Title:

Variational Autoencoders for Anomaly Detection in IoT Data

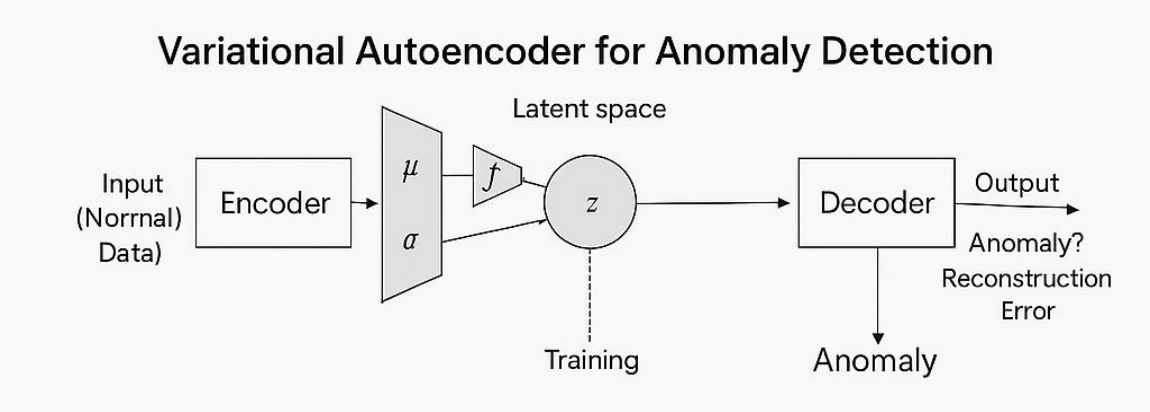
Problem Statement

IoT devices, such as sensors, collect large amounts of data every day. Sometimes, things can go wrong like sensor malfunctions, sudden attacks, or unusual behavior. Finding these unusual activities (called **anomalies**) is important to keep the system working safely and correctly.

In real life, we often **don't know what abnormal data looks like**, and manually labeling all data is difficult. Therefore, we need an intelligent method that can **learn what normal data looks like** and automatically detect anything different. This project solves the problem by using a **Variational Autoencoder (VAE)** — a deep learning model that learns from normal data and tells us if new incoming data is abnormal, based on its reconstruction accuracy.

ABSTRACT

The Internet of Things (IoT) includes smart devices that collect and send data. Sometimes, these devices may send strange or incorrect data due to problems like faults or attacks. To keep the system working well, we need to find these unusual patterns (called **anomalies**) quickly. In this project, we use a type of deep learning model called a **Variational Autoencoder** (VAE) to find such anomalies. We train the model using only the normal (correct) data. After learning, if the model has trouble recreating (reconstructing) a data point, it means the data might be abnormal. This method helps us **detect problems automatically**, even if we don't have examples of abnormal data. It's a smart and useful way to keep IoT systems safe and working properly.



Dataset Collection

For this project, we created a synthetic IoT sensor dataset that simulates real-world environmental conditions in a smart building or industrial setup. The dataset is designed to help detect anomalies that could indicate faults, unusual activity, or failures in the system.

Sensor Type	Description
Temperature	Measures ambient temperature (in °C)
Humidity	Measures moisture content in the air (in %)
Light	Indicates ambient light intensity (in lumens)
CO_2	Monitors carbon dioxide concentration (in ppm)
Noise	Measures sound level in the environment (in dB)
Motion	Detects movement using a motion sensor (binary: 0/1)
Pressure	Atmospheric pressure values (in hPa)
Vibration	Measures vibrations using an accelerometer (in m/s^2)
Label	Indicates whether the data point is normal (`0`) or anomalous (`1`)

Preprocess the Data

- Clean missing values.
- Normalize or standardize.
- Split into training (normal data only) and testing (normal + anomalous).

Variational Autoencoders

A VAE is a **generative model** that learns to compress data (like temperature, pressure, etc.) into a smaller, meaningful form (called the **latent space**) and then reconstruct it.

- **Normal data** can be compressed and reconstructed well.
- **Abnormal data (anomalies)** are reconstructed poorly, and this poor reconstruction is what helps us **identify anomalies.**

Summary of VAE Model Results:

- **Anomaly Detection Threshold**: 0.2132
- **Accuracy**: 97%
- **Precision for Normal (0)**: 1.00 → All predicted normals are actually normal.
- **Recall for Anomaly (1)**: 1.00 → It caught *every single anomaly*!
- **F1-Score for Anomaly**: 0.74 → Solid performance, especially for a small anomaly class (only 50).