ML-based Distributed Decision Making with Convergent Replicated Data Types:

An Application for Distributed Pollution Monitoring

M.Tech Project

Under the supervision of

Dr.Sandip Chakraborty

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

Presented by: Amit Kumar

20CS30003

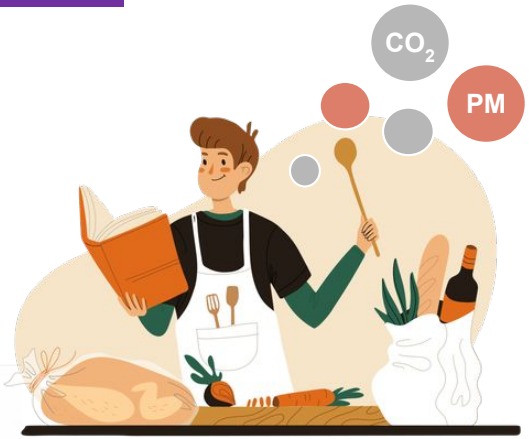
Indoor Pollution & Air Quality Monitors

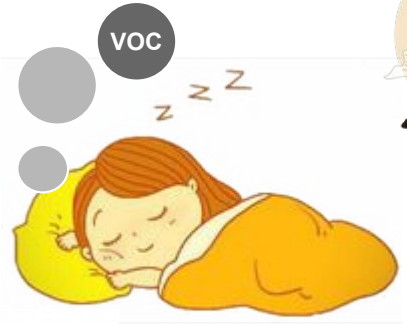
Household air pollution(WHO Report):

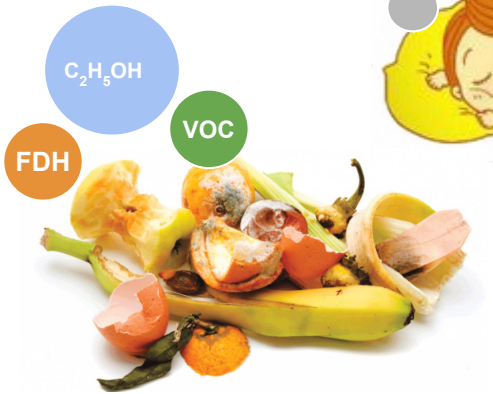
- ❏ estimated 3.2 million deaths per year in 2020
- ❏ including over 237,000 deaths of children under the age of 5

Low cost air quality monitors:

- ❏ isolated measurements (location dependent)
- ❏ fail to capture complex dynamics of air pollution propagation across interconnected indoor spaces.







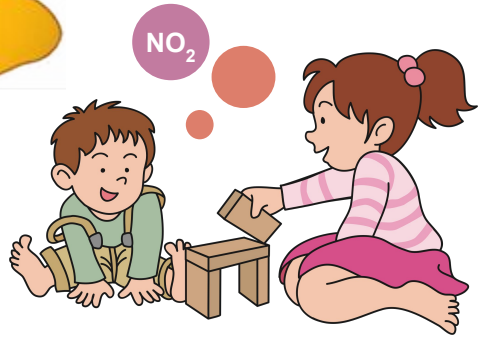
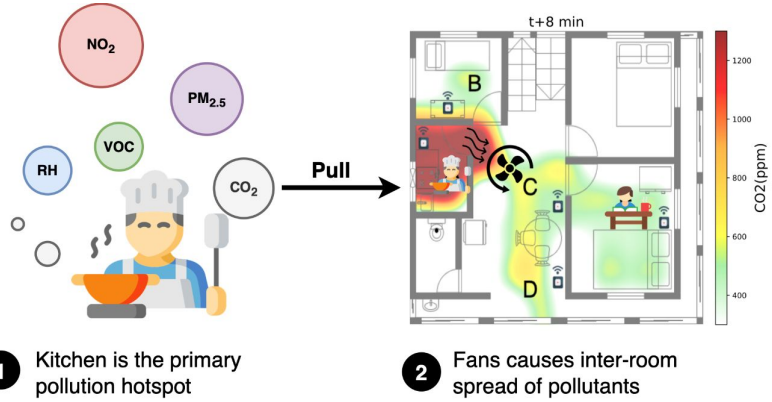


Image sources: https://www.flaticon.com/

Air Pollution Propagation



1 Kitchen is the primary pollution hotspot

2 Fans causes inter-room spread of pollutants

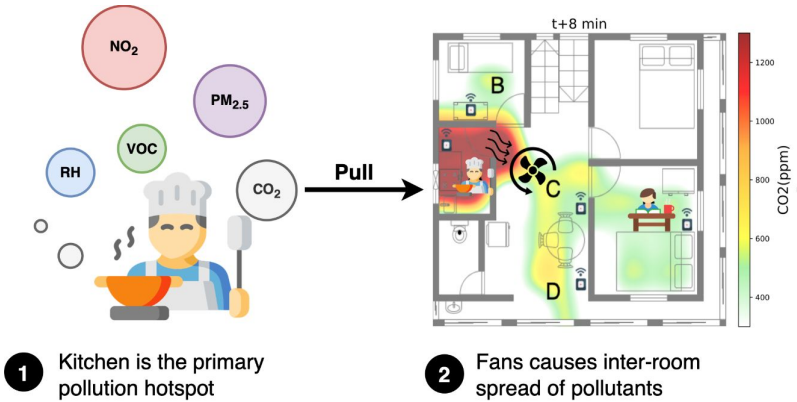
- Different sensor(location) sense the same activity at **different time**. How **Real time ?**
- Can we **exploit the location specific informations** from the different areas of the house to keep a **global context**?
- How it can help?
  - Air management strategies eg. **exhaust** on/off
  - Preventive actions eg. **fire**
- current monitoring systems lack the capability to track propagation patterns or provide timely, coordinated responses

Image sources: https://www.flaticon.com/

Air Pollution Propagation

Objective of the study

Develop distributed system architecture that transforms standalone air quality monitors into a coordinated network of air quality monitors.



How?

- ESP32-based monitoring device (DALTON[1])
- consistent global context across all monitors
  - SetCVRDT (Conflict-free Replicated Data Type)
  - **prediction updated** by each node with time. **Sequence of updates** may not be same.
- distributed coordination and decision making
  - Raft consensus algorithm
  - ESP-32 - **Limited resources**
  - Different decision making **(lagging behind)**
- ML models to predict real-time activity detection

Image sources: https://www.flaticon.com/

[1] Karmakar, P., Pradhan, S. and Chakraborty, S., 2023. Exploring Indoor Health: An In-depth Field Study on the Indoor Air Quality Dynamics. *arXiv preprint arXiv:2310.12241*.

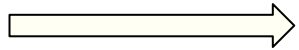
Information Flow

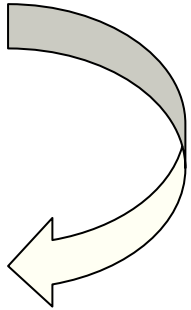
Air monitoring sensor (**DALTON**[1]) capture the Air composition.

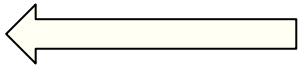
Individual nodes(ESP32 microcontroller) use local ML models to generate activity predictions based on the Air composition.

Predictions are shared across nodes using Set CvRDT module.

The Raft-elected leader processes the aggregated predictions







[1] Karmakar, P., Pradhan, S. and Chakraborty, S., 2023. Exploring Indoor Health: An In-depth Field Study on the Indoor Air Quality Dynamics. *arXiv preprint arXiv:2310.12241*.

Set Cvrdt Module

Implement

- ❏ State-based Conflict-free replicated data type system using the **set** data structure.
- ❏ Each node to receive insert and remove items requests from the **main_set** that is **stored locally**.

Requires

- ➔ When an item is inserted to the set using the /add endpoint, the item must reflect in other nodes' set too.
- ➔ When an item is removed from the set using the /rem endpoint, the item must be removed from the other nodes' set too.

What items are we storing

- ❏ Unique id(ip address + random number), Activity prediction at each sensors.

Properties

- ➔ Convergence (same state)
- ➔ Commutativity (order of operation)
- ➔ Associativity (grouping/rearranging)
- ➔ Idempotency (reinsertion, redeletion)

Testing

- ➔ Multiple Instance Deletion Problem
- ➔ Multiple Node Insert-Remove Operations
- ➔ Concurrent Reinsertion at nodes
- ➔ Multiple remove and add in parallel.

Raft

Implement

- ❏ For leader election.
- ❏ Instead of all nodes processing and analyzing the global state, the elected leader takes primary responsibility for complex decision-making.
- ❏ This is particularly important for resource-constrained ESP32 devices
- ❏ Fire → Necessary precautions

Leader:

- ➔ Sends heartbeat
- ➔ Broadcasts decisions to all follower nodes.

Node states

- ❏ Follower, Candidate, Leader

Election Process

- ➔ Node discovery by joined message.
- ➔ Voting mechanism: Self voting, Vote request broadcasting, processing

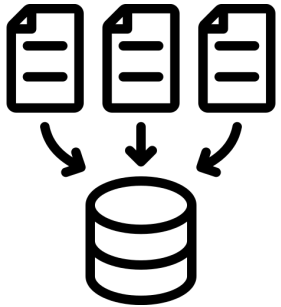
Testing and challenges

- ➔ Infinite election problem: Term number in elections with no result.
- ➔ Three node configuration
- ➔ four node configuration.
- ➔ Shutting down follower, leader and turning it on , multiple times.

Data Collection

Lab Activity Dataset [2]

- A three-month study conducted in an academic research laboratory setting
- To see the correlation between indoor activities and air composition. We can use this for precautions and decision making.



Cooking Dataset [3]

- comprising cooking type and food measurements from a kitchen in a suburban food canteen in India.
- To see whether smaller changes air composition can be easily predicted or not. We can use this for notifications (What's being cooked right now).

[2] Prasenjit Karmakar, Swadhin Pradhan, and Sandip Chakraborty. 2024. Exploiting Air Quality Monitors to Perform Indoor Surveillance: Academic Setting. In 26th International Conference on Mobile Human-Computer Interaction (MOBILEHCI Adjunct '24), September 30–October 03, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/ 3640471.3680243

[3] Sandip Chakraborty Prasenjit Karmakar, Swadhin Pradhan. Indoor air quality dataset with activities of daily living in low to middle-income communities. arXiv:2407.14501, v3, 2024.

Image sources: https://www.flaticon.com/

ML Model | ESP-32 | Emlearn Library

ML-Models used

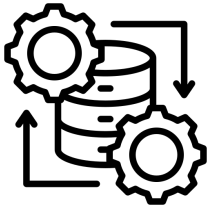
- ❏ Decision Trees, Random Forest, Extra Trees, Gaussian Naive Bayes, Neural Network (sklearn mlp)

Why Emlearn Library

- ➔ We cannot train ML-models on ESP-32
- ➔ Converts ML Models to efficient C code.
- ➔ Easy to deploy to any microcontroller with c99 compiler

Challenges

- ➔ ESP-32 supports only 16 bit integers
- ➔ floating point operations till 1 decimal place (Very inefficient).



Data Processing

- ❏ Converted the Air composition(input data) values from floating point to integral values within the 16 bit range.

Image sources: https://www.flaticon.com , https://emlearn.readthedocs.io/en/latest/source/README.html

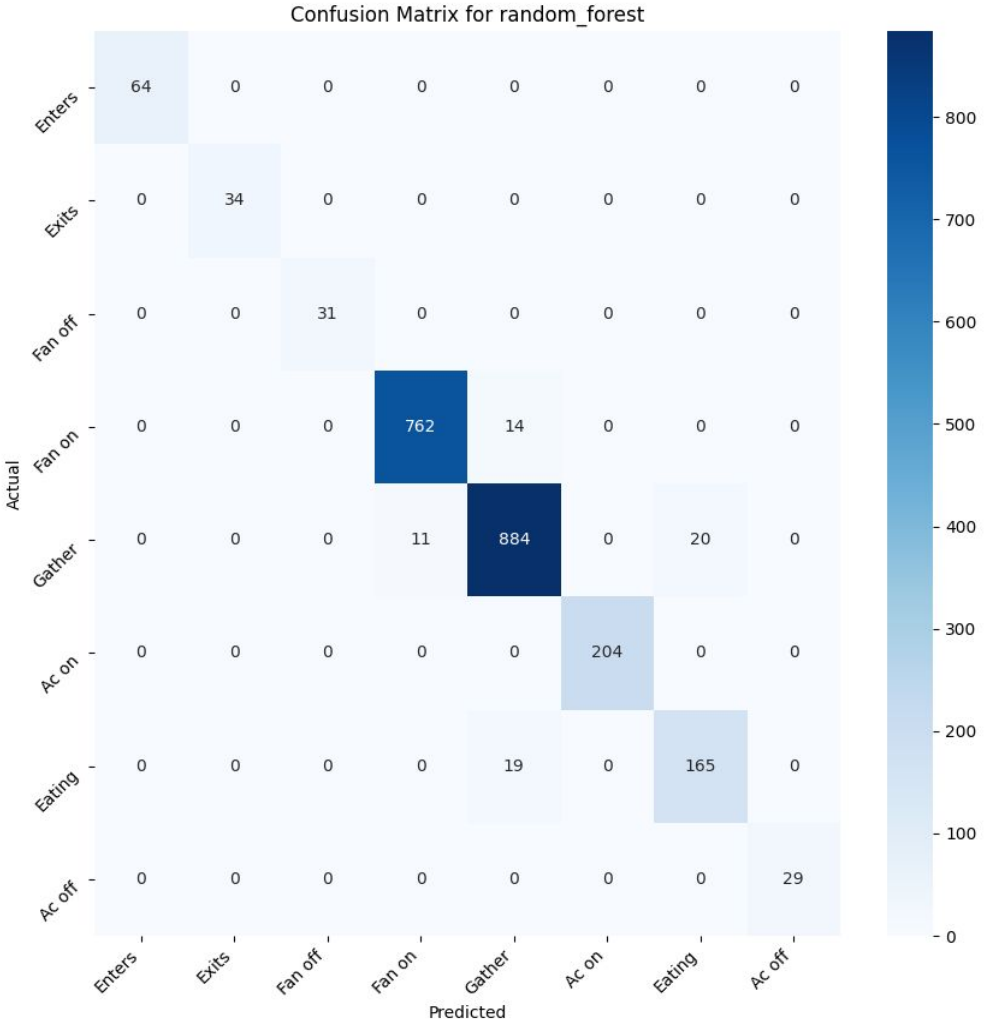
Results - Lab Activity

TABLE 6.1: Comprehensive Model Performance Analysis for Lab Activity Classification

Model	Accuracy	Avg. Time (µs)	Max Time (µs)	ESP32 Errors
Decision Tree	93.56%	9	42	0
Random Forest	97.41%	34	330	0
Extra Trees	97.36%	62	463	0
Gaussian NB	48.28%	356	382	0
Neural Network	49.45%	89	438	0

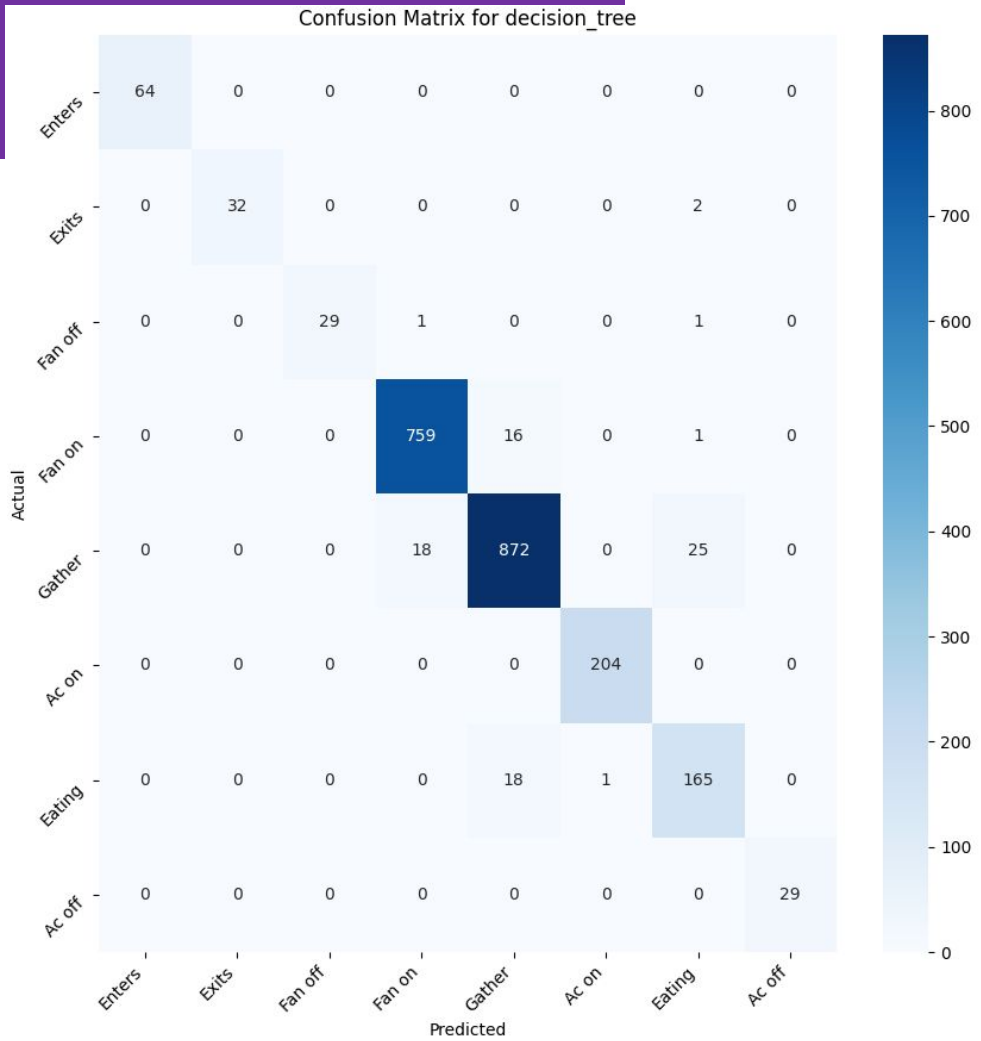
Results - Lab Activity

Confusion Matrix for
Random Forest



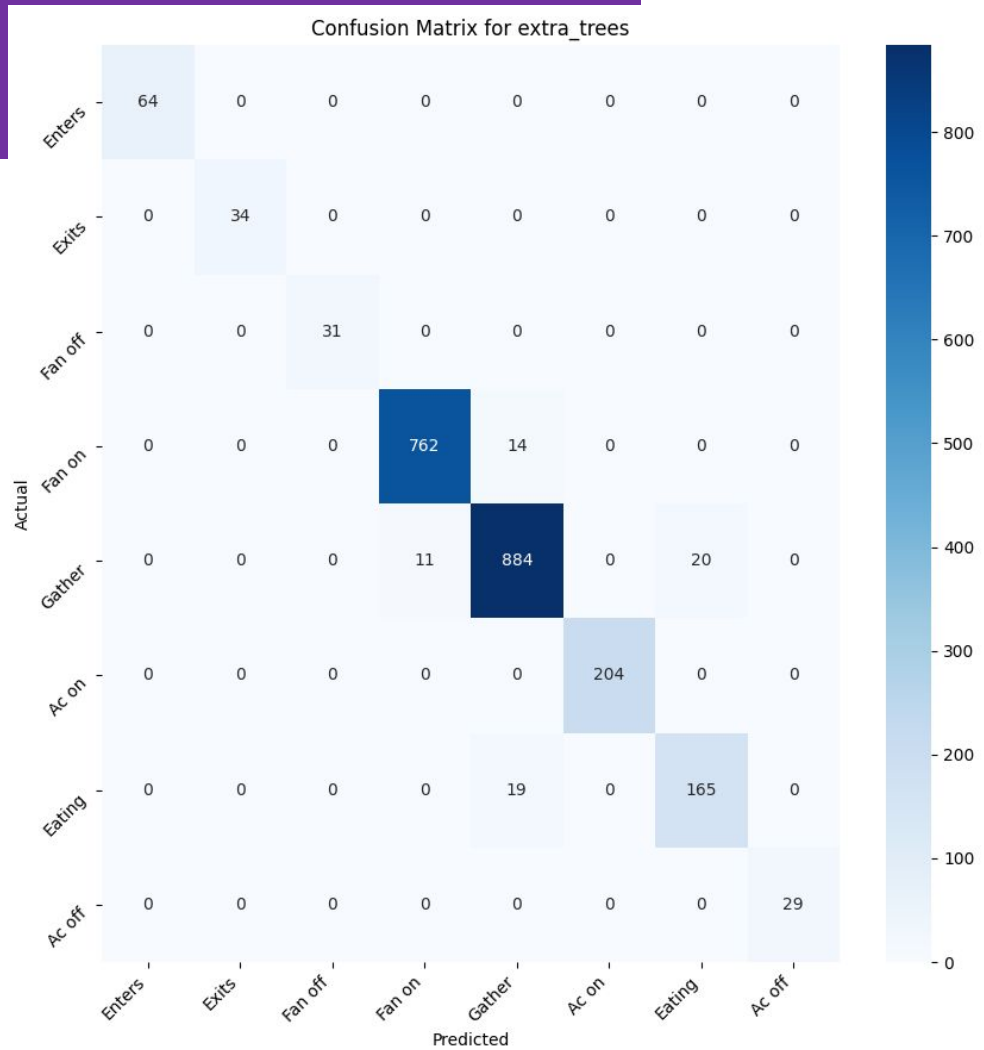
Results - Lab Activity

Confusion Matrix for
Decision Trees



Results - Lab Activity

Confusion Matrix
Extra Trees



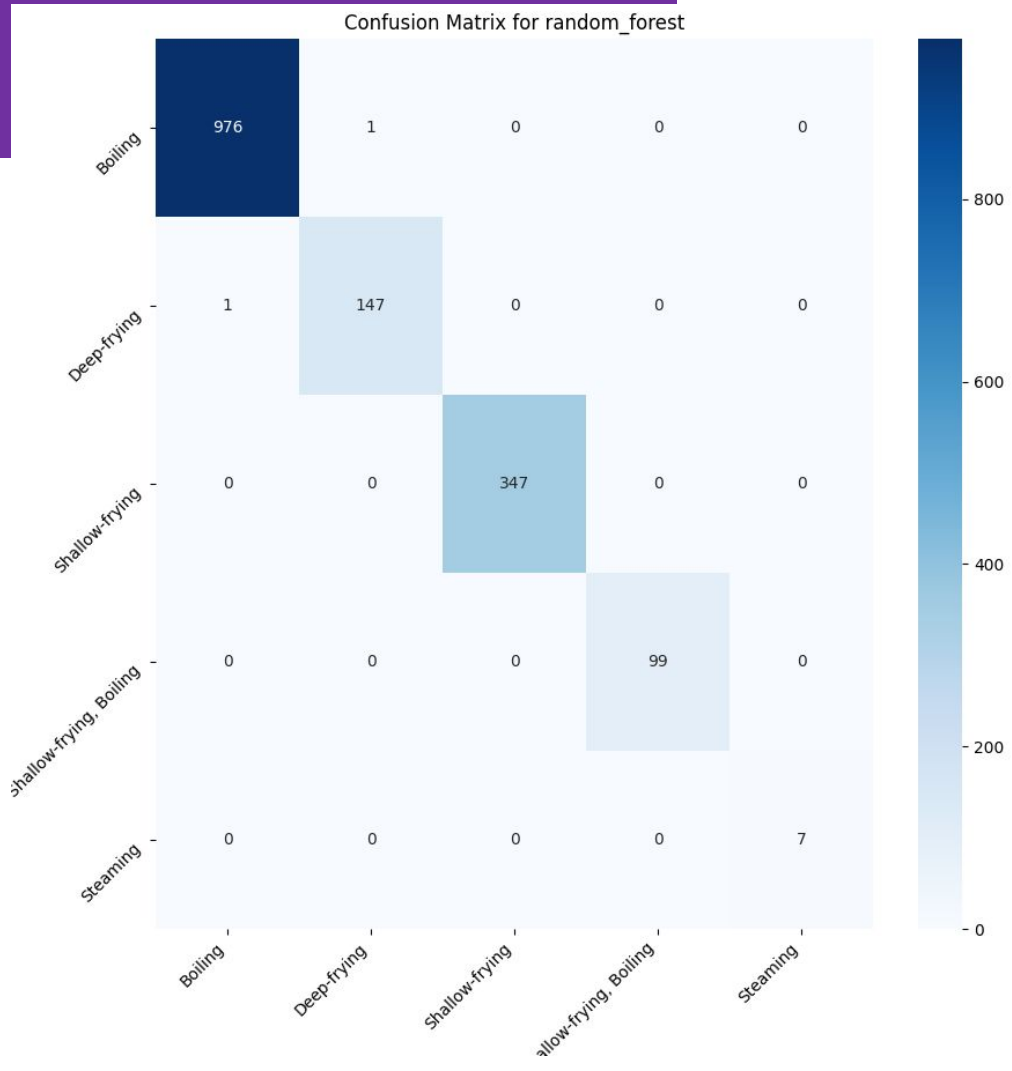
Results - Cooking Activity (Cook_Type)

TABLE 6.2: Model Performance for Cooking Type Classification

Model	Accuracy	Avg. Time	Max Time	ESP32 Errors
Decision Tree	97.28%	6 µs	39 µs	0
Random Forest	99.68%	16 µs	271 µs	0
Extra Trees	99.81%	43 µs	339 µs	0
Gaussian NB	47.72%	157 µs	181 µs	0
Neural Network	44.36%	61 µs	352 µs	0

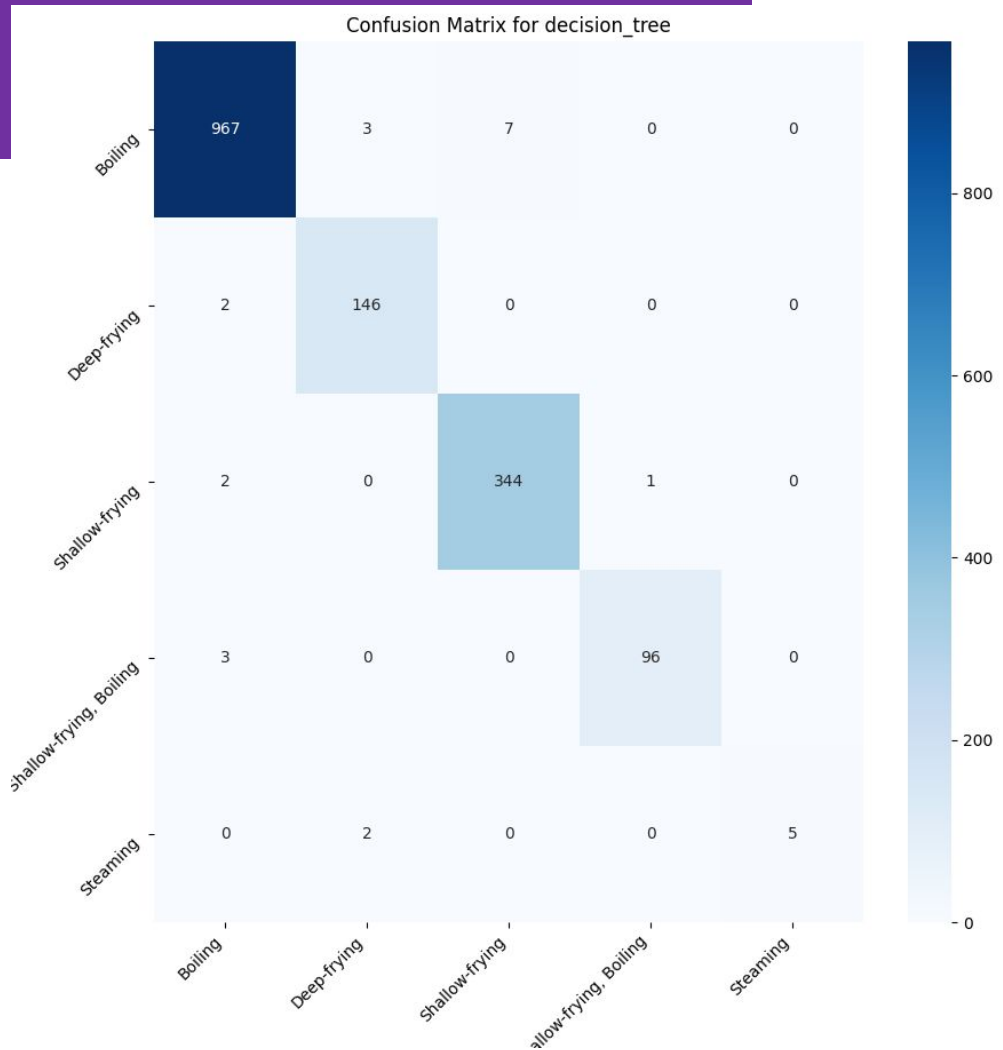
Cooking Activity

(Cook_Type)
Confusion Matrix for
Random Forest



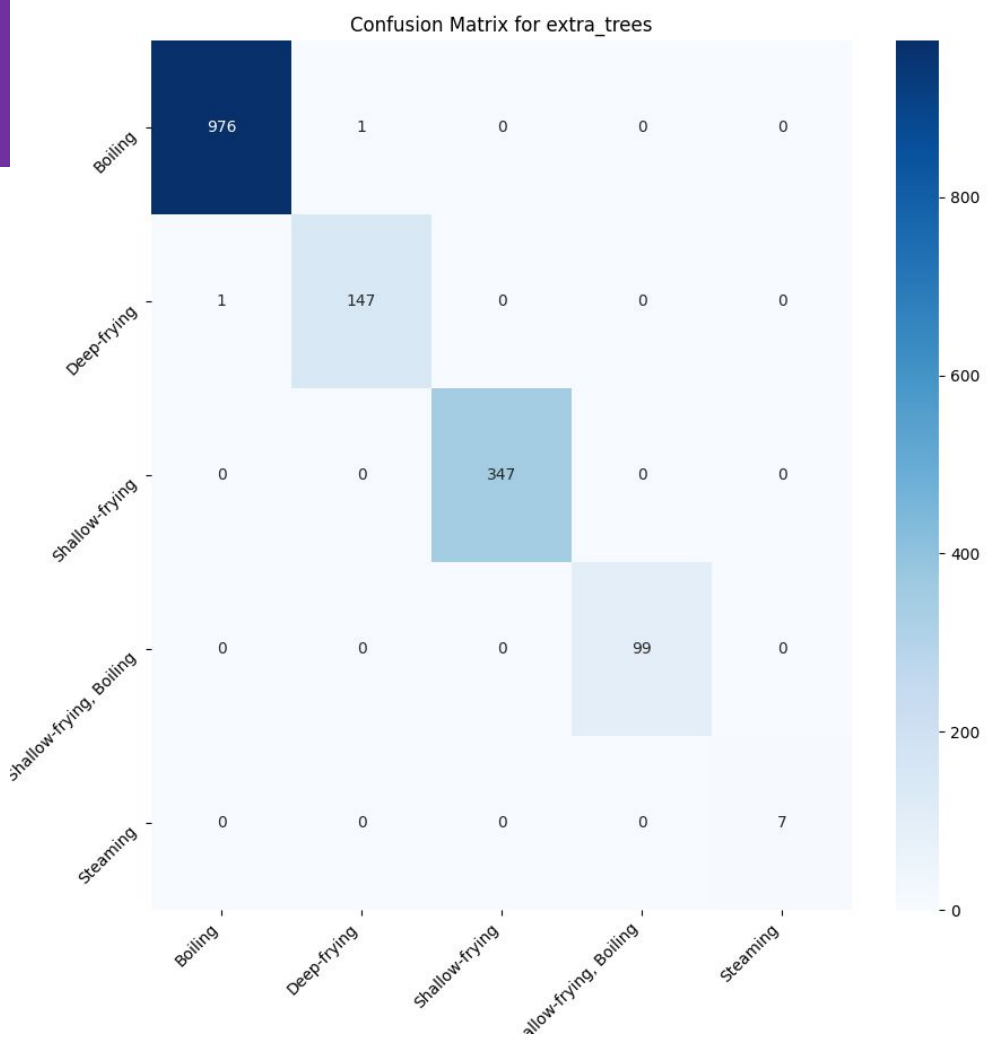
Cooking Activity

(Cook_Type)
Confusion Matrix for
Decision Trees



Cooking Activity

(Cook_Type)
Confusion Matrix for
Extra Trees



Results - Cooking Activity (Food_Type)

TABLE 6.3: Model Performance for Food Type Classification

Model	Accuracy	Avg. Time	Max Time	ESP32 Errors
Decision Tree	94.17%	7 µs	38 µs	0
Random Forest	99.49%	23 µs	289 µs	0
Extra Trees	99.68%	64 µs	311 µs	0
Gaussian NB	46.52%	178 µs	209 µs	0
Neural Network	45.71%	83 µs	391 µs	0

Cooking Activity

(Food_Type)
Confusion Matrix for
Random Forest

Cooking Activity

(Food_Type)
Confusion Matrix for
Decision Trees

Cooking Activity

(Food_Type)
Confusion Matrix for
Extra Trees

Gaussian Naive Bayes, sklearn mlp

- While previous research[2] demonstrated high accuracy with (64,64)-neuron hidden layers (achieving 97.7% accuracy), EmLearn's C code generation limitations forced us to use a reduced architecture (10,10 neurons hidden layers) for ESP32 deployment .
- Despite successful deployment, this reduced architecture showed poor performance (around 45% accuracy) due to less number of hidden layers.
- Floating-Point Operations: Both Neural Network and Gaussian NB struggled with ESP32's floating point limitations, showing significantly higher prediction times and poor accuracy. The necessary integer conversion of sensor data further impacted model performance.

[2] Prasenjit Karmakar, Swadhin Pradhan, and Sandip Chakraborty. 2024. Exploiting Air Quality Monitors to Perform Indoor Surveillance: Academic Setting. In 26th International Conference on Mobile Human-Computer Interaction (MOBILEHCI Adjunct '24), September 30–October 03, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640471.3680243

Conclusion

Successfully implemented and tested three key components that can work together to create a robust and efficient system:

- **Set CvRDT**: ensures consistent data sharing across distributed nodes.
- **Raft** Consensus Algorithm: enables reliable leader election and coordinated decision making.
- **Machine Learning Models**: successfully implemented and validated multiple machine learning models on ESP32 devices using the EmLearn library. The most efficient model, Random Forest, achieved over **97% accuracy for lab activity detection** and **99% for cooking activity detection**, while maintaining **minimal prediction times** suitable for real-time operation

Scope of Future Work

01

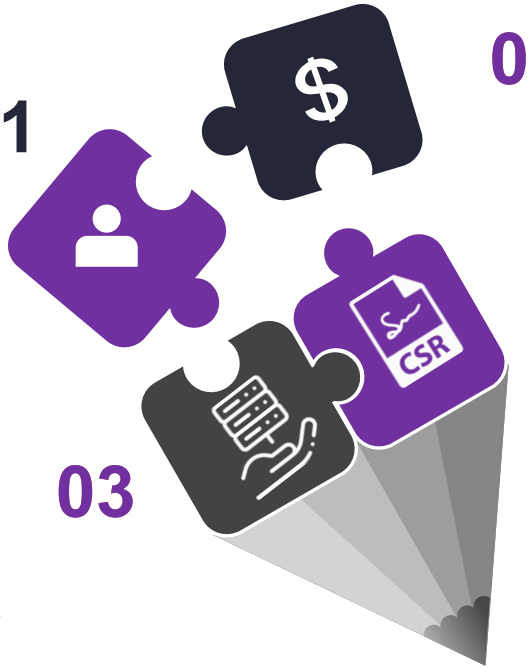
System Integration

Implementing unified codebase combining Set CvRDT, Raft, and ML models on ESP-32

02

Dataset Enhancement

Expanding the dataset to include more diverse activities. Currently lab activity dataset has only 8 activity.



03

Real-world Deployment

Testing in various residential and commercial settings. Gathering user feedback for system refinement.

04

using bluetooth

Currently we are using common wifi for communication for UDP. Low power consumption., Low latency in short range, Better for periodic transfers,etc.

THANK YOU