect CO_2 concentration in air. Polyatomic gas molecules have absorptive wavelengths in the infrared region, hence when they come in the path of infrared waves, the intensity of the incident wave decreases proportional to the concentration of the gas.

Figure 3.6: Wuhan Cubic CM1107 NDIR CO_2 sensor [27]

The CM1107 has two infrared light beams, one for detection and the other as reference, which give the module a high CO_2 detection accuracy. The module makes use of this absorption spectrum of CO_2 to detect the amount of CO_2 molecules in air. NDIR sensors do not have the selectivity problems and are a reliable gas sensing method.

Autocalibration: As of April 2018, the outdoor concentration of CO_2 is 400 parts per million (ppm), and the concentration indoors does not fall below this level. This fact is used for the autocalibration algorithm embedded in the CM1107, that every 7 days of continuous operation, the lowest signal is considered 400 ppm. To avoid calibration shifts in our tests, this feature was turned off by sending the appropriate command through the sensors UART interface.

3.1.4 Amphenol SM-PWM-01C (Particulate Matter)

The concentration of dust particles suspended in air can be detected through optical sensing. The method uses an infrared LED and a photo sensor to emit infrared waves and detect the reflected rays, respectively

Figure 3.7: Amphenol SM-PWM-01C Particulate Matter sensor [28]

Both small dust particles (<2.5 μm) such as cigarette ash and dry organic particles and large particles (<10 μm) can be detected by the photosensor and differentiated in the output. The output pulse narrows or widens according to the size of the particles; hence the varying output pulse enables the system to distinguish between the type and size of dust.

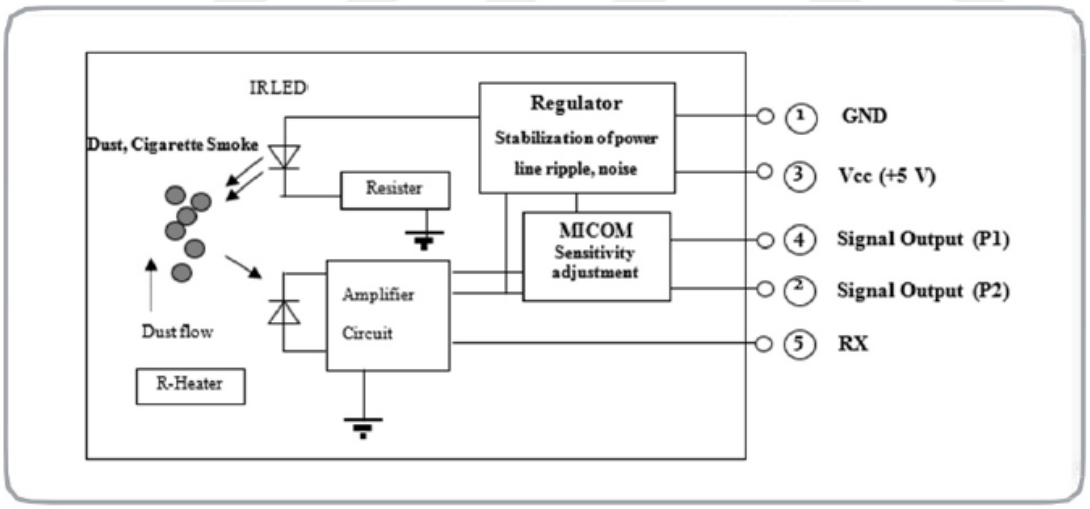


Figure 3.8: Internal block diagram of SM-PWM-01C

The output pins P1 and P2 (Figure 3.8) give the final processed output for large and small particles respectively.

3.2 Prototype Design

3.2.1 Electronic Design

To be do the measurements and to publish the obtained data to the database, the Linkit One prototyping board was used as our central development platform. The board has an ARM7 chipset and has 802.11 b/g/n WIFI transceiver.

Figure 3.9: Linkit One IoT prototyping board [29]

The board has an Arduino bootloader, and can be programmed using the Arduino IDE. There was a need to maintain parallel tasks for example our dust sensor relies on decoding Pulse Width Modulation (PWM) signals, for this reason additional Microcontrollers (Arduino Nano) were utilized. Due to the number of connections, a PCB shield was prepared to be mounted on top of the Link-it One board and which had connectors to all sensors.

The GPS and Bluetooth features of the board were not utilized for the current application, but Bluetooth will be useful for future works. Many smart-home devices use Bluetooth, and our device can possibly be used to send triggers to them. A possible scenario maybe adjusting the settings of a connected air-conditioner, for instance.

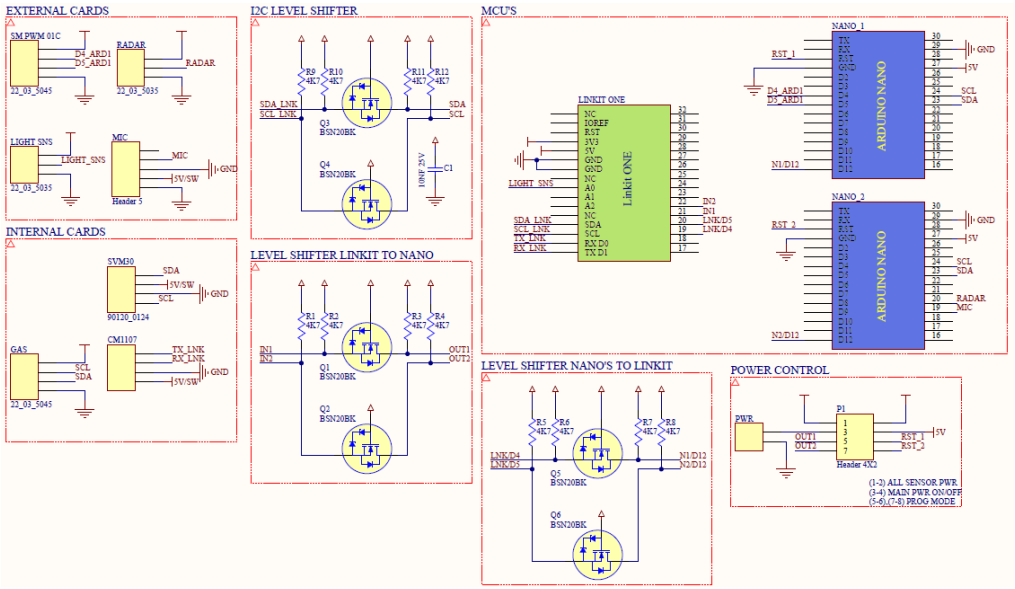


Figure 3.10: Schematic of the sensor shield (Light sensor, Microphone and Radar were not used in this study)

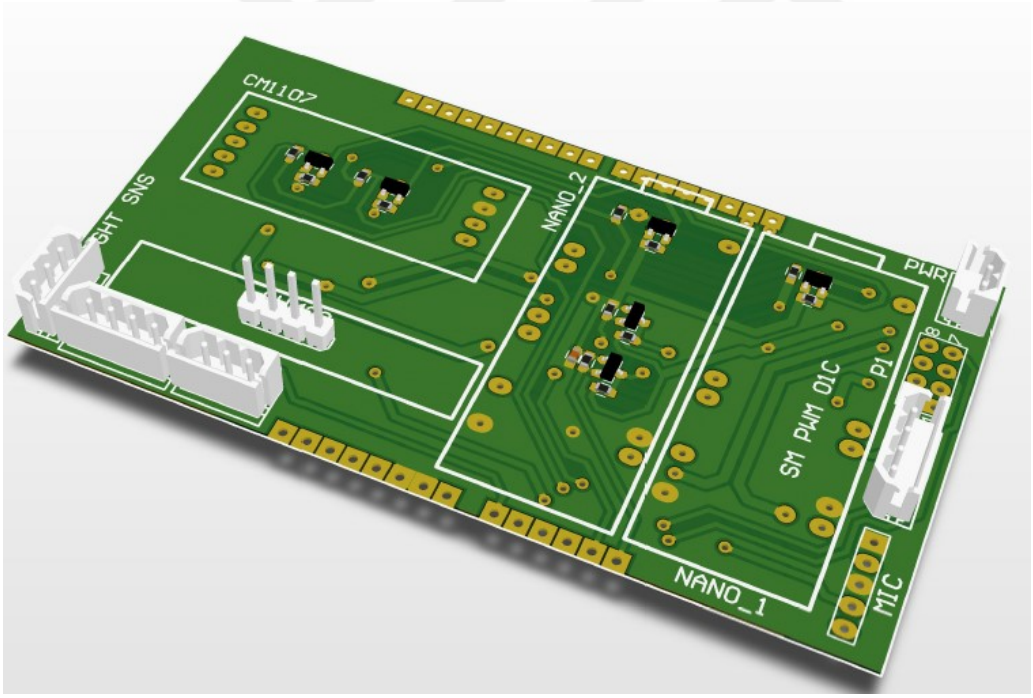


Figure 3.11: 3D view of the sensor shield PCB

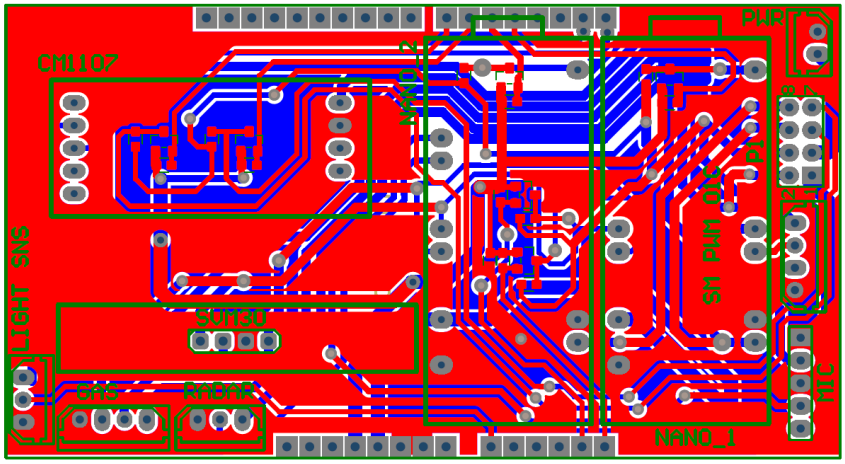


Figure 3.12: Dual layer PCB layout for the sensor shield

The circuit was designed to be able to handle multiple communication protocols are required by out hardware. The sensors were using I^2C, UART, and PWM protocols and the MCUs were using I^2C for their communication with each other. Only a single I^2C bus was maintained, and the sharing was made possible by a I^2C busy pin which was pulled to HIGH when a node was using the bus. Other nodes would wait until the bus was available. Also, we were using boards with multiple logic levels, (3.3V and 5V), therefore all communication buses had two versions on the board maintained by a level shifter circuit shown in Figure 3.10: Schematic of the sensor shield (Light sensor, Microphone and Radar were not used in this study)

It should be noted that this circuit could also had connections for a microphone, a low-cost radar and light sensor, with measurements programs integrated. The measurements were not required for the current study, so their integration is not detailed.

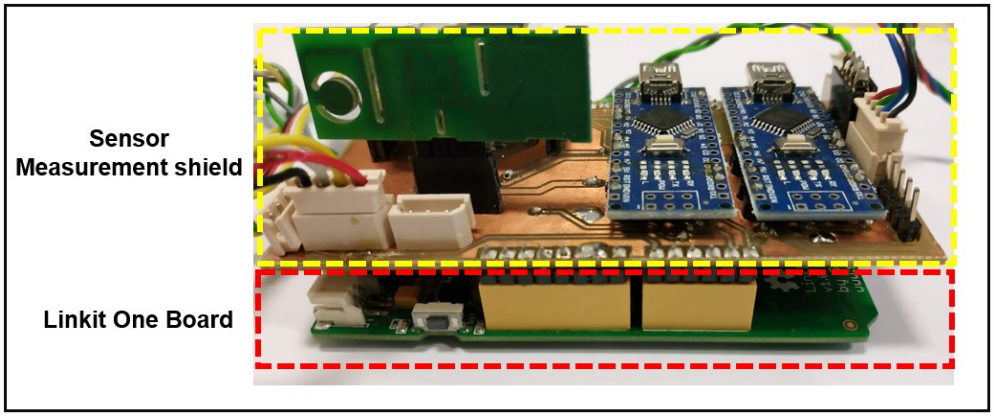


Figure 3.13: Linkit One Board with custom Sensor Shield

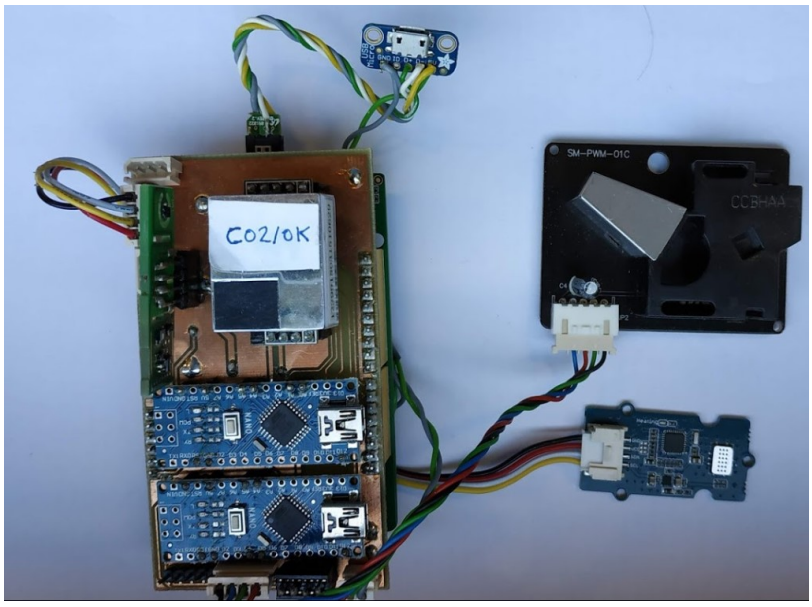


Figure 3.14: Complete assembly of the sensor platform

3.2.2 Protective Casing

To protect our sensors from harsh environmental conditions and handling between tests, a protective casing was designed in Siemens NX CAD software.

Figure 3.15: 3D drawing of the protective housing (top), electronics housed inside a 3D printed protype (bottom)

Care was taken to fit all the components in a way they had access to air-flow in the sensor manifold. The temperature/humidity sensor was specially placed at the bottom of the casing so that the heaters of gas sensors do not inflate the temperature reading in a way that they differ from the room-temperature.

3.2.3 Data Acquisition

Once data was measured by the electronic system, it was needed to publish this data to a MySQL server. In IoT systems, Message Queuing Telemetry Transport (MQTT) is the current standard for topic-based communication. The server has a MQTT broker running which is responsible for routing data when message a device publishes or subscribes to a 'topic'.


```
{
  "function":"setData",
  "id":"Indoor",
  "co":"0.20",
  "no2":"0.80",
  "nh3":"2.85",
  "co2":"585",
  "temperature":"28.63",
  "humidity":"44.07",
  "voc":"74",
  "pm25":"5.69",
  "pm10":"19.79"
}
```


Figure 3.16: MQTT payload sent to the Air Quality topic

Another script subscribes to the 'Air Quality' topic and executes and INSERT query to a MySQL table. The table is already prepared to receive the specified values, and each row corresponds to data captured at each timestep. The timestamp is recorded in the UNIX format (number of seconds since midnight January 1, 1970) to avoid allocating separate fields for the hour minutes dates etc.

The second layer in Figure 3.17 is of Database Administration, for this phpMyAdmin framework is used. This allows us to visualize the tables, and conveniently execute 'SELECT ' queries to extract data in CSV format for the time periods where the tests are performed.

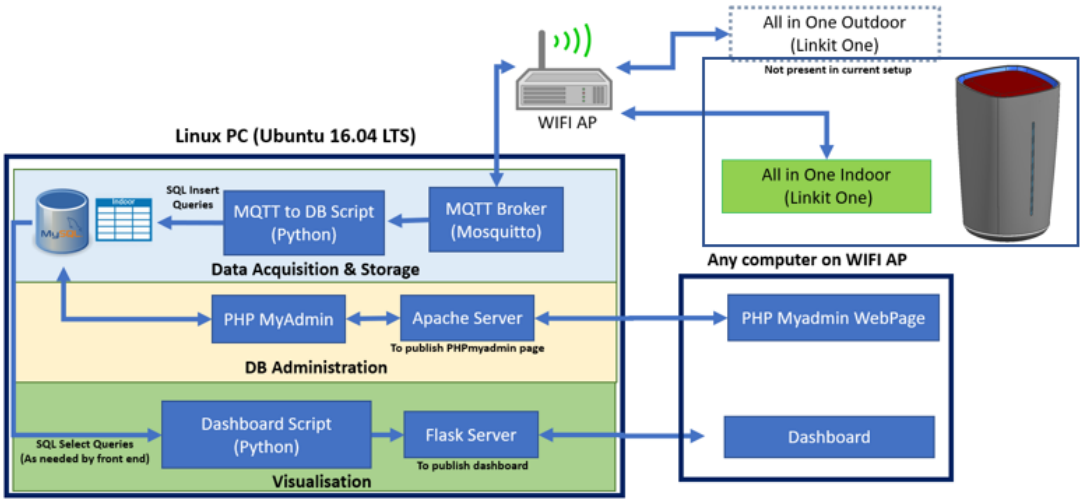


Figure 3.17: Data Acquisition Architecture

Finally, the Visualization layer was used to host a dashboard webpage on the server where all the measurements could be viewed as the experiments were taking place.

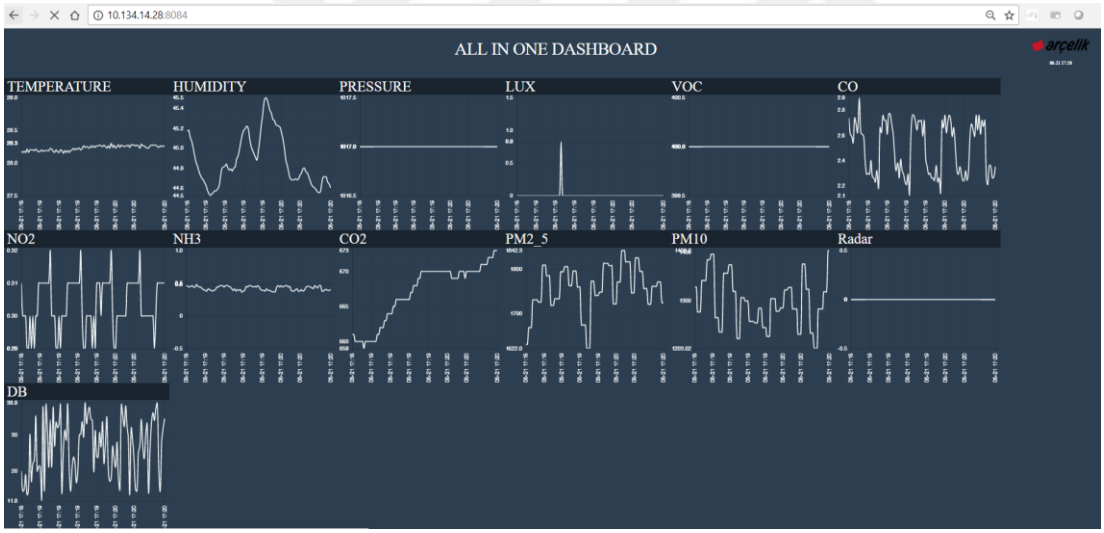


Figure 3.18: Dashboard for data visualization

To make sure that the communication is continuous, throughout the development stages the system was tested at a sample period of 500 milliseconds, even though the sample period can be as a high as 60 seconds. This is due to the fact that air diffuses in the air slowly and the response times of sensors are also nearer to those ranges. To make sure transients and sensor errors are noticed, the sample period for our tests was set to 6 seconds.

4. DATA ANALYSIS

This section covers the details of how popular techniques in the field of machine learning are applied to the sensor data obtained from our Air Quality device. We are particularly interested in detecting the signature of certain household activities, and the feasibility for conducting further research in the area.

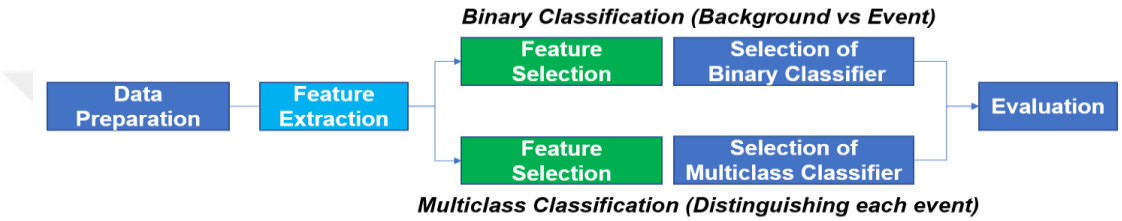


Figure 4.1: Analysis Workflow

The scope of the problem is similar to [18] with our classes being background, cleaning, cooking, and smoking. As shown in Figure 4.1 we have also evaluated the case of binary classification for use cases where we are only interested in the deviations from background. This can be useful because in a real-life scenario, an event may not fall into the finite number of classes the multiclass model is trained with. There, just detecting 'some' pollution event can be a sufficient hypothesis to generate control rules for the smart-home.

4.1 Experiment Design

To be able to develop our algorithm we required a test setup and a plan where we could test our hypotheses repeatedly. Tests were conducted in the form of 'test units' where each activity target activity was performed (Figure 4.2). The test room was ventilated in between tests. Note that gas sensor readings often take hours to return to background levels.

Start	*Ventilation*	*Cleaning*	*Ventilation*	*Cooking*	*Ventilation*	*Smoking*	*Ventilation*	*Test*

Figure 4.2 Sequence of 1 test unit

The ventilation step between tests varied and was often as short as 30 minutes, so there were often lasting effects of the last activity. In order to not make the experiment too controlled as done in [20] [19] these durations were not strictly controlled.

4.1.1 Test Setup

Figure 4.3 shows the layout of the test room. The room had one air inlet and the door was used as the outlet. A fan was used to create the air flow in between experiments. The sensor hub was placed as one would place the product in a kitchen counter top. Since we want to study the feasibility of this device in smart home applications, we deliberately did not optimize the layout in a way to give maximum sensor reading.

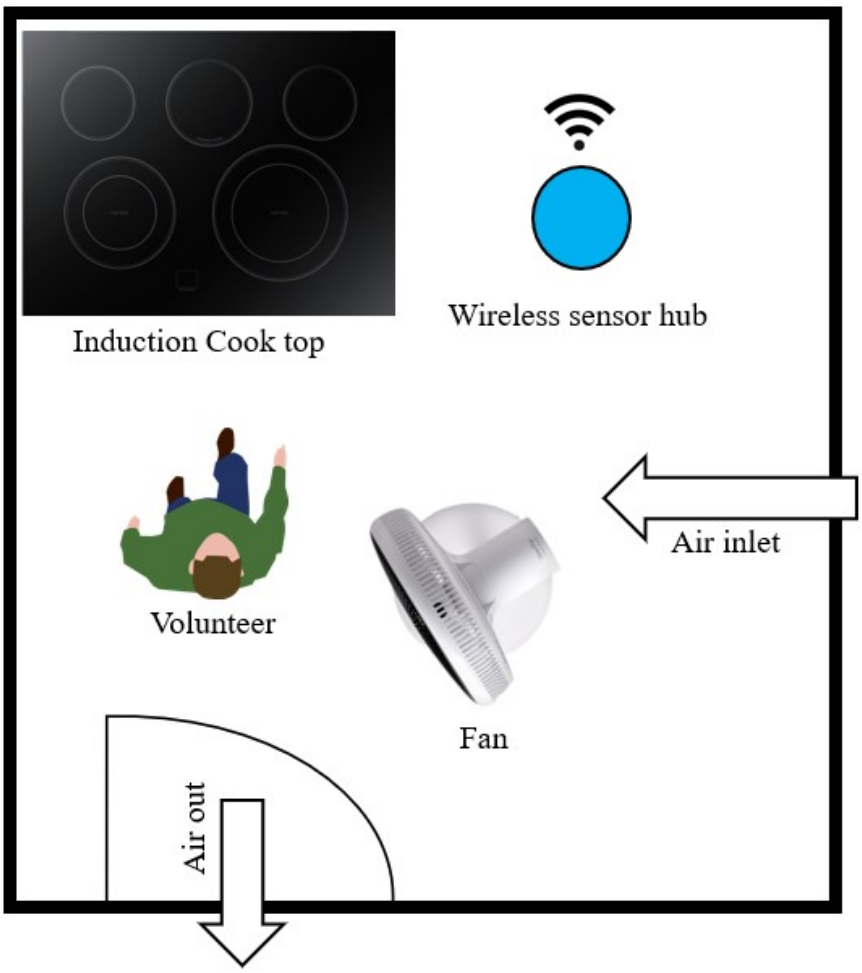


Figure 4.3: Layout of test room

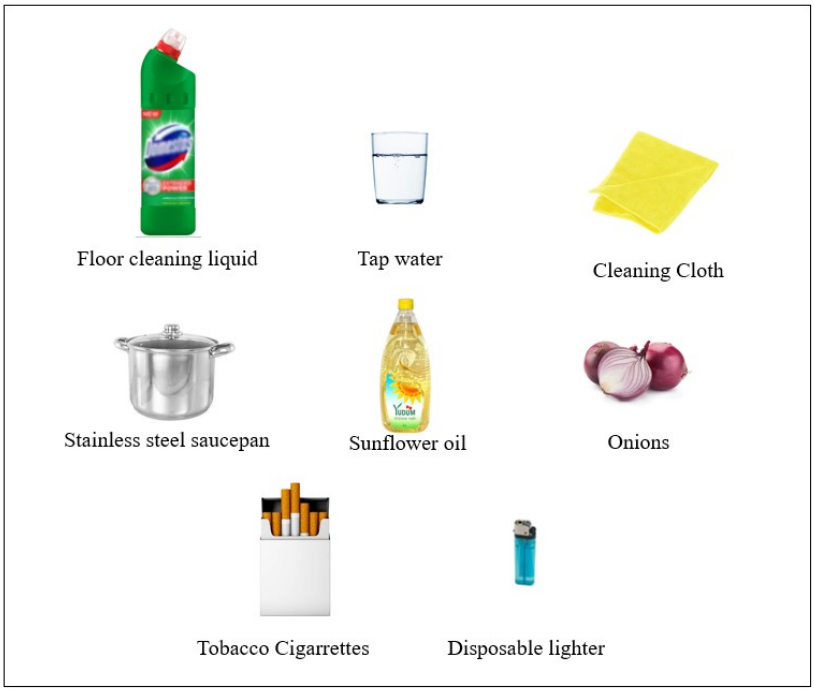


Figure 4.4: Materials used for simulating activities

4.1.2 Preparing data for Supervised Learning

There were a total of 22 tests conducted, in each test unit each of the 3 activities was performed. The duration of the activity was variable, and volunteers were only given rough guidelines on how to conduct the experiment. Some key observations from the raw data are listed as follows.

1. As seen in figure 4.5, the parameters return to background levels after 10 minutes of ventilation.
2. The NO_2 channel shows peaks during the cleaning activity.
3. Cooking causes increase in nearly all the parameters, mainly PM2.5 and humidity.
4. Smoking also effects PM levels, but the rise is very fast and diminishes quickly after.

These observations were seen repeatedly for all tests and it pointed to the fact that pattern recognition techniques seen in the literature can be applied to detect these signatures automatically.

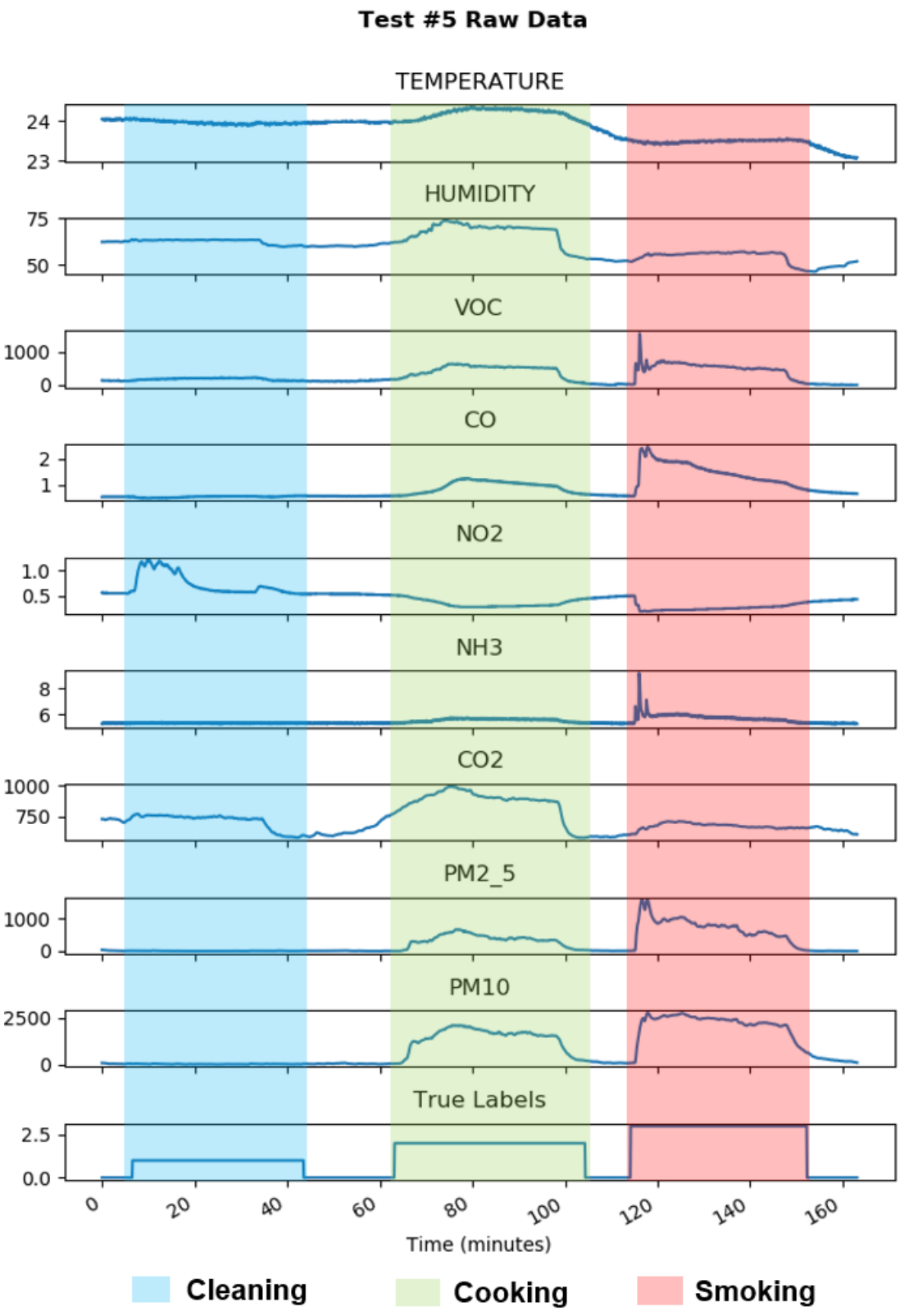


Figure 4.5: Raw Data for a single test unit

To prepare this as a supervised learning problem the data was split in a manner shown in Figure 4.5. The data was shuffled and split into training, and 2 test sets.

Test 1 was utilized in the algorithm development stage, where Test 2 was only tested at the end. Throughout this section however, test 2 performance is shown for each development step.

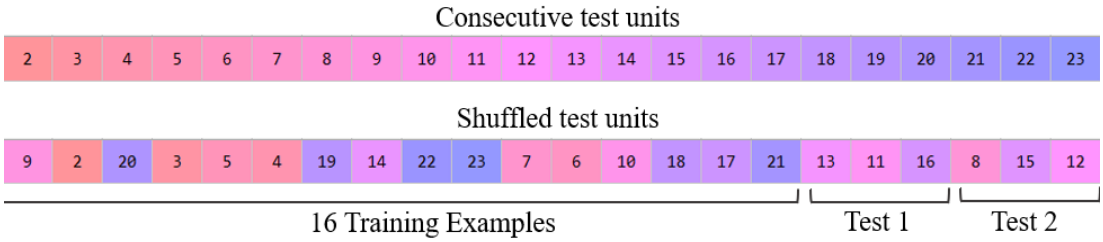


Figure 4.6: Shuffling and splitting of test units

Note that shuffling and splitting was done with respect to separate test units, and not samples picked randomly from the entire dataset. There were 10 additional sets of background data obtained from same days as the class experiments. Since it is time series data, it is important to preserve the order of points. The succeeding sections will provide details of how sliding window feature extraction methods were used to assist our algorithm.

Table 4.1: Length of data set in time units

	Background	**Cleaning**	**Cooking**	**Smoking**	**All**
Training Set (Hrs)	**22.3**	**9.0**	**12.7**	**10.8**	**54.9**
Test Set 1 (Hrs)	**9.8**	**1.0**	**2.2**	**1.3**	**14.3**
Test Set 2 (Hrs)	**7.2**	**0.8**	**2.7**	**1.9**	**12.6**
Total Hours	**39.3**	**10.8**	**17.6**	**14.0**	**81.8**
Sample Rate	**6 seconds**				

A total of 81.8 hours of air quality was collected and most of it was used for training as shown in Table 4.1. Most of the captured data was background since it is easy to capture and also enables the models to have a good baseline.

4.1.3 Evaluation Metrics for Classifier Performance

$$\textbf{Accuracy:}\frac{\textit{True Positive}}{\textit{Number of Samples}} = \frac{\textit{True Positive}}{\textit{True Positives+True Negatives+False Positives+False Negatives}} \quad (4.1)$$

As we can see in Table 4.1, we have imbalance in the number of samples we have for each class. A one-shot accuracy metric would be skewed towards a class with more support (number of samples), and our design decisions for the model will be driven by a false metric. Therefore, the use of Precision and Recall is a widely used technique in literature concerning classifiers [30]

$$\textbf{Recall:}\ \frac{\textit{True Positives}}{\textit{True Positives+False Negatives}} \quad (4.2)$$

Consider a binary classifier, recall is its ability to detect positive instances of a class. The problem arises when a classifier with perfect recall gives too much a false alarm.

$$\textbf{Precision:}\ \frac{\textit{True Positives}}{\textit{True Positives+False Positives}} \quad (4.3)$$

Intuitively precision is the ability of the classifier to reject false alarm. The precision would be 1.0 if there are no false positives. The problem here is that we do not account for the negatives, and this can lead to dangerous results based on what the classifier is used for

$$\textbf{F1-score: } 2 * \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision+Recall}} \quad (4.4)$$

The F1-score is a harmonic mean of the precision and recall that gives equal importance to both metrics. Other F-score techniques such as F-beta exist when there is separate importance given to each of the metrics. For the purpose of our study F1 score suffices.

Extension of F1-score to multiclass case:

F1-score for each class is calculated treating the target class as positive and all other as negative. The total F1-score is then calculated as an average of all F1-scores. Some studies add weights to the average calculation to account for varying class support. In our study we treat all the classes as equally important, so we use uniformly weighted F1-scores.

For the remainder of this study, uniformly weighted F1 Score will be used as the metric of choice to evaluate classifier performance and select appropriate algorithms.

4.2 Feature Extraction

As detailed in section 2, the problem poses many challenges including, sensor baseline drift, cross sensitivity of MOS sensors, and the varying properties of the activity in question. These variations include.

1. **Baseline concentrations of pollutants:** Due to influence from outdoor air quality and seasonal trends, background concentrations can vary from day to day.
2. **The length of response:** Each activity (cleaning, cooking, smoking) can have varying length of sensor response, depending on the distance of the activity on the sensor, the concentration of stimulant used and the way the activity is performed.
3. **The shape of the response:** For the same reasons as above, the shape of response can also show variation.
4. **Cross sensitivity:** As we can see in the correlation plot below, response in each channel influences other channels. This is because of cross-sensitivity of our sensors, and since all our sensors measure through air as the proxy, the response of our sensor is not orthogonal.

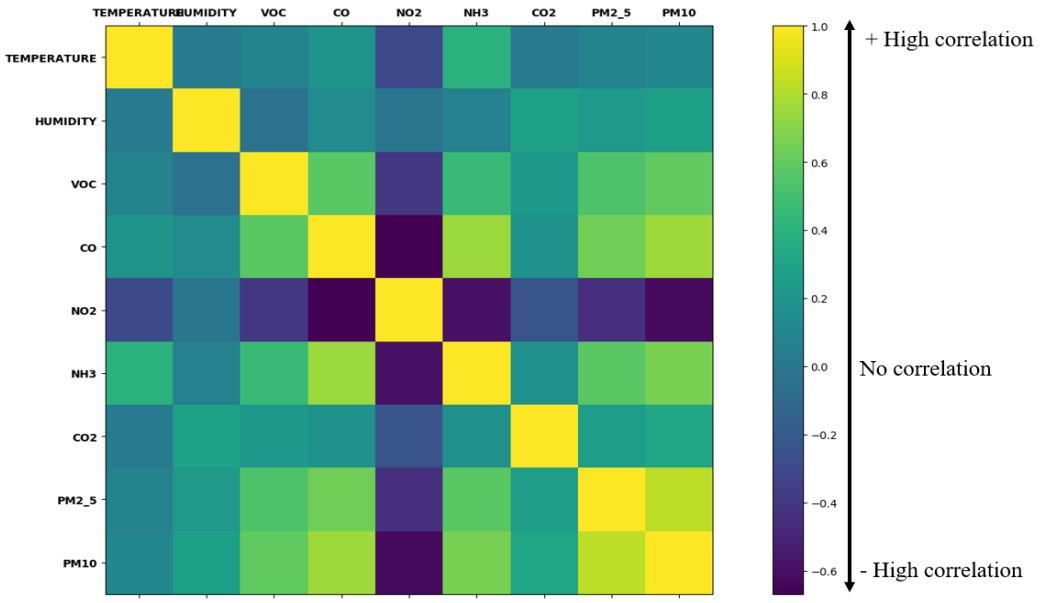


Figure 4.7: Correlation plot of training set

Among the techniques discussed in literature, we saw using a feature extraction-based approach over Convolutional Neural Networks and other techniques. The main reason for this is the amount of data we have this problem. CNNs are powerful, if

there is enough data, with 22 tests at our disposal, CNNs will likely fail to generalize for the unpredictable test sets that may come up.

4.2.1 Vector Array Normalization

$$VAN_i = \frac{X_i}{\sqrt{\sum_{j=i-w}^{i}(X_j)^2}} \tag{4.5}$$

Equation 1: Vector Array Normalization for measurement parameter X. *W* denotes the window size **[23]**

The Normalization techniques include Zero Mean, Standard Scaling, Min Max Scaling, and others. For time series data however, the Vector Array Normalization is used. This feature technique was used in [19] for pollutant source identification and [23] for odor classification in an electronic nose.

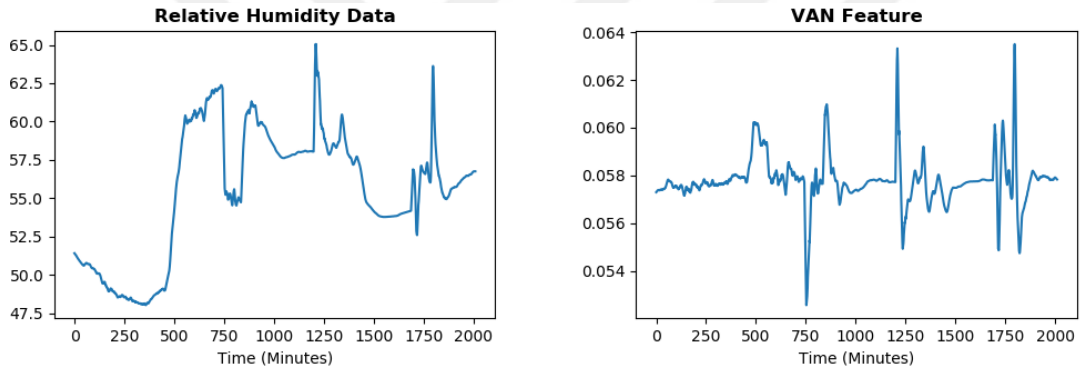


Figure 4.8: Example of VAN feature extraction with fixed window size of 30 minutes

As seen in Figure 4.8 the VAN feature is able to remove trends from the raw data and able to focus on sharp changes in the data (t=500, t=750), in case the change is slower (t=1300) the magnitude is much lower. It is also interesting to note that the direction of VAN from its mean value denotes whether there is a rise or a fall in the original data.

4.2.2 Normalized Moments

$$NSTD_i = \frac{1}{\max(W)}\sqrt{\frac{1}{w}\sum_{j=i-w}^{i}(X_j - \mu_W)^2} \tag{4.6}$$

The Normalized Standard Distribution (NSTD) metric is due to the work in [18] where it was used on PM2.5 data to distinguish background versus events by setting

a threshold based on NSTD distribution for each class. As we can see on our example data this metric is very good at detecting sharp changes in data. Normalizing by the maximum value of each window makes sure that the data is detrended.

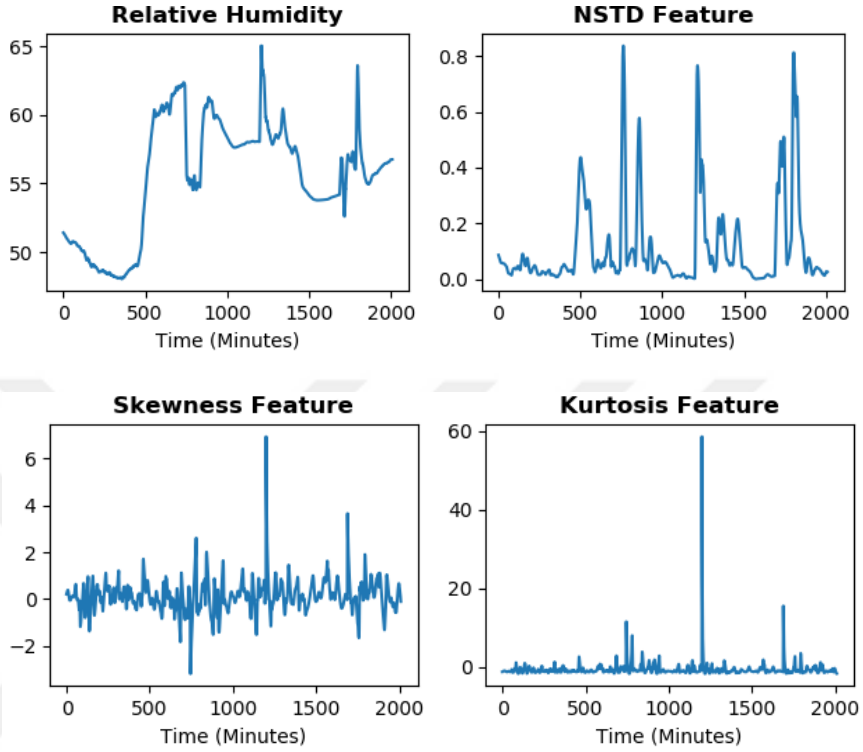


Figure 4.9: Example of rolling window moments with fixed window size of 30 minutes

Extension to third and fourth moment: This approach was extended to calculate the third and fourth moments which correspond to Skewness and Kurtosis behavior of data to add more features to our analysis.

$$NSKW_i = \frac{1}{\max(W)} \sum_{j=i-w}^{i} \frac{(X_j - \mu_W)^3}{\sigma^3}, \quad NKRT_i = \frac{1}{\max(W)} \sum_{j=i-w}^{i} \frac{(X_j - \mu_W)^3}{\sigma^4} \tag{4.7}$$

4.2.3 Estimated Hurst Exponent

The Hurst exponent is measure of Self-Similarity of a time varying series. The following approximation is due to [31]

$$\langle |\log(t+\tau) - \log(t)|^2 \rangle \sim \tau^{2H} \tag{4.8}$$

Table 4.2: Hurst exponent and how it describes time series

Value of Hurst Exponent	Implication
H=0.5	Series is Geometric Brownian motion
H>0.5	Series is trend reinforcing
H<0.5	Series is mean reverting

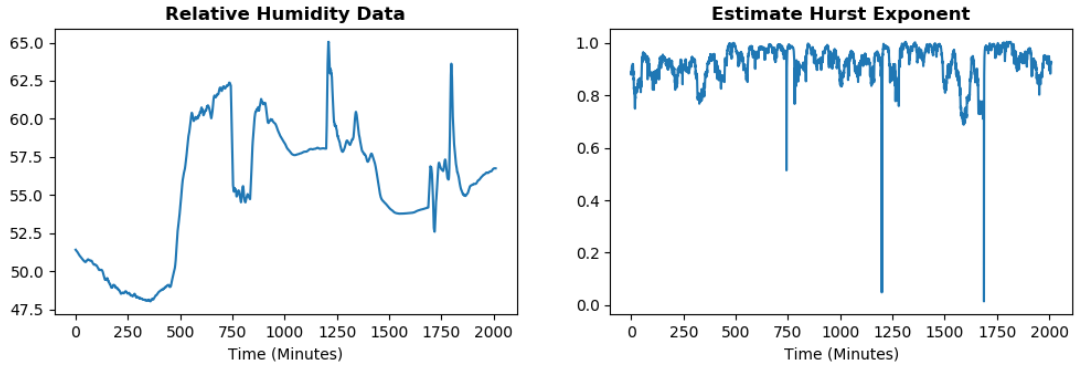


Figure 4.10: Example of rolling window Hurst Exponent with fixed window size of 30 minutes

The current example shows that the Hurst exponent drops below 0.5 at t=1200 and t=1700. If we expect the raw data at those points, we see the series is returning to its rolling mean values. Similarly, the noisy section at t=750 is detected by H=0.55 which corresponds to random walk.

Using the hurst exponent and other features, we expect to have a set of robust feature set to use in our classification algorithms.

4.3 Classification

Once we have extracted our features, we need to evaluate them for classification of the target events. A total of 45 features extracted from the 9 primary sensor measurement outputs were used to evaluate classification algorithms (Table 4.3). Note that no cross-sensor features were calculated, and each feature is a window operation on individual sensor arrays as detailed in the previous section.

Table 4.3: Summary multivariate feature matrix

Measured Parameter	Extracted features
Temperature (°C)	5
Relative Humidity (%)	5
Volatile Organic Compounds (ppb)	5
CO_2 (ppm)	5
CO (ppm)	5
NO_2 (ppm)	5
NO_3 (ppm)	5
PM 2.5 ($\mu g/m^3$)	5
PM 10 ($\mu g/m^3$)	5
Total Number of columns in univariate feature matrix	**45**

The temperature features were ultimately not used in development of our algorithms. The reason being the air conditioner in the adjacent lab. Since our tests were conducted during work hours, temperature could be the only determining factor for our binary classification and deeming the analysis unfair. Therefore, temperature was eliminated from the algorithm development.

4.3.1 Binary Classification

The problem was prepared as a binary classification problem by treating background labels as 0 and saturating the other classes to 1. In this section we will first explore the behavior of our extracted features in the context of class separation by the use of distribution plots. Later combinations of classifiers and features will be evaluated on test data.

4.3.1.1 Feature Distributions for the binary case (2 classes)

All 5 features relating to temperature were discarded after seeing the following plot. The main reason is that none of our activities significantly effect temperature (even cooking causes slow change of 1 degree). Also, the mean value of both distributions overlap near the mean. Humidity data was kept as is as cleaning and cooking both generate humidity.

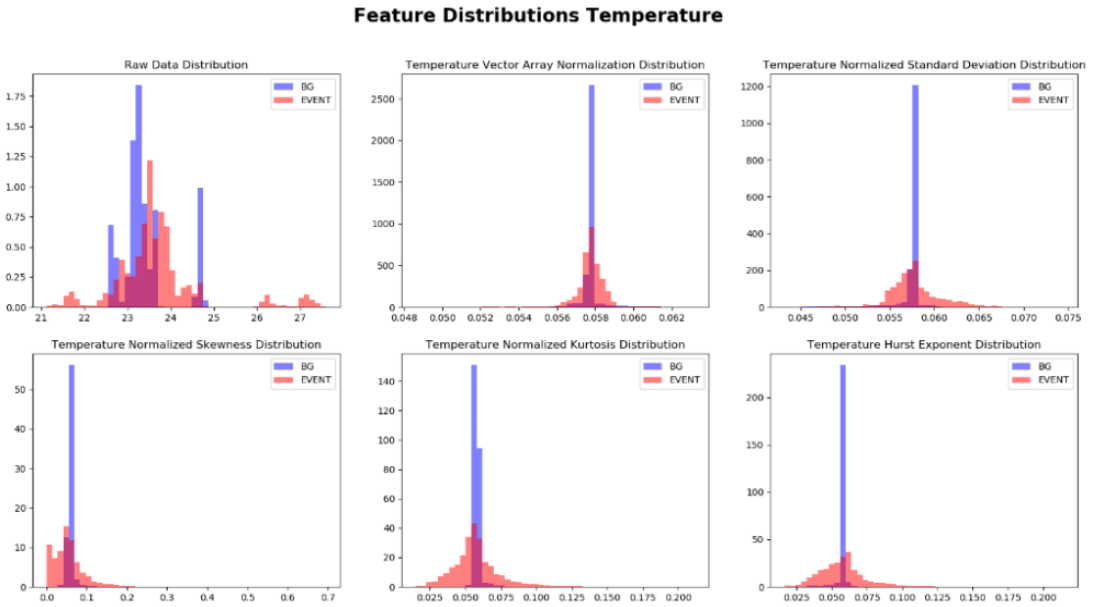


Figure 4.11: Class wise Temperature distribution for all features (Binary case)

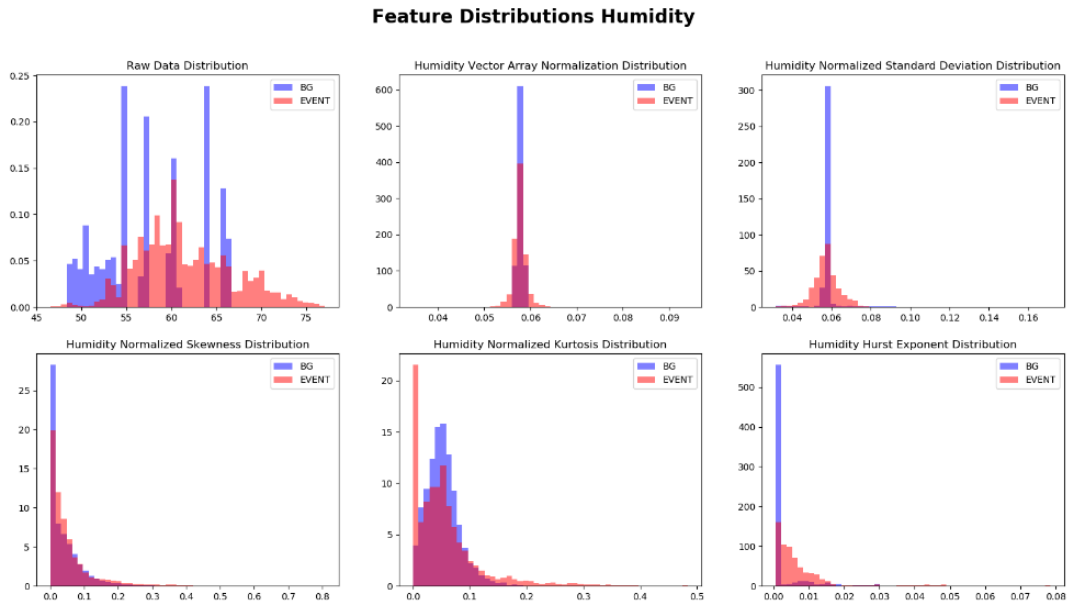


Figure 4.12: Class wise Humidity distribution for all features (Binary case)

The VOC features already show that the class separation is made apparent after feature extraction is performed. The PM10 also shows a similar behavior where class separation exists specially in the skewness feature.

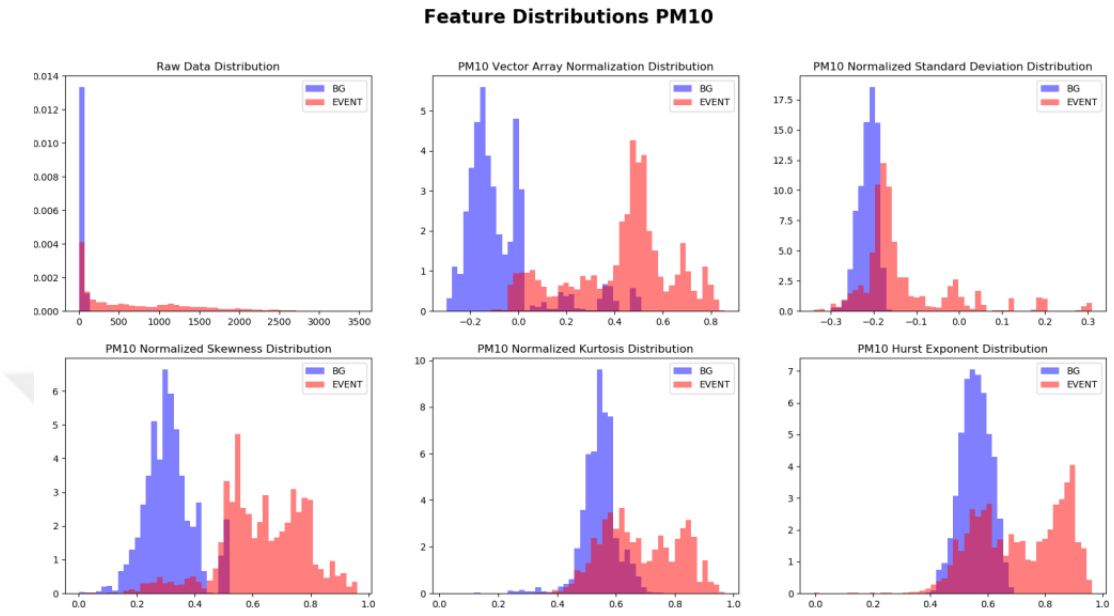


Figure 4.13: Class wise PM10 distribution for all features (Binary case)

As apparent from the distribution a thresholding technique implied in [15] may not performed since there is overlap in the distributions. Still, we can see that a feature set, may be able to perform well on the task.

4.3.1.2 Performance of various classifiers on each feature set

We put our features to the test by feeding them into four different classifiers. Decision Tree is a tree-based classifier which assigns thresholds using the Gini index [32]. K-Nearest Neighbors uses distance metrics from points in the training set. Linear Discriminant Analysis [33] is a projection-based technique that we have detailed in the multiclass section.

Table 4.2 shows the F-scores of the candidate classifiers when fitted with single feature sets. Each feature set contains 8 arrays corresponding to the unique sensor measurements excluding temperature and processed through the specified extraction technique.The combination with the best performance was the Normalized Skewness used with the Decision tree classifier.

Table 4.4: Comparison of classifier F-scores with single feature sets

	Decision Tree Classifier			K Nearest Neighbors (K=100)			Linear Discrimination Analysis			Gaussian Naive Bayes			
	Train	Test 1	Test 2	Train	Test 1	Test 2	Train	Test 1	Test 2	Train	Test 1	Test 2	
Raw	1.00	0.94	0.87	0.99	0.91	0.88	0.79	0.71	0.78	0.83	0.74	0.74	0.85
VAN	1.00	0.91	0.83	0.99	0.92	0.84	0.88	0.91	0.83	0.94	0.91	0.84	0.90
NSTD	1.00	0.92	0.91	0.98	0.92	0.84	0.88	0.89	0.73	0.94	0.90	0.77	0.89
NSKW	**1.00**	**0.94**	**0.90**	**0.99**	**0.94**	**0.90**	**0.94**	**0.93**	**0.86**	**0.95**	**0.95**	**0.86**	**0.93**
NKRT	1.00	0.91	0.87	0.99	0.90	0.87	0.92	0.88	0.87	0.91	0.90	0.77	0.90
Hurst	1.00	0.91	0.86	0.98	0.93	0.85	0.94	0.93	0.80	0.91	0.91	0.77	0.90
Avg.	**1.00**	**0.92**	**0.87**	0.99	0.92	0.86	0.89	0.87	0.81	0.91	0.89	0.79	

It was of our interest to evaluate if multiple feature sets could be used to make the scores better. Figure 4.14 shows a plot where all possible combinations (5! =120 in total) were evaluated, and the best combination selected at each combination step.

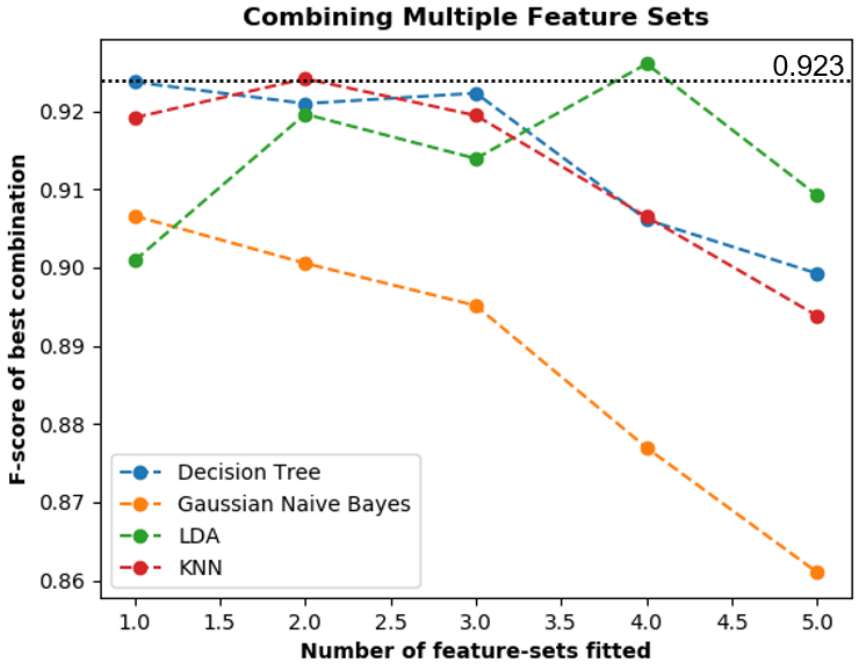


Figure 4.14: Performance of classifiers with multiple feature sets fitted

It was observed that all classifiers other than LDA had poorer performance as more and more feature sets were fitted. LDA only surpassed Decision tree when 4 feature sets were fitted and deteriorated at the next step. Therefore, we validated that decision tree with the NSKW feature set is the best choice. In the succeeding section

4.3.1.3 Decision Tree Classifier

The selected classifier is a decision tree classifier which has the following performance on test sets.

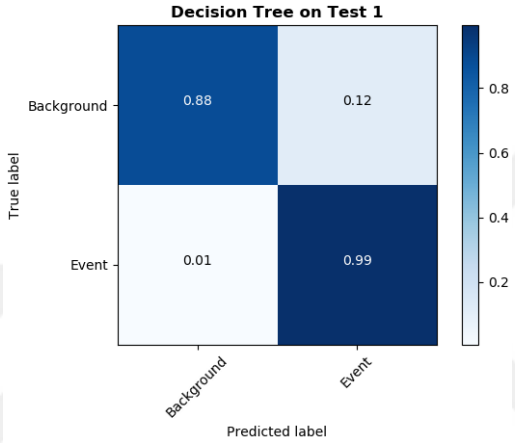


Figure 4.15: Decision Tree Classifier fitted on the Skewness Feature Set

Table 4.5: Performance of Decision Tree Classifier on Skewness Feature

Metric	Value
Training Accuracy	1.00
Cross validation Accuracy (10 fold)	0.95
Precision	0.95
Recall	0.94
F1 score	0.94

Figure 4.16 shows the resulting Decision Tree classifier on the skewness feature. This tree of heuristics can easily be deployed on a micro-controller and is very cheap in terms of computation costs.

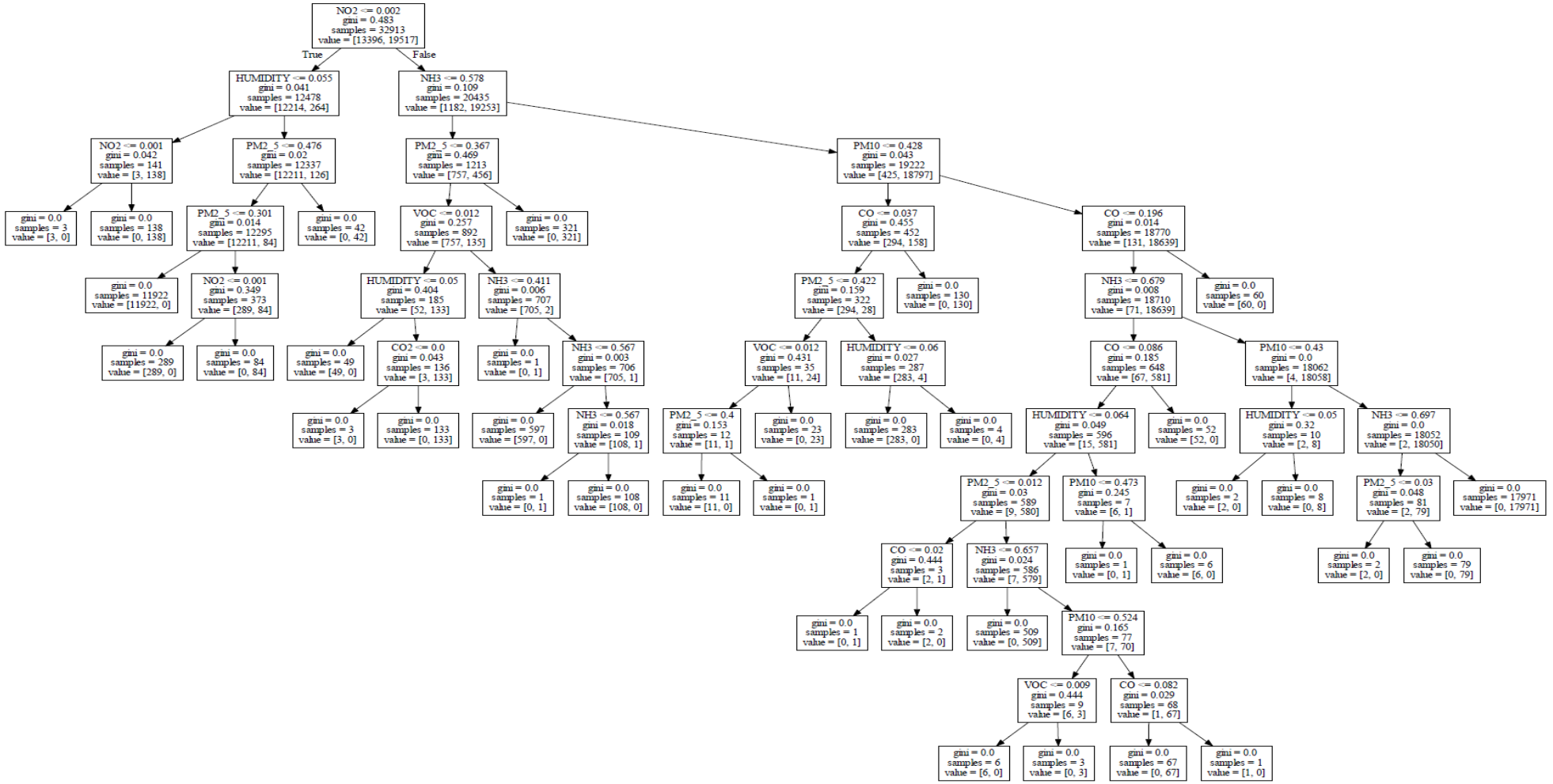


Figure 4.16: Decision Tree Classifier fitted on the Skewness feature set

4.3.2 Multiclass Classification

As discussed in the last section, the binary classification case was successfully accomplished with single feature sets. Plotting the univariate distributions of the two classes on page 37 made it obvious that a single feature set may be sufficient for classification. This section will show how this is not the case for Multiclass classification, and we will need to rely on the whole feature matrix defined in the feature extraction section.

4.3.2.1 Feature Distribution of 4 Classes

Plotting the distributions for all four classes shows that background has a different distribution than all the foreground classes, we see that no single feature is enough to classify all the classes.

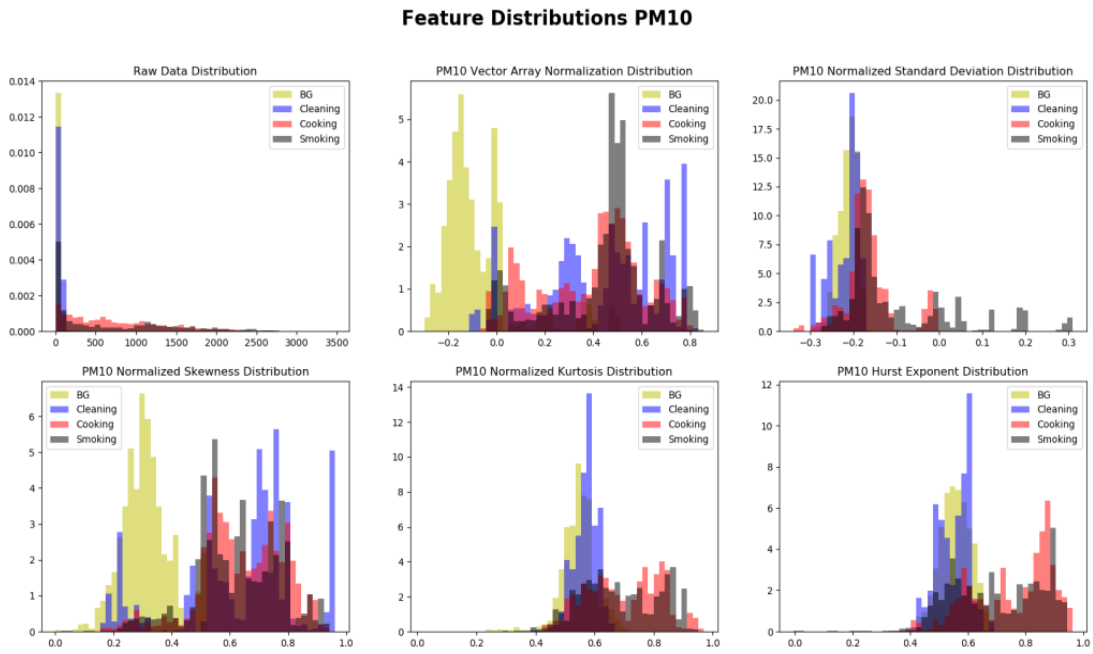


Figure 4.17: PM10 feature distribution for all four classes.

By running the classifier models on each feature set one by one, as done for the binary case Figure 4.17 shows how single features are inadequate.

Table 4.6: Weighted F-scores of multiple classifiers on each feature set

	Decision Tree Classifier			K Nearest Neighbors (K=100)			Gaussian Naïve Bayes			Linear Discriminant Analysis			
	Train	Test1	Test 2	Train	Test 1	Test 2	Train	Test 1	Test 2	Train	Test 1	Test 2	Avg.
Raw	1.00	0.65	0.50	0.91	0.64	0.59	0.62	0.56	0.44	0.78	0.65	0.49	0.65
VAN	1.00	0.69	0.65	0.89	0.70	0.59	0.66	0.60	0.45	0.68	0.74	0.54	0.68
NSTD	1.00	0.70	0.66	0.79	0.65	0.49	0.62	0.52	0.40	0.79	0.81	0.49	0.66
NSKW	1.00	0.70	0.68	0.92	0.72	0.65	0.65	0.52	0.39	0.73	0.66	0.50	0.68
NKRT	1.00	0.64	0.64	0.89	0.67	0.60	0.62	0.51	0.38	0.72	0.61	0.41	0.64
Hurst	1.00	0.67	0.63	0.90	0.63	0.50	0.60	0.50	0.36	0.79	0.72	0.57	0.66
Avg.	1.00	0.68	0.63	0.88	0.67	0.57	0.63	0.54	0.40	0.75	0.70	0.50	

4.3.2.2 Classification on Univariate Feature Matrix with Linear Discriminant Analysis (LDA)

LDA Methodology [34]: The Linear Discriminant Analysis is a projection/classification method that projects the n-dimensional feature space to a C dimensional space in order to simultaneously:

1. Increase the distance **between** class means (S_B)
2. Decrease the distance **within** each class (S_W)

$$S_B = \frac{1}{N}\sum_{i=1}^{C} N_i \left(\mu_i - \mu\right)\left(\mu_i - \mu\right)^T \tag{4.9}$$

$$S_W = \sum_{i=1}^{C} \sum_{\mathbf{x}_k \in C_i} \left(\mathbf{x}_k - \mu_i\right)\left(\mathbf{x}_k - \mu_i\right)^T \tag{4.10}$$

Where C is the total number of unique class labels. This optimization is done by obtaining a projection matrix W.

$$W = \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}, \tag{4.11}$$

This procedure is very similar to Principal Component Analysis (PCA), but the key difference is that LDA is a supervised method and maximizes the class separation in the projected space, whereas PCA has no knowledge of true labels, and works to project features to maximize the total variance [35]

LDA then generates a linear decision boundary, decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule.

Transforming our feature matrix of n=40 and 4 classes to a 3-dimensional space using LDA we see that there is clear separation of classes for both training and test sets.

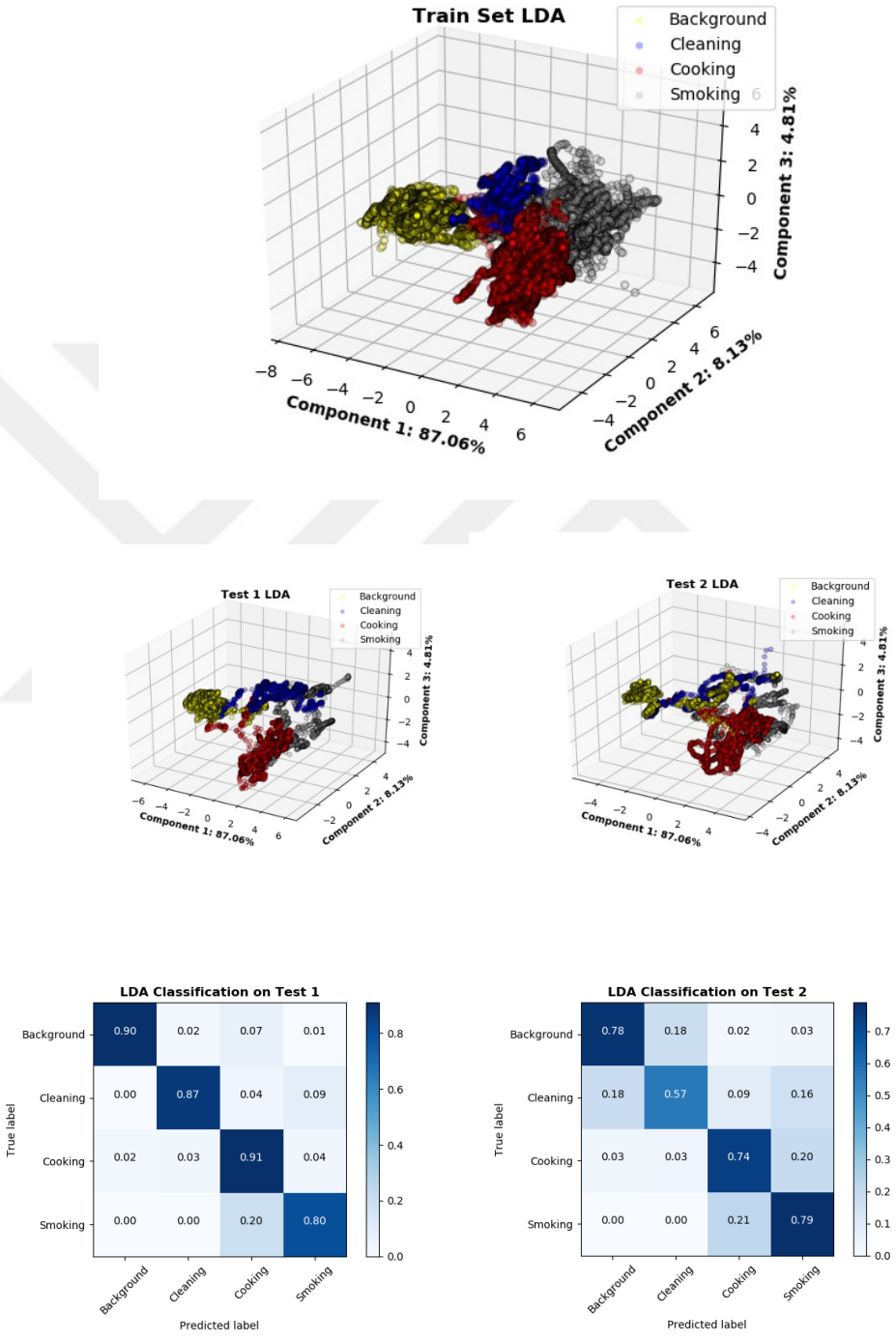


Figure 4.18: Performance of LDA classifier on all features. LDA Projections (Top), Confusion Matrix (Bottom)

Table 4.7: Fitting metrics for LDA classifier

Data Set	Accuracy
Training	0.91
Cross Validation (5 fold)	0.80
Test	0.81

Table 4.8: Class-wise evaluation of LDA classifier

Class	Precision	Recall	F1 Score
Background	92	84	88
Cleaning	73	71	72
Cooking	75	76	75
Smoking	76	79	77
Total	**79**	**80**	**79**

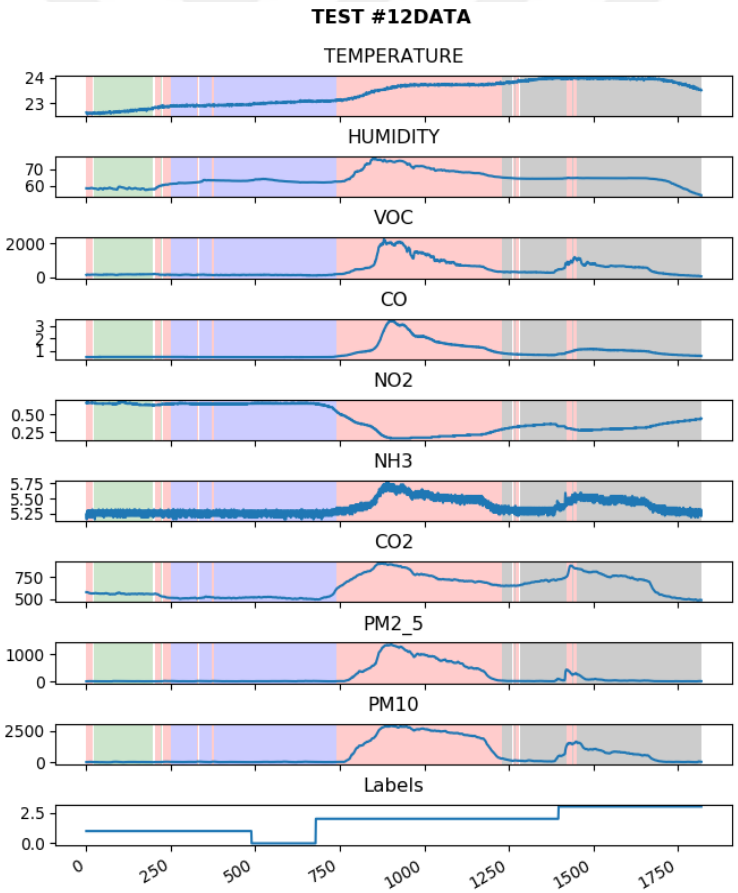


Figure 4.19:LDA Classifier Hypothesis on a unit from Test 1 (Blue=Background, Green=Cleaning, Red=Cooking, Black=Smoking)

Recursive Feature Elimination (RFE)

After obtaining results on the full feature matrix it was analyzed if some features can be eliminated to improve the results.

Algorithm 1: Recursive feature elimination

1.1 Tune/train the model on the training set using all predictors
1.2 Calculate model performance
1.3 Calculate variable importance or rankings
1.4 **for** *Each subset size* S_i, $i = 1 \ldots S$ **do**
1.5 Keep the S_i most important variables
1.6 [Optional] Pre–process the data
1.7 Tune/train the model on the training set using S_i predictors
1.8 Calculate model performance
1.9 [Optional] Recalculate the rankings for each predictor
1.10 **end**
1.11 Calculate the performance profile over the S_i
1.12 Determine the appropriate number of predictors
1.13 Use the model corresponding to the optimal S_i

Figure 4.20: Recursive Feature Elimination algorithm (RFE) [36]

RFE is a wrapper feature selection technique, which means that it works by recursively fitting the features to obtain the best combination of features [37]. The operation was conducted for previously discussed LDA, and Random Forest classifier for comparison.

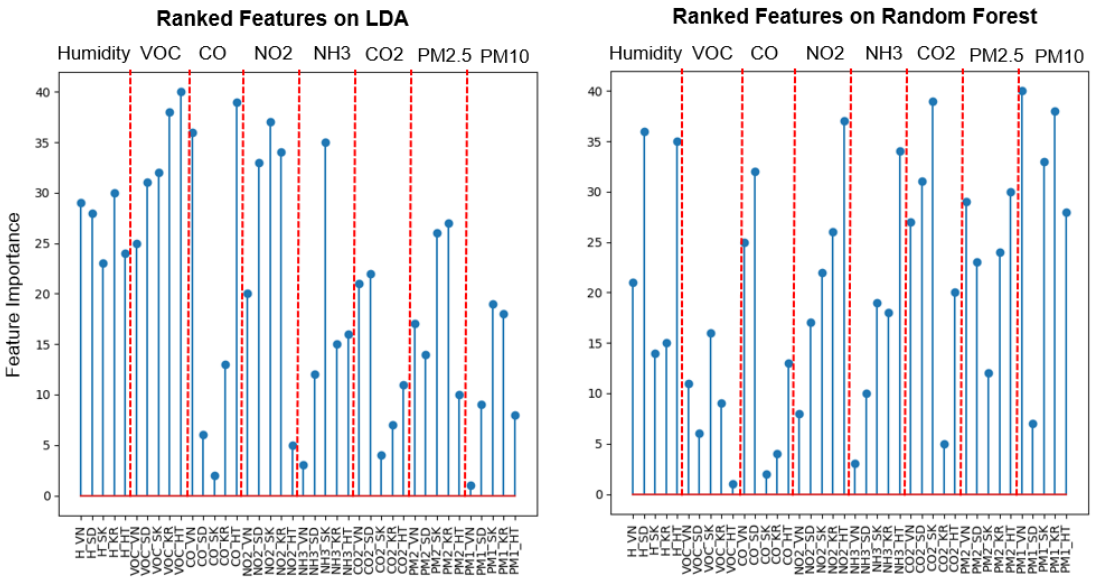


Figure 4.21: RFE feature importance for LDA and Random Forest Models

Both the classifiers give more importances to different features, where random forest is giving high importance to PM10, LDA is giving low to medium importance. The figure was divided into the associated measurement parameters to see if any of the sensor needs to be eliminated. In both plots, most Carbon Monoxide features have low importance except one or two which have very high importance. Therefore, the measurement cannot be eliminated.

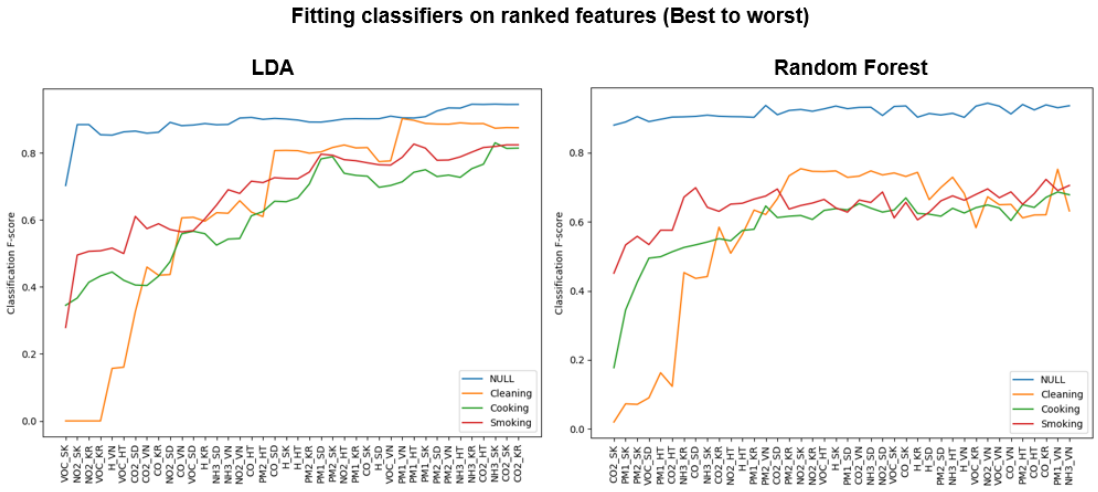


Figure 4.22: Adding ranked features one by one (Best first)

Figure 4.22 shows test scores as worse features were added to training matrix. Both classifiers had their respective ranking for the matrix, it is of interest to see if eliminating the worst features would have an effect on test performance. As we can see in the plot for Random forest, we could easily eliminate features where the F-scores are not improving. LDA on the other hand shows an upwards trend as more and more features are added.

4.3.2.3 Classification on Feature Matrix with Neural Networks

Neural networks is a technique that is extensively used to solve machine learning problems such as regression, classification, prediction and even signal synthesis.

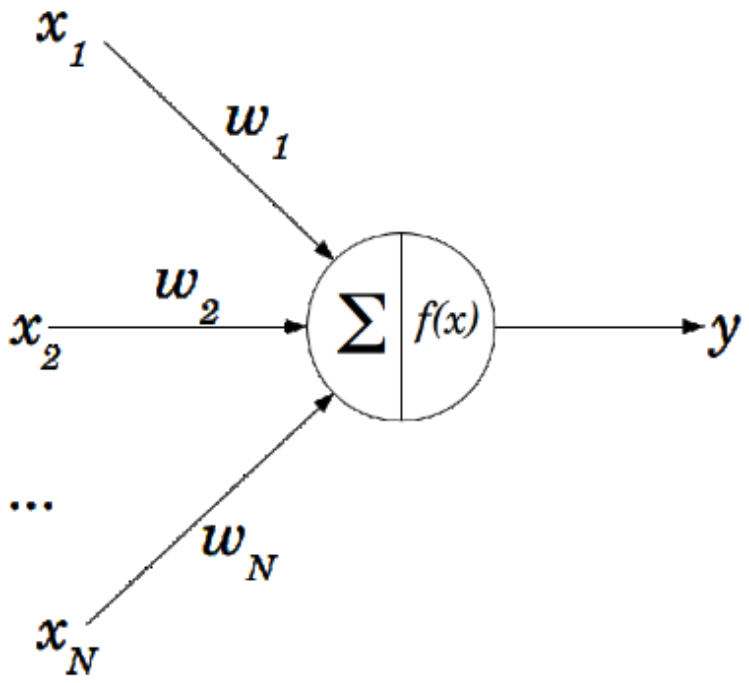


Figure 4.23: The Perceptron Model [38]

The biologically inspired perceptron unit takes a linear combination of the inputs, puts them through a non-linear function, and obtains the output. These perceptrons are connected in complicated 'black-box' type architectures. The large number of tuning parameters for in a Neural Network are automatically 'trained' using the training set examples through algorithm similar Stochastic Gradient Descent [39].

The training is an iterative procedure, where inputs are propagated forward through the network, the output, for an untrained network is not equal to the true labels of the training set, and the error is used in the backward propagation to update the parameters.

The hyperparameters selected for our model were as follows:

Activation function: Identity
Optimizer: Adam
Loss function: Cross Entropy Loss
Regularization: 10^{-3}
Validation fraction: 0.2

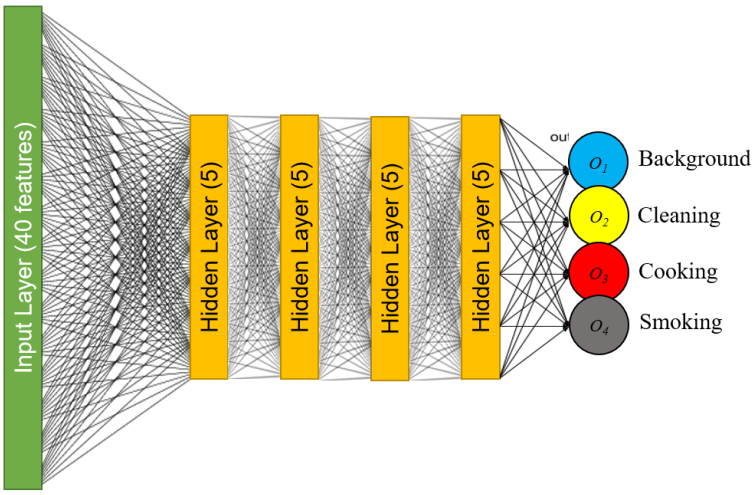


Figure 4.24: Architecture of Multi-layer perceptron used

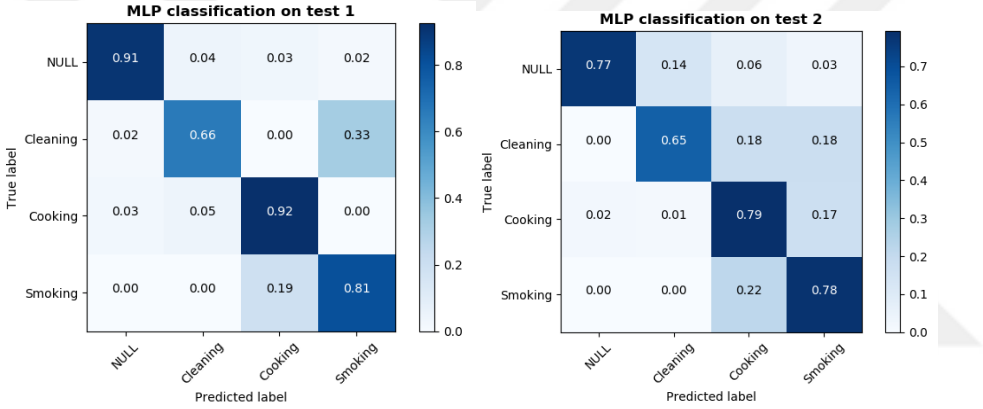


Figure 4.25: Confusion Matrices of MLP classifier on each test set

Table 4.9: Fitting metrics for MLP classifier

Data Set	Accuracy
Training	0.94
Cross Validation (5 fold)	0.83
Test	0.79

Table 4.10: Class-wise evaluation of MLP classifier

Class	Precision	Recall	F1 Score
Background	92	83	87
Cleaning	71	70	70
Cooking	74	69	71
Smoking	61	85	71
Total	**77**	**77**	**76**

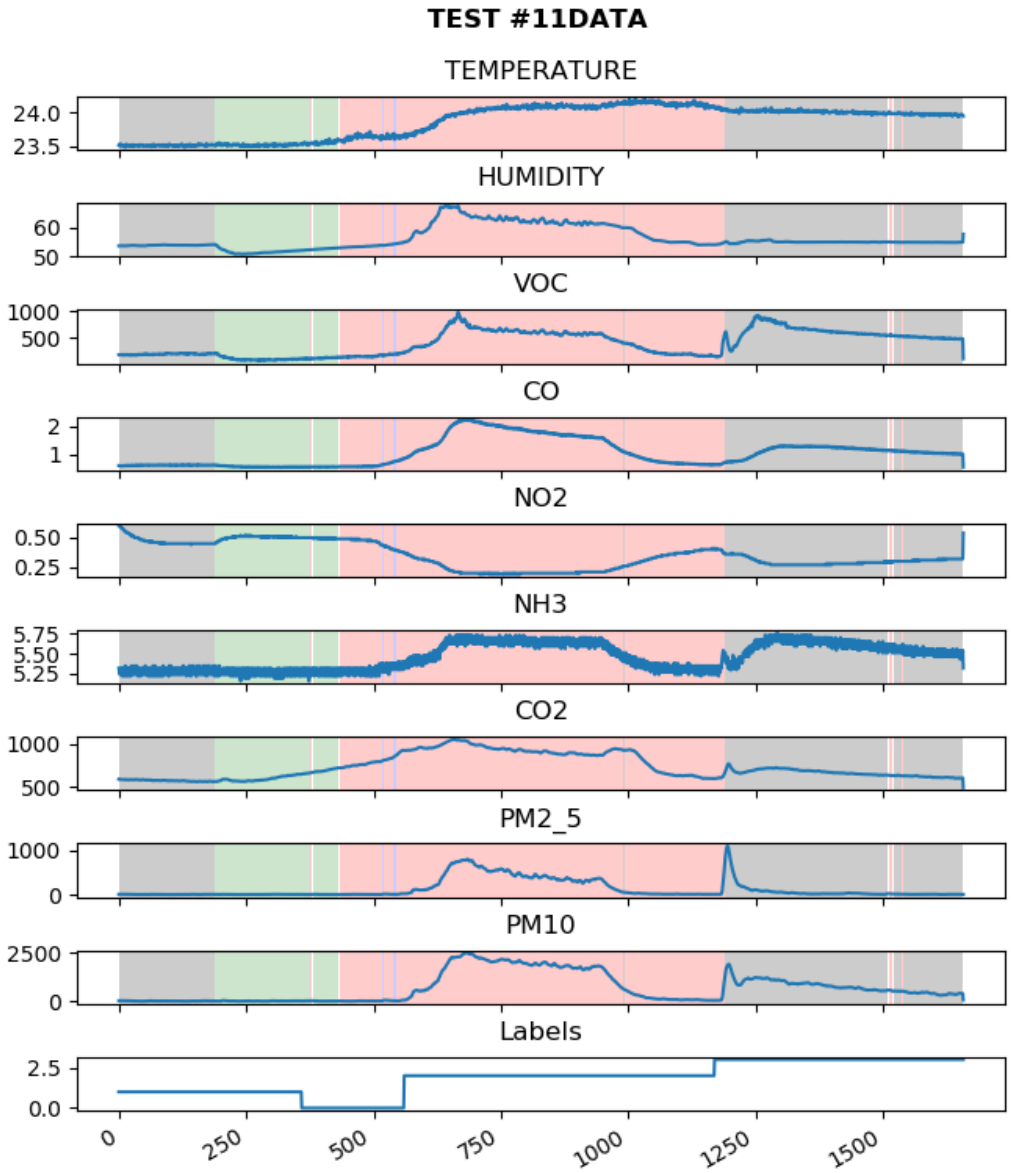


Figure 4.26: MLP Classifier Hypothesis on a unit from Test 1 (Blue=Background, Green=Cleaning, Red=Cooking, Black=Smoking)

4.3.2.4 Comparison of Multiclass Classifiers

Evaluating various classifiers, we concluded that cleaning is by a small margin the most difficult to detect. This can be traced to the fact that as compared to smoking and cooking, the convection of chemicals is much slower through the air. The selected classifier is LDA because of consistent scores on all metrics and classes.

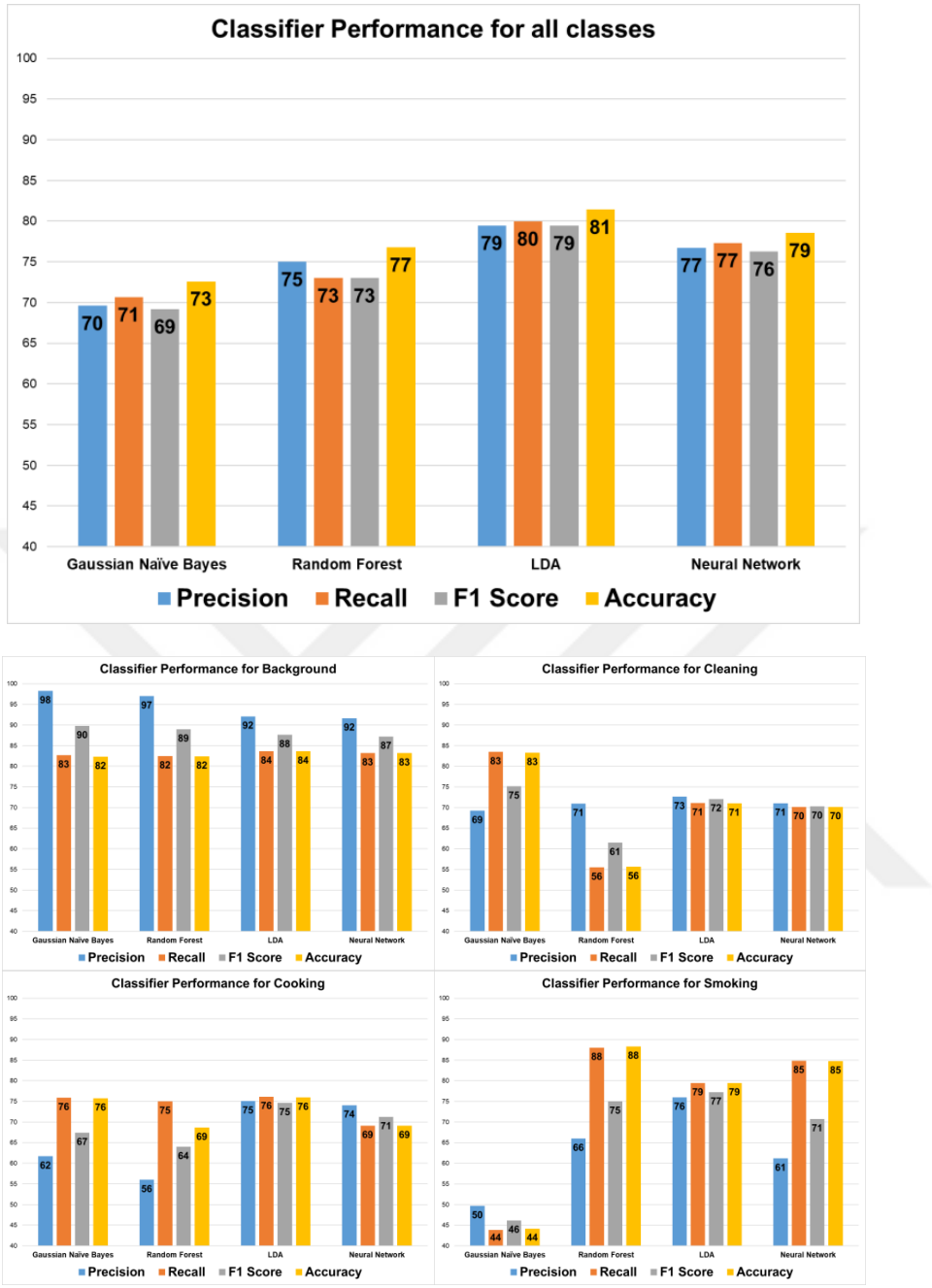


Figure 4.27: Comparison of classifiers using different metrics. Total (top), Class wise (bottom)

5. CONCLUSIONS AND FUTURE WORK

This study successfully achieved the following deliverables:

i. A sensor platform that measures 9 different air quality parameters, and has the hardware and software required to publish the data to a MySQL server at fixed intervals.

ii. Indoor Activity Dataset, which is labelled for four classes; background, cleaning, cooking and smoking. The total instances of each activity are 22, with additional 10 tests just for background data. In total of 81 hours of Air Quality data was captured throughout a course of 2 months, with metadata documented.

iii. Data analysis techniques from literature were evaluated with our own data.

iv. A binary classifier was proposed based on Skewness features and Decision Tree algorithm, with a test accuracy of 94%.

v. Two multiclass classifiers were proposed based on all extracted features. The classifiers provided a test accuracy of 79% (Linear Discriminant Analysis), and 81% (Multi-layer perceptron). LDA was preferred considering its performance on Precision-Recall Metrics.

The current scope of the study however was limited to 3 activities only, and in more or less similar test conditions. The next step would be to study these problems in a more uncontrolled setting, possibly of a real household. Also, a closed loop system that can predict activities based on long term forecast and perform actions can also be studied.

Figure 5.1 Shows how this work could be extended to a closed loop smart-home environment management system where activity detection could be combined with forecasting and device control. Also, additional sources of data such as openly accessible weather information, and additional measurement modules could be combined to provide a truly automatic ambience control in the living space. To

develop such a system requires a test living space, where additional sensors such as cameras and door sensors could be installed for data labelling.

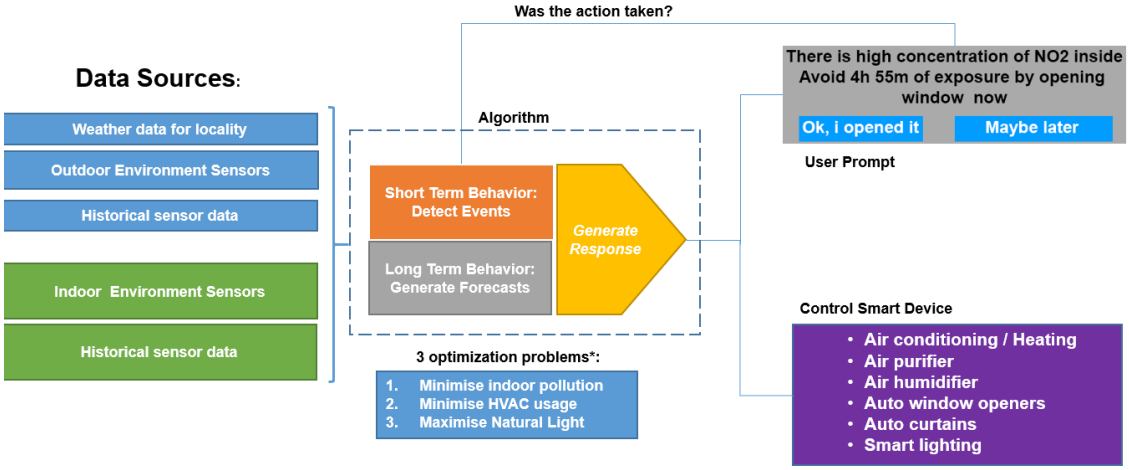


Figure 5.1: Closed loop environmental system for future studies

Another direction that can be worked on is the evaluation of Deep Neural Network architectures such as Recursive Neural Networks, Long and Short-Term Memory Neural networks which are known to work well for time series modelling, and Convolutional Neural Networks, which could really work well in handling dimensionality of data by organizing each time step as a frame and each sensor as a pixel therein.

REFERENCES

[1] **Kim, S. and Paulos**, **E.,** "inAir: measuring and visualizing indoor air quality," Ubiquitous Computing, 2009. [Online]. Available: https://dl.acm.org/citation.cfm?id=1620557. [Accessed 13 November 2018].

[2] **Kim, S. and Paulos**, **E.** "inAir: measuring and visualizing indoor air quality," , 2009. [Online]. Available: https://dl.acm.org/citation.cfm?id=1620557. [Accessed 13 11 2018].

[3] Centers for Disease Control and Prevention, "Carbon Monoxide Poisoning," [Online]. Available: https://ephtracking.cdc.gov/showCoRisk.action. [Accessed 13 November 2018].

[4] **Debbie, F. J. K., Jarvis**, **J. ,** "WHO Guidelines for Indoor Air Quality: Selected Pollutants," World Health Organisation, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK138707/. [Accessed 13 November 2018].

[5] **Alberts**, **W. ,** "Indoor air pollution: NO, NO2, CO, and CO2," *Journal of Allergy and Clinical Immunology,* vol. 94, no. 2, 1994.

[6] "Particulate matter (PM10 and PM2.5)," National Pollutant Inventory Australia, [Online]. Available: http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25. [Accessed 13 November 2018].

[7] **Netatmo**, "Healthy Home Coach Product Page," [Online]. Available: https://www.netatmo.com/en-us/aircare/homecoach. [Accessed 13 November 2018].

[8] **Büke, T. and Köne**, **A. Ç.,** "Assessing Air Quality in Turkey: A Proposed, Air Quality Index," *Sustainability,* vol. 8, no. 1, pp. 1-8, 2016.

[9] **Philip, A. , Peterson**, **J. D.** "Practical Use of Metal Oxide Semiconductor Gas Sensors for Measuring Nitrogen Dioxide and Ozone in Urban Environments," *MDPI Sensors,* July 2017.

[10] **Hwang, K.-I. ,Kang**, **J.,** "A Comprehensive Real-Time Indoor Air-Quality Level Indicator," *MDPI Sustainability,* 1 September 2016.

[11] **Fonollosa, J. , Rodriguez, I., Abhijeet**, **L. ,** "Human activity monitoring using gas sensor arrays," *Sensors and Actuators B. Chemical,* 2014.

[12] **Candanedo, L. M. and Feldheim**, **V.,** "Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models," *Energy and Buildings,* vol. 112, no. , pp. 28-39, 2016.

[13] **Jin, M., Bekiaris.-Liberis. N., Weekly. K., Spanos C.** "Sensing by proxy: Occupancy detection based on indoor CO2 concentration," in *UBICOMM* , 2015.

[14] **Gutmacher, D., Hoefer, U. and Wöllenstein, J.** , "Gas sensor technologies for fire detection," *Sensors and Actuators B-chemical,* vol. 175, no. , pp. 40-45, 2012.

[15] **Zhuang, Y., Lin, F. ,Yoo, ,E.-H. and Xu, W.,** "AirSense: A Portable Context-sensing Device for Personal Air Quality Monitoring," , 2015. [Online]. Available: https://dl.acm.org/citation.cfm?id=2757293. [Accessed 13 11 2018].

[16] **Bhattacharya, P. R**., "Indoor air quality monitoring using wireless sensor network," *IEEE. Sixth International Conference on Sensing Technology (ICST),* pp. 422-427, 2012.

[17] **Peng, P., Zhao, X., Pan X. and Ye, W.,** "Gas Classification Using Deep Convolutional Neural Networks," *Sensors,* vol. 18, no. 1, p. 157, 2018.

[18] **Biyi Fang, Z.** "AirSense: an intelligent home-based sensing system for indoor air quality analytics," *UbiComp Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing,* 2016.

[19] **Andrew, M., Saad, S. M.** "Pollutant Recognition Based on Supervised Machine Learning for Indoor Air Quality Monitoring Systems," *Applied Sciences,* vol. 7, no. 8, 2017.

[20] **Dang, C. T.,** "Theodor: A Step Towards Smart Home Applications with Electronic Noses," in *Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction, Article No. 11* , Berlin, Germany — September 20 - 21, 2018 , 2018.

[21] **Milna K., Kalit, T.,** "Application of electronic nose and electronic tongue in the dairy industry," in *Mljekarstvo : journal for dairy production and processing improvement*, 2014.

[22] SGX Sensortech, "MICS 6814 A multipixel gas sensor for the detection of CO, NO2, and NH3".

[23] **Jia Yan, Z.,** "Electronic Nose Feature Extraction Methods: A Review," *MDPI Sensors,* 2015.

[24] **Xu, P., Song, K., Xia, X., Chen, Y., Wang Q and Wei, G.,** "Temperature and Humidity Compensation for MOS Gas Sensor Based on Random Forests," , 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-10-6373-2_14. [Accessed 13 11 2018].

[25] **Sensirion AG**, *Preliminary datasheet SVM30, Multigas, humidity and tempereature combo module.*

[26] **SGX Sensortech**, *MICS-6814 Multipixel MOS sensor datasheet.*

[27] **Wuhan Cubic Optoelectronics Ltd**., *CM1107 Dual-Beam CO2 Sensor Datasheet.*

[28] **Amphenol Corporation**, *SMART Dust Sensor SM-PWM-01C Datasheet.*

[29] **Seeed Studio**, "Linkit One Development board," [Online]. Available: http://wiki.seeedstudio.com/LinkIt_ONE/#specifications. [Accessed 14 November 2018].

[30] **Su, L. T.**, "The relevance of recall and precision in user evaluation," *Journal of the Association for Information Science and Technology,* vol. 45, no. 3, pp. 207-217, 1994.

[31] **Quantstart**, "Basics of Statistical Mean Reversion Testing," [Online]. Available: https://www.quantstart.com/articles/Basics-of-Statistical-Mean-Reversion-Testing. [Accessed 14 November 2018].

[32] **Breiman L. and Friedman, J.**, "Classification and Regression Trees (Wadsworth Statistics/Probability)," , 1984. [Online]. Available: https://science-open.com/document/vid/8626444e-8b4f-494b-a185-b0c6530c53b5. [Accessed 14 11 2018].

[33] **Silva A. P. D. and Brito, P.**, "Linear discriminant analysis for interval data," *Computational Statistics,* vol. 21, no. 2, pp. 289-308, 2006.

[34] **Choi, S., Eom T. and Jeong, G.-M.**, "Gas Classification Using Combined Features Based on a Discriminant Analysis for an Electronic Nose," *Journal of Sensors,* vol. 2016, no. , pp. 1-9, 2016.

[35] **Smith, L. I.**, "A tutorial on Principal Components Analysis," , 2002. [Online]. Available: http://iro.umontreal.ca/~pift6080/h09/documents/papers/pca_tutorial.pdf. [Accessed 14 11 2018].

[36] "20 Recursive Feature Elimination," [Online]. Available: https://topepo.github.io/caret/recursive-feature-elimination.html. [Accessed 14 November 2018].

[37] **Guyon, I., Weston, J., Barnhill S. and Vapnik, V.**, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning,* vol. 46, no. 1, pp. 389-422, 2002.

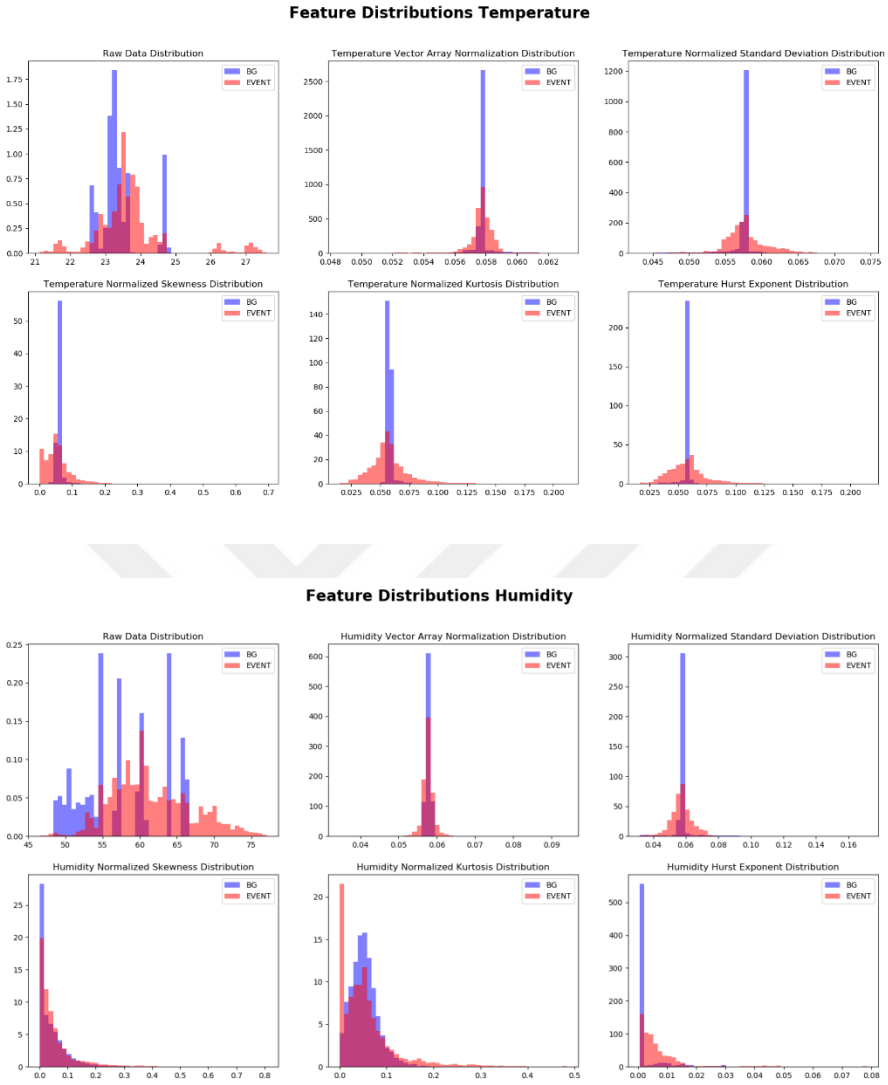
[38] **Lucidarme, P.**, "Perceptron update rule demonstration," [Online]. Available: https://www.lucidarme.me/perceptron-update-rule-demonstration/. [Accessed 14 November 2018].

[39] **Bottou, L.**, "Large-Scale Machine Learning with Stochastic Gradient Descent" [Online]. Available: http://leon.bottou.org/publications/pdf/compstat-2010.pdf. [Accessed 14 11 2018].

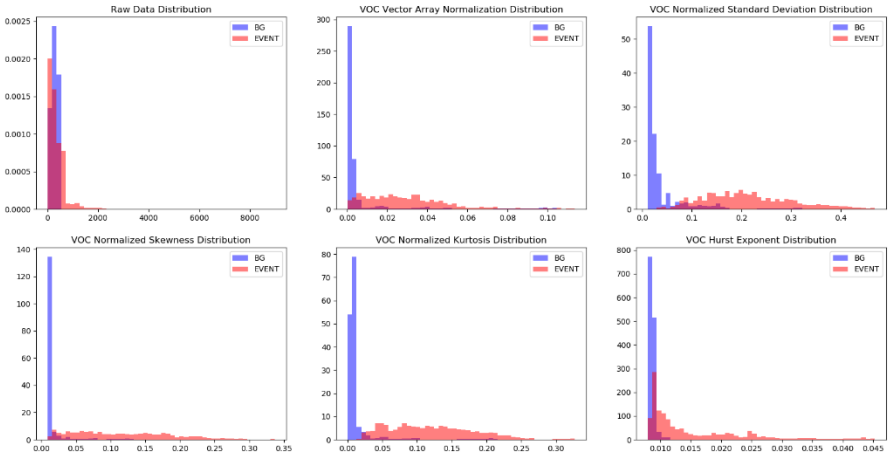
[40] **Environmental Protection Agency USA**, "Exposure Assessment Tools by Media - Air," 13 November 2018. [Online]. Available: https://www.epa.gov/expobox/exposure-assessment-tools-media-air#indoorair.

APPENDICES

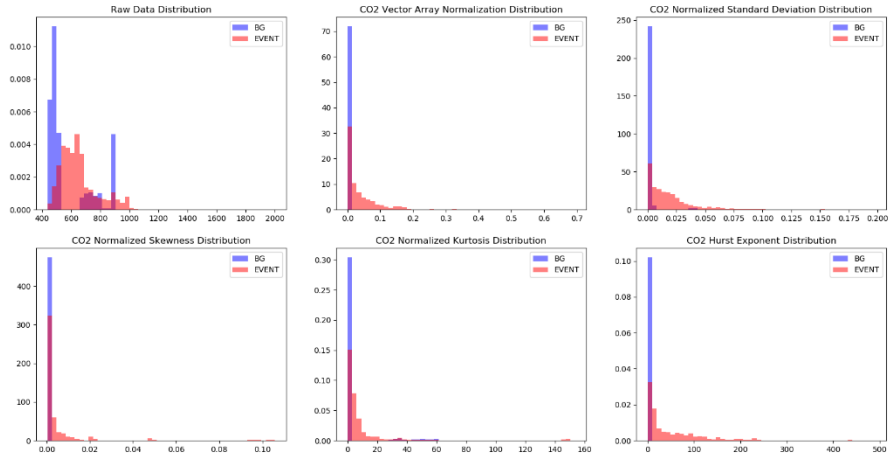
APPENDIX A: Feature distributions for the Binary case



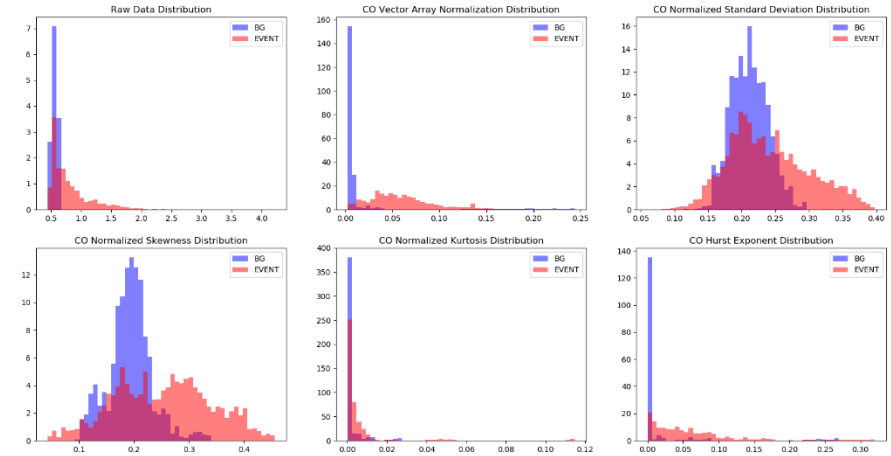
Feature Distributions VOC



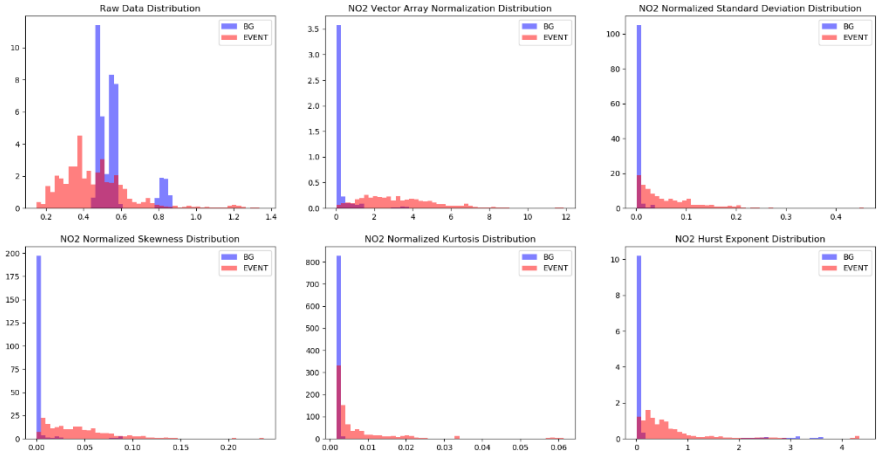
Feature Distributions CO2



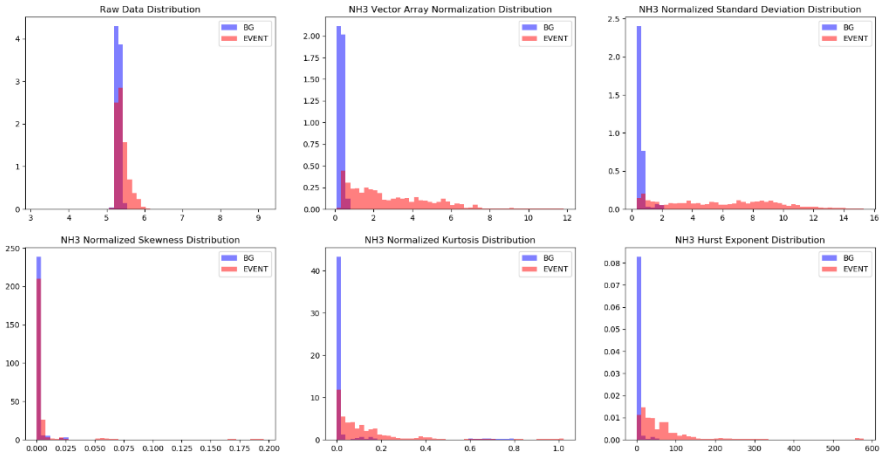
Feature Distributions CO



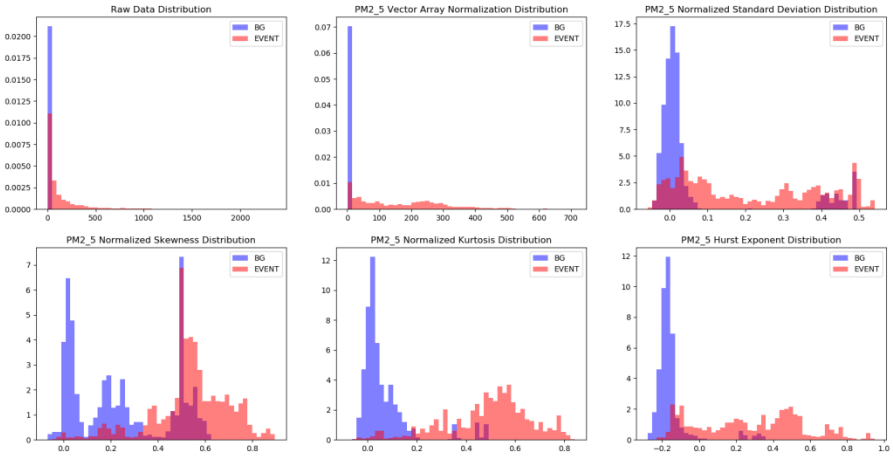
Feature Distributions NO2



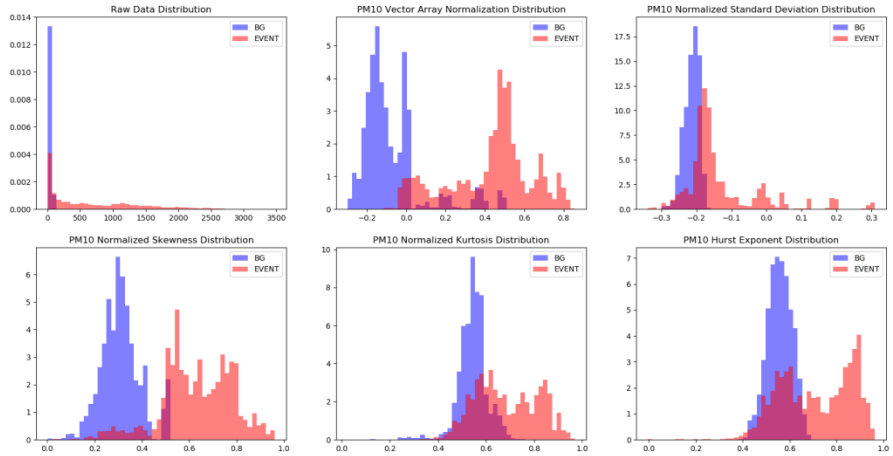
Feature Distributions NH3



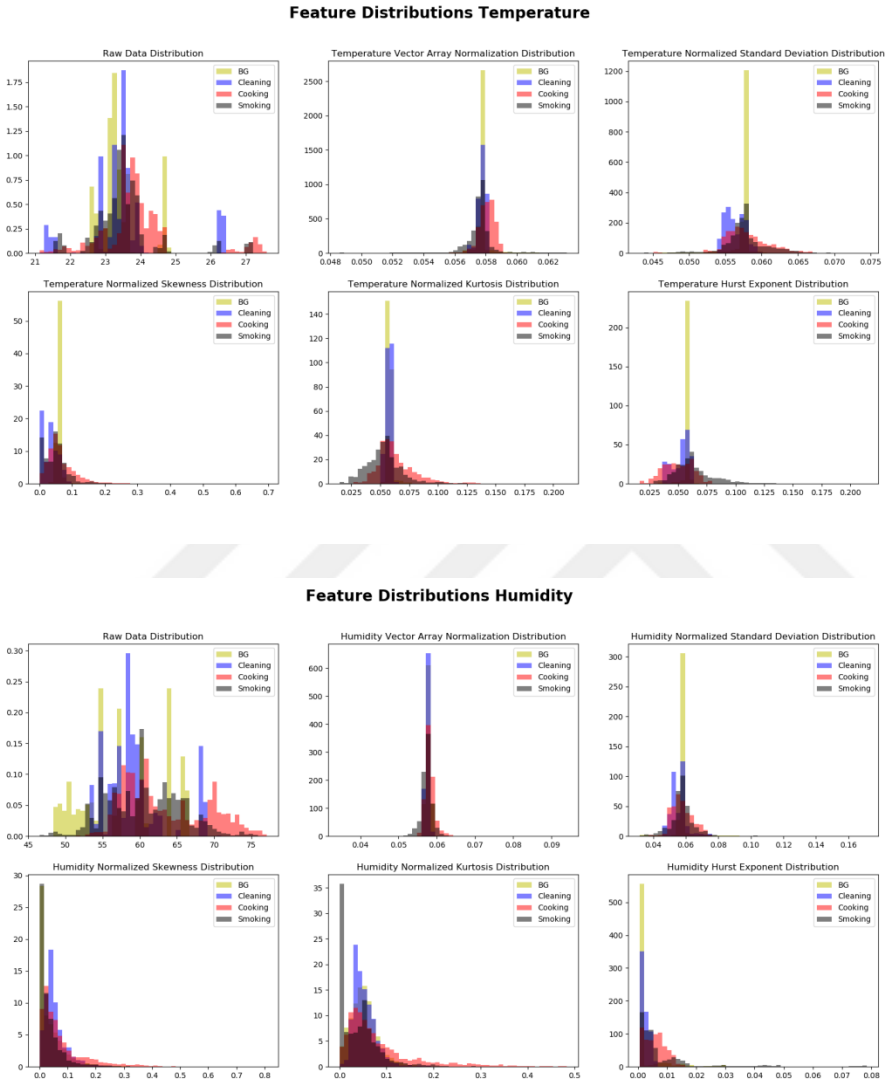
Feature Distributions PM2_5



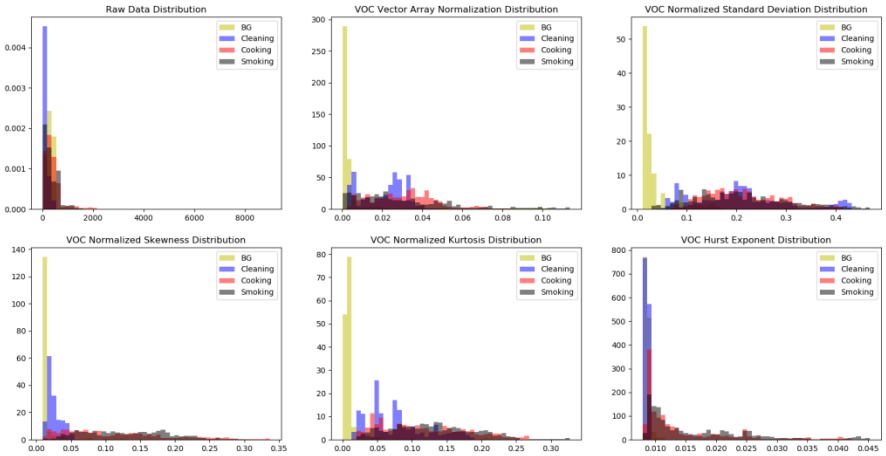
Feature Distributions PM10



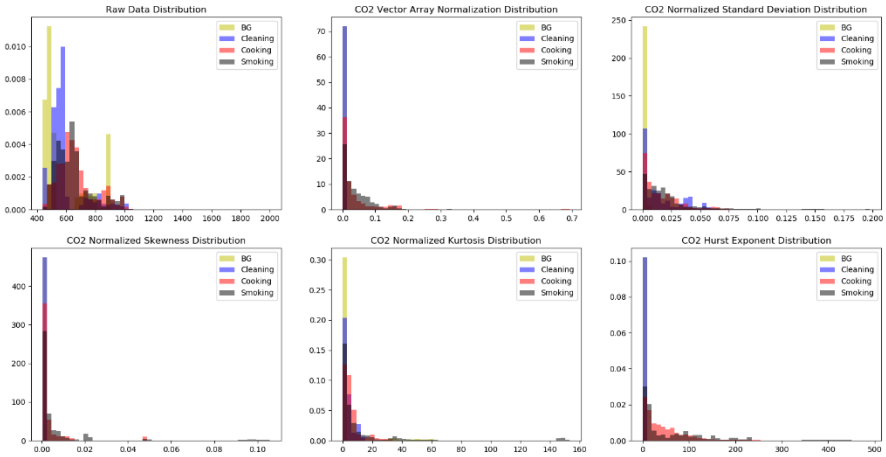
APPENDIX B: Feature Distributions for the Multiclass case



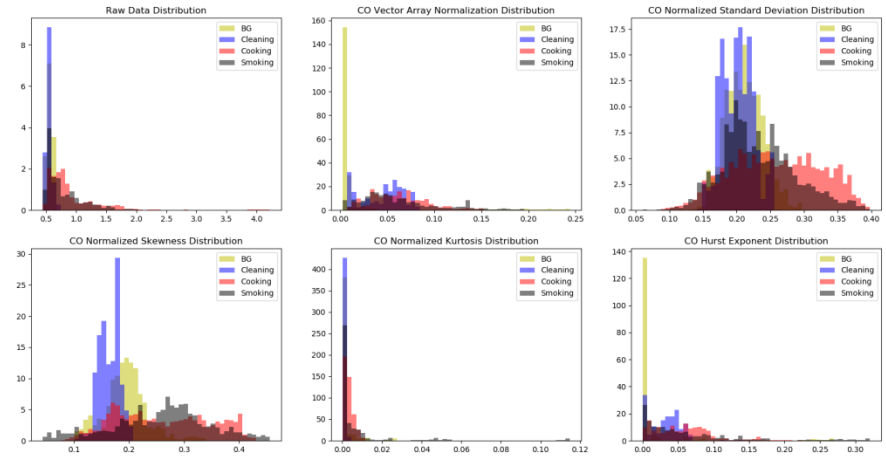
Feature Distributions VOC



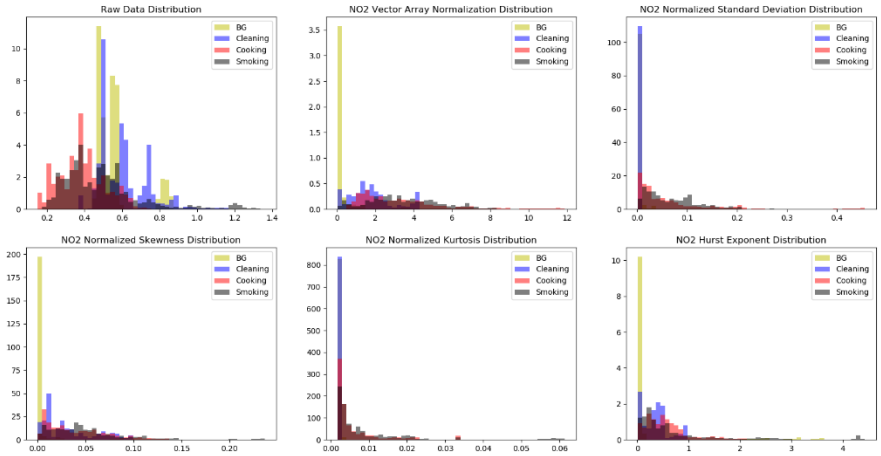
Feature Distributions CO2



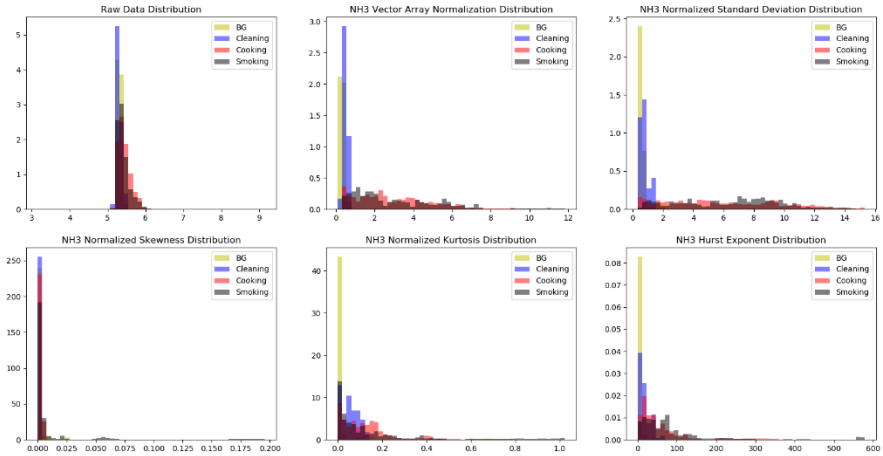
Feature Distributions CO



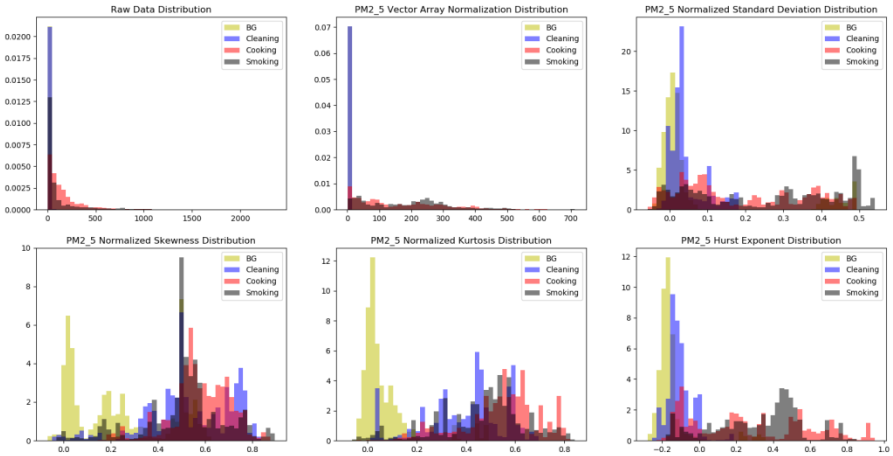
Feature Distributions NO2



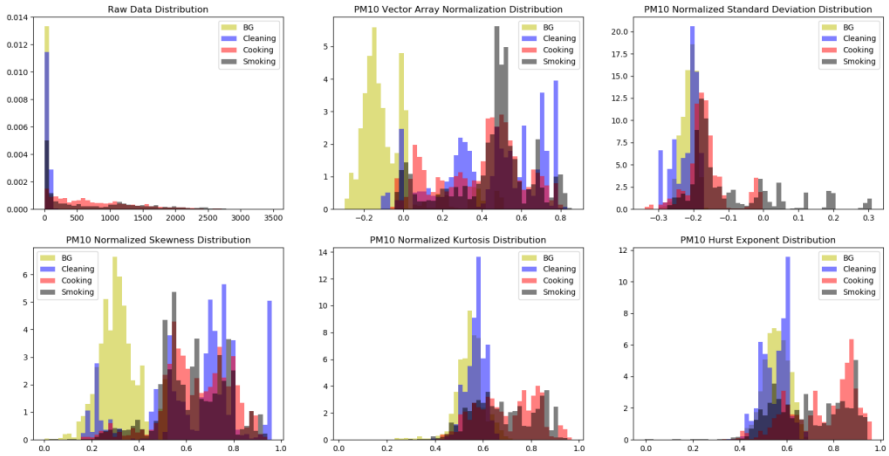
Feature Distributions NH3



Feature Distributions PM2_5



Feature Distributions PM10



CURRICULUM VITAE

Name Surname **: Muhammad Ali NAGARIA**

Place and Date of Birth **: 1993 Karachi**

E-mail **: alinagaria@gmail.com**

EDUCATION

- **B.Sc.** **:** 2016, Sabanci University, Mechatronics Engineering
- **B.Sc. (exchange)** **:** 2015, Hong Kong University of Science & Technology

PROFESSIONAL EXPERIENCE:

- 2017-present Arçelik A.Ş. Sensor Technologies, R&D Project Engineer
- 2015 LANXESS Hong Kong CAE and Part-testing intern