 **NAMAL UNIVERSITY**

**Department of Computer Science**

**BS Computer Science Program**

**6th Semester, Session 2024 Semester Project**

**CS-352 Data Analysis and Visualization**

**Data Analysis and Visualization Project Proposal**

# Dataset Source and Description

The dataset used in this project is titled "PIRvision Office Dataset 1". It is a publicly available dataset associated with the research paper "Promoting Occupancy Detection Accuracy Using On-Device Lifelong Learning." The dataset comprises time-stamped motion detection data collected from multiple Passive Infrared (PIR) sensors deployed in office environments. Each row includes sensor values, environmental features, and the ground truth label indicating whether the room was occupied or not.

The data is well-suited for classification and time-series analysis, offering rich insights into motion-based occupancy patterns.

# Research Questions

This project aims to answer the following research questions using data analysis and machine learning techniques:

1. How accurately can occupancy be detected using PIR sensor readings?

- This will involve building and evaluating supervised classification models using motion data as input.

1. How does the model performance vary over time, and is there evidence of concept drift?
   * This will be analysed by training models on earlier data and testing them on later time segments to examine if accuracy declines over time.
2. Which PIR sensors or features contribute most significantly to predicting occupancy?
   * Feature importance techniques such as those provided by tree-based models (e.g., XGBoost) or SHAPvalues will be used to interpret the model's decision-making.

# Preliminary Thoughts on Challenges and Solutions

* Challenge 1: Handling Imbalanced Classes

The dataset may have more data points for 'unoccupied' states than 'occupied,' which can bias the model. Solution: Apply resampling techniques (SMOTE, under sampling) or use metrics like F1-score and ROC-AUC instead of plain accuracy.

* Challenge 2: Temporal Splitting for Concept Drift Analysis

Splitting data chronologically may lead to insufficient training/testing coverage.

Solution: Carefully segment data using time windows (e.g., daily or hourly) and use sliding windows for training/testing.

* Challenge 3: Feature Interpretation

With many sensor inputs, interpreting the most useful ones could be difficult.

Solution: Use XGBoost's built-in feature importance and SHAP for clearer visual insights.