

IEEE Globecom Workshop on Communication, Computing and
Networking in Cyber Physical Systems, Abu Dhabi, 2018

Machine Learning-based Occupancy Estimation using Multivariate Sensor Nodes

Adarsh Pal Singh

Signal Processing and Communication Research Center
IIIT Hyderabad

December 13, 2018

The Team



- **Adarsh Pal Singh**
4th Year, B. Tech. + MS by Research in Electronics and Communication
- **Vivek Jain**
4th Year, B. Tech. in Electronics and Communication
- **Sachin Chaudhari**
Assistant Professor, IIIT Hyderabad
- **Vishal Garg**
Professor, IIIT Hyderabad



Norwegian University of
Science and Technology

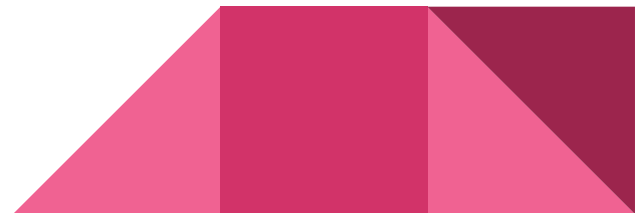
- **Stefan Werner**
Professor, NTNU Gløshaugen
 - **Frank Alexander Kraemer**
Associate Professor, NTNU Gløshaugen
-

Section I

Basics, Motivation and Contribution

In A Nutshell...

- The aim of this research project is to accurately estimate the number of occupants in a room using non-intrusive sensors only.
- Supervised machine learning models are used to gain inference about occupancy count from the deployed sensor nodes in the room.



Motivation

- In buildings, a large chunk of energy is spent on HVAC and lighting systems.
- Studies have demonstrated energy savings upto 30% in buildings where the occupancy pattern was known.
- Existing approaches of using camera, RFID, smart wearables and WiFi are either intrusive or privacy invasive.
- A lot of research has already been done in the field of occupancy detection.



Our Contributions

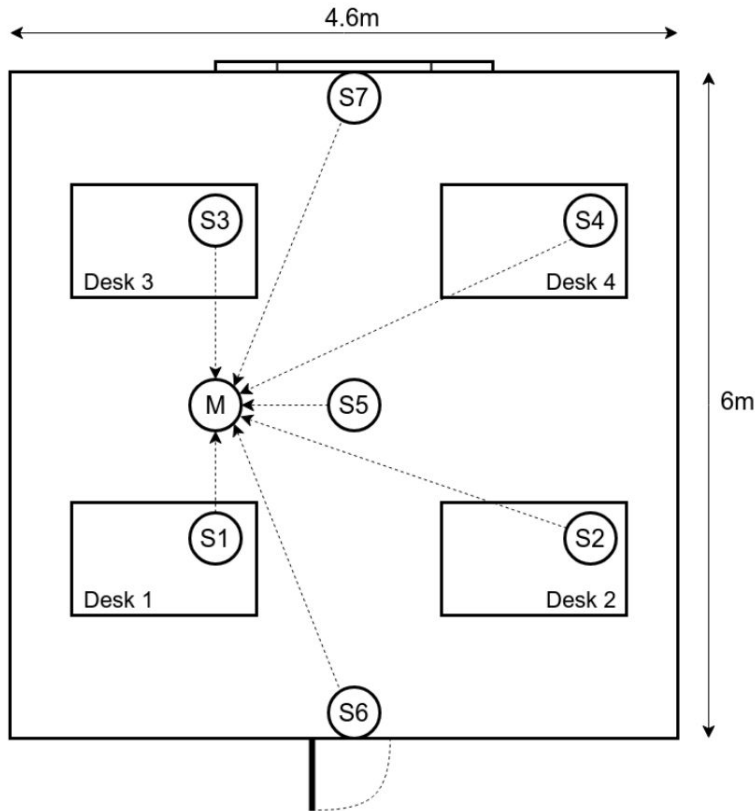
- A new dataset formed by deploying multiple multivariate sensor nodes in a lab.
- Accuracy and F1 score reported for different supervised ML models trained on various combinations of features, both homogeneous and heterogeneous.
- Data preprocessing and feature engineering techniques discussed. In particular, a new feature was derived from CO₂ readings.
- Analysis with PCA (unsupervised learning) done to see the performance of a reduced-dimensional dataset.



Section II

Experimental Setup and Data Collection

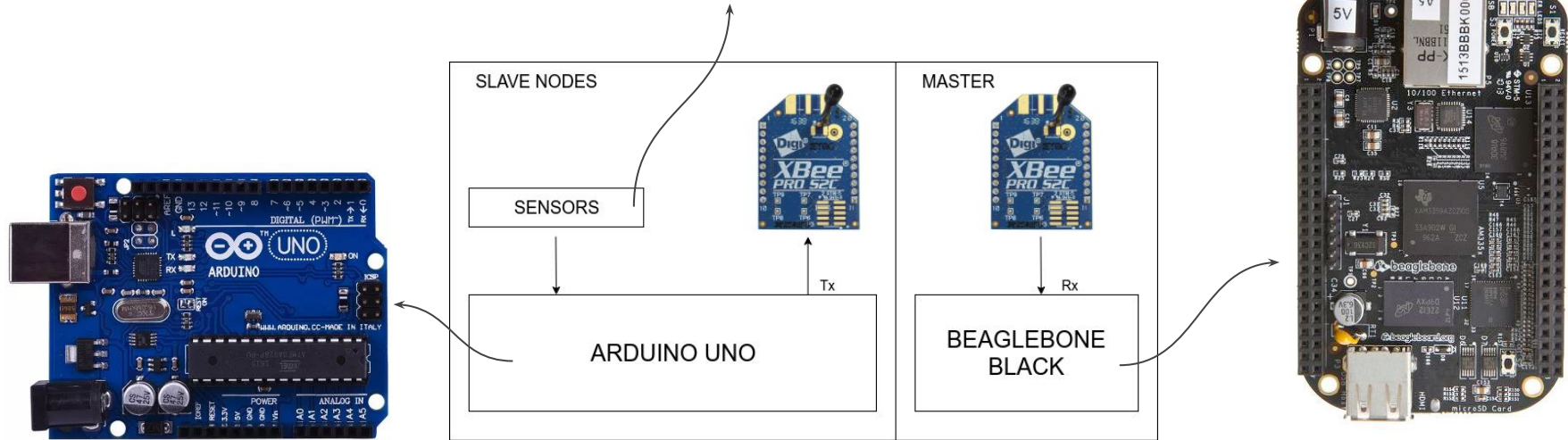
Experimental Setup




- A WSN with star topology was deployed in the lab.
- Data logging network with one Master and 7 Slave sensor nodes.
- Slaves transmit to sink every 30s.
- **S1 - S4**: Temperature, light, sound sensors
- **S5**: CO₂ sensor
- **S6 - S7**: PIR sensor

Architecture of Wireless Nodes

Sensor	Parameter	Resolution	Accuracy
BH1750	Light	1 Lux	1.2 times
MAX4466	Sound	0.01V*	-
MH-Z14A	CO ₂	5ppm	±50ppm
Digital PIR	Motion	-	-



Sensor Data Sampling and Transmission

- Data from sensor nodes transmitted to sink every 30s.
 - Temperature, Light and CO₂ sampled once in 30s by Arduino.
 - For sound, continuous sampling done and the maximum peak-to-peak value achieved in 30s is sent.
 - PIR polled every 2.5 seconds. For even a single motion event in 30s, a '1' is sent (else, '0' sent).
 - Master appends data packets from each slave into corresponding .txt files after appending the time-stamp.
- 

Section III

Data Preprocessing and Feature Engineering

Dataset Formation

- The data was collected continuously for four days => over 10,000 data points!

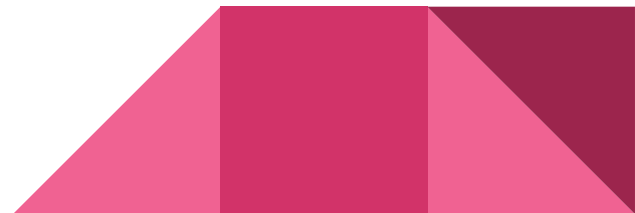
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Date	Time	Temp-1	Temp-2	Temp-3	Temp-4	Light-1	Light-2	Light-3	Light-4	Sound-1	Sound-2	Sound-3	Sound-4	CO2-5	PIR-6	PIR-7
2	2017/12/22	10:49:41	24.94	24.75	24.56	25.38	121	34	53	40	0.08	0.19	0.06	0.06	390	0	0
3	2017/12/22	10:50:12	24.94	24.75	24.56	25.44	121	33	53	40	0.93	0.05	0.06	0.06	390	0	0
4	2017/12/22	10:50:42	25	24.75	24.5	25.44	121	34	53	40	0.43	0.11	0.08	0.06	390	0	0
5	2017/12/22	10:51:13	25	24.75	24.56	25.44	121	34	53	40	0.41	0.1	0.1	0.09	390	0	0
6	2017/12/22	10:51:44	25	24.75	24.56	25.44	121	34	54	40	0.18	0.06	0.06	0.06	390	0	0
7	2017/12/22	10:52:14	25	24.81	24.56	25.44	121	34	54	40	0.13	0.06	0.06	0.07	390	0	0
8	2017/12/22	10:52:45	25	24.75	24.56	25.44	120	34	54	40	1.39	0.32	0.43	0.06	390	1	0
9	2017/12/22	10:53:15	25	24.81	24.56	25.44	121	34	54	41	0.09	0.06	0.09	0.05	390	0	0
10	2017/12/22	10:53:46	25	24.81	24.56	25.5	122	35	56	43	0.09	0.05	0.06	0.13	390	0	0
11	2017/12/22	10:54:17	25	24.81	24.56	25.5	101	34	57	43	3.84	0.64	0.48	0.39	390	1	1
12	2017/12/22	10:54:47	25.06	24.81	24.56	25.44	122	35	57	43	2.2	0.31	0.33	0.21	390	1	1
13	2017/12/22	10:55:18	25.06	24.81	24.56	25.5	123	35	57	44	0.42	0.13	0.14	0.09	390	1	0
14	2017/12/22	10:55:49	25.06	24.88	24.63	25.5	123	35	57	43	0.21	0.15	0.07	0.06	390	1	0
15	2017/12/22	10:56:19	25.06	24.81	24.63	25.56	123	35	57	44	1.66	0.21	0.12	0.09	390	1	0
16	2017/12/22	10:56:50	25.06	24.88	24.63	25.56	123	35	58	44	0.57	0.17	0.21	0.13	390	1	0
17	2017/12/22	10:57:21	25.06	24.88	24.63	25.56	123	35	58	44	0.14	0.17	0.15	0.06	390	1	0
18	2017/12/22	10:57:51	25.06	24.88	24.63	25.56	122	35	58	44	0.24	0.05	0.1	0.05	395	0	0
19	2017/12/22	10:58:22	25.06	24.81	24.63	25.56	122	35	57	44	0.25	0.07	0.07	0.05	395	1	0
20	2017/12/22	10:58:52	25.06	24.81	24.63	25.56	123	35	58	44	0.23	0.07	0.08	0.06	395	1	0
21	2017/12/22	10:59:23	25.06	24.88	24.63	25.5	123	35	58	45	0.13	0.06	0.06	0.06	395	0	0
22	2017/12/22	10:59:54	25.13	24.88	24.63	25.5	123	35	58	44	0.13	0.04	0.06	0.06	390	0	0
23	2017/12/22	11:00:24	25.06	24.88	24.63	25.56	123	35	57	44	0.15	0.05	0.06	0.06	395	0	0
24	2017/12/22	11:00:55	25.13	24.88	24.63	25.56	123	35	58	45	0.11	0.05	0.06	0.06	395	0	0

Dataset Formation

- 15 sensory features and no **derived features**.
- The raw dataset collected by the Master had **missing values, time skews and no ground truth**.

Feature
Engineering

Preprocessing

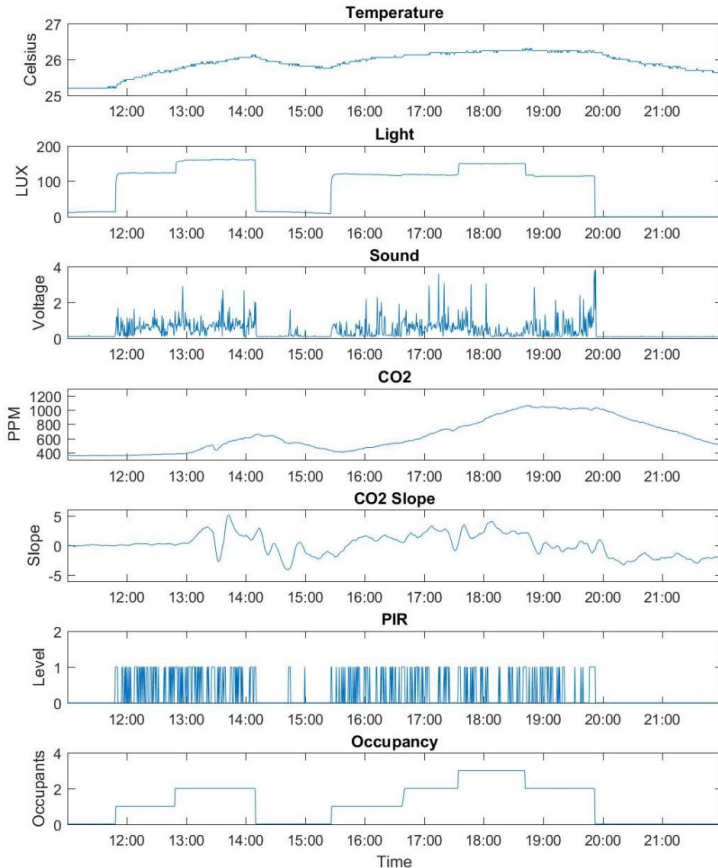


Data Preprocessing

- A few seconds of variation was present between the arrival times of packets of different nodes. We merged the time-stamps within a given time frame into a common vector.
- Feature vectors with missing data were simply deleted.
- The manually logged ground truth of occupancy count was appended in the last column.



Feature Engineering



- CO₂ gives excellent deviation to the number of occupants.
 - However, it takes several minutes for the readings to rise/fall to a steady state.
 - A new feature was derived in the form of slope of CO₂ by which we can infer the rising or falling trend of CO₂ values.
 - This was calculated by fitting a linear regression in a window of 25 points at each instance and calculating the slope of the line.
-

Section IV

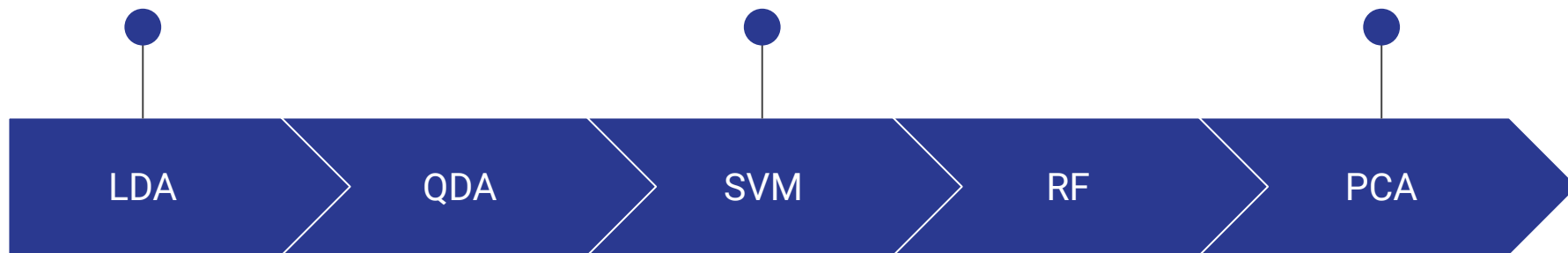
Machine Learning Algorithms

Machine Learning Algorithms

Linear Discriminant
Analysis

Support Vector
Machine

Principal Component
Analysis



Quadratic Discriminant
Analysis

Random Forest

Supervised

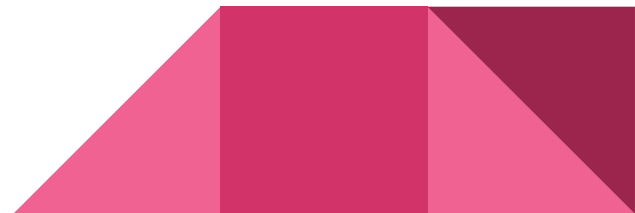
Unsupervised

Linear Discriminant Analysis (LDA)

- Assumes a multivariate Gaussian distribution with μ_k and Σ for class conditional probability $P(X|y = k)$.
- The predictions are made using Bayes' rule-

$$P(y = k|X) = \frac{P(X|y = k)P(y = k)}{\sum_i P(X|y = i)P(y = i)}$$

- The class prior probability $P(y = k)$ is learned from the training data along with μ_k and Σ .
- No tunable hyperparameter.



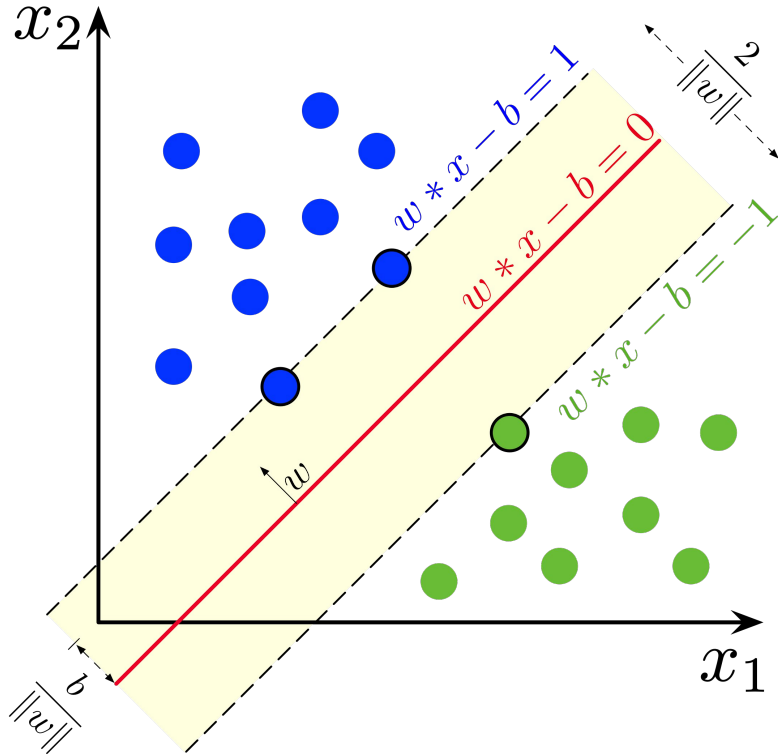
Quadratic Discriminant Analysis (QDA)

- Assumes a multivariate Gaussian distribution with μ_k and Σ_k for class conditional probability $P(X|y = k)$.
- The predictions are made using Bayes' rule-

$$P(y = k|X) = \frac{P(X|y = k)P(y = k)}{\sum_i P(X|y = i)P(y = i)}$$

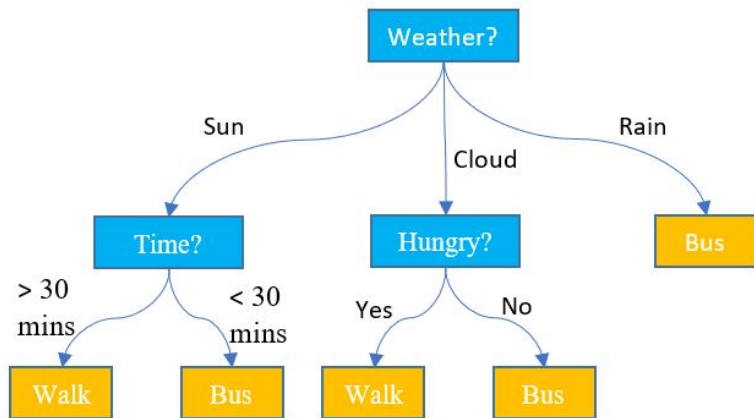
- The class prior probability $P(y = k)$ is learned from the training data along with μ_k and Σ_k .
 - No tunable hyperparameter.
- 

Support Vector Machine (SVM)



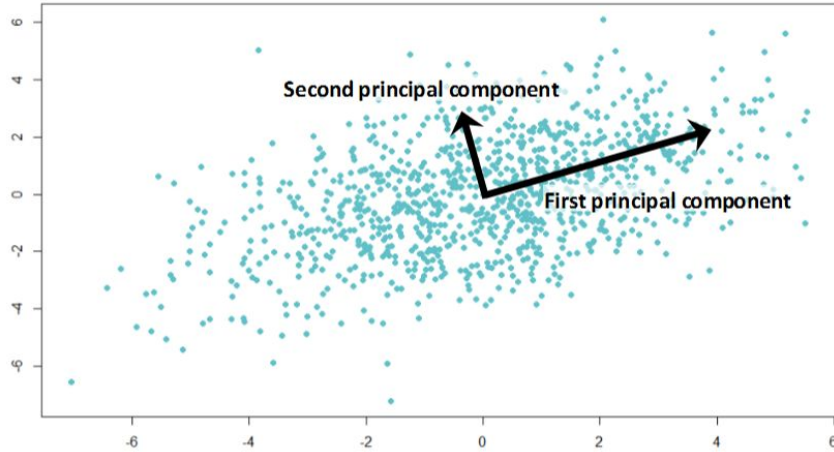
- No assumptions about the data.
- Each n-dimensional feature vector is a point in an n-dimensional feature space.
- SVM attempts to fit an optimal hyperplane between 2 classes with the help of support vectors.
- For k classes, $k(k-1)/2$ classifiers.
- C: penalty hyperparameter.
- Non-linear boundary with kernel.

Random Forest (RF)



- A decision tree is a tree-shaped structure in which each internal node represents a logical test on some feature and each leaf node represents an outcome class.
- RF is an ensemble of DTs grown on some part of the dataset.
- Classification and Regression Tree (CART) algorithm used for tree formation (training).
- Hyperparameters: `n_trees`, `min_samples_split`, `max_depth`.

Principal Component Analysis (PCA)



- PCA is a dimension-reducing unsupervised procedure in which a dataset having multiple and possibly correlated features is decomposed into a set of orthogonal components that capture the maximum amount of variance.
- This procedure is scale variant.
- The reduced dataset is then fitted with supervised models for performance evaluation.

Section V

Experiments and Results

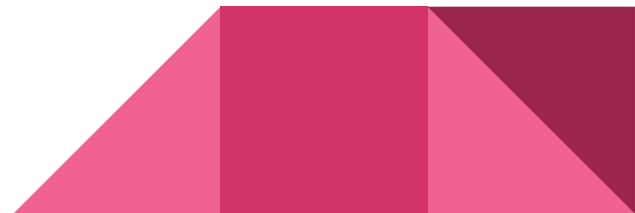
Experimentation Logistics

- 10-fold cross validation.
- Accuracy and macro F1 score (necessary since dataset is skewed) reported.
- Scikit-learn in Python.

- Normalized training data used for SVM.
- Linear and RBF kernel used.
- Penalty hyperparameter tuned for each feature set.

- “n_trees” tuned for RF and kept at 30.
- “min_samples_split” tuned to prevent overfitting.

- Phase I: Homogeneous fusion.
- Phase II: Heterogeneous fusion.
- Phase III: Dimensionality reduction using PCA.



Phase I Results

Feature	Metric	LDA	QDA	SVM (Linear)	SVM (RBF)	RF
Temp{1,2,3,4}	A	0.840	0.862	0.866	0.895	0.869
	F1	0.479	0.590	0.554	0.730	0.657
Light{1,2,3,4}	A	0.973	0.919	0.973	0.973	0.972
	F1	0.928	0.854	0.929	0.927	0.925
Sound{1,2,3,4}	A	0.851	0.879	0.875	0.885	0.887
	F1	0.449	0.544	0.542	0.591	0.601
PIR{6,7}	A	0.869	0.869	0.870	0.870	0.870
	F1	0.474	0.474	0.466	0.460	0.460
CO ₂	A	0.809	0.808	0.812	0.812	0.763
	F1	0.383	0.409	0.286	0.314	0.329
Slope	A	0.852	0.831	0.870	0.870	0.876
	F1	0.387	0.394	0.462	0.510	0.564
CO ₂ , Slope	A	0.891	0.867	0.890	0.888	0.873
	F1	0.556	0.590	0.592	0.635	0.559

- CO₂ slope feature shows promising result.
- CO₂ + slope fusion further increases the performance.
- Light gives the best performance with an F1 score of 0.929 (Linear SVM).
- Temperature, second best with an F1 score of 0.73 (RBF SVM).

Phase II Results

Feature	Metric	LDA	QDA	SVM (Linear)	SVM (RBF)	RF
Temp{1,2,3,4}, CO ₂ , Slope	A	0.903	0.881	0.904	0.912	0.894
	F1	0.653	0.680	0.667	0.750	0.684
Temp{1,2,3,4}, CO ₂ , Slope, Sound{1,2,3,4}	A	0.920	0.908	0.933	0.924	0.918
	F1	0.735	0.749	0.793	0.782	0.731
Temp{1,2,3,4}, CO ₂ , Slope, Sound{1,2,3,4}, PIR{6,7}	A	0.922	0.910	0.934	0.924	0.919
	F1	0.737	0.748	0.793	0.780	0.734
Temp{1,2,3,4}, CO ₂ , Slope, Sound{1,2,3,4}, PIR{6,7}, Light{1,2,3,4}	A	0.980	0.957	0.982	0.984	0.978
	F1	0.946	0.911	0.948	0.953	0.933

- One sensor type at a time added in a greedy manner (light considered only in the end).
- Complete dataset gives best performance with F1 score of 0.953 (RBF SVM).
- Without light, F1 score of ~0.8 achievable.

Confusion Matrices obtained in Phase II

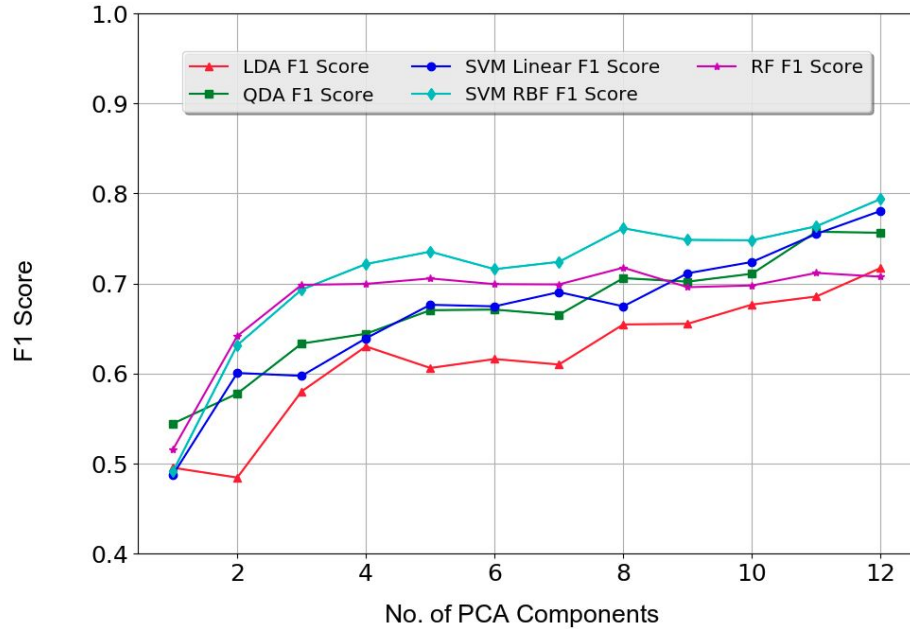
	Predicted 0	Predicted 1	Predicted 2	Predicted 3
Actual 0	8117	43	41	27
Actual 1	104	336	19	0
Actual 2	65	48	502	133
Actual 3	21	7	154	512

Linear SVM case for the complete dataset devoid of light features.

	Predicted 0	Predicted 1	Predicted 2	Predicted 3
Actual 0	8196	1	3	28
Actual 1	0	453	6	0
Actual 2	0	0	712	36
Actual 3	10	1	67	616

SVM with RBF kernel case for the complete dataset.

Phase III Results




- PCA done for complete dataset devoid of light features.
- Just 4 components give 92% accuracy and 0.72 F1 with RBF SVM.

Section VI

Conclusion and Future Work

Conclusion

- A deployment scheme involving multiple multivariate sensor nodes for occupancy count estimation was proposed.
 - Various methods to process large amounts of data obtained from the WSN discussed.
 - The proposed slope of CO₂ feature improved the accuracy and F1 metric.
 - The results show a promising 98.4% accuracy of occupancy estimation and a high F1 score of 0.953 using SVM with RBF kernel.
 - An accuracy of 92% and a moderate F1 score of 0.72 achievable with just 4 PCA components without light features.
- 

Future Work

- Extension of this model to large workspaces.
- Experiments with real-time extension of this model.
- Transfer learning approaches for faster training in multiple rooms.

