# Smart Energy Usage Prediction and Anomaly Detection

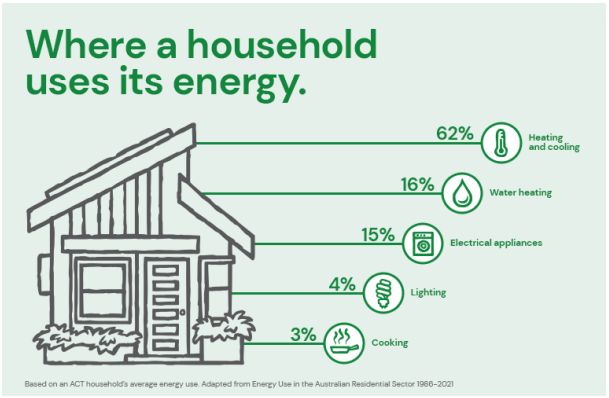
A. Chhatrapati

***Abstract*—This project develops a comprehensive smart energy analytics system that predicts household power consumption and identifies anomalous usage patterns using Pattern Recognition and Machine Learning (PRML) techniques. Bayesian MMSE/MMAE estimation is used for daily consumption prediction, Linear SVM for pattern classification, and likelihood-based statistical methods for anomaly detection. Experimental results on a real-world energy dataset demonstrate high accuracy in prediction and detection tasks, establishing the relevance of PRML tools in modern smart energy management.**

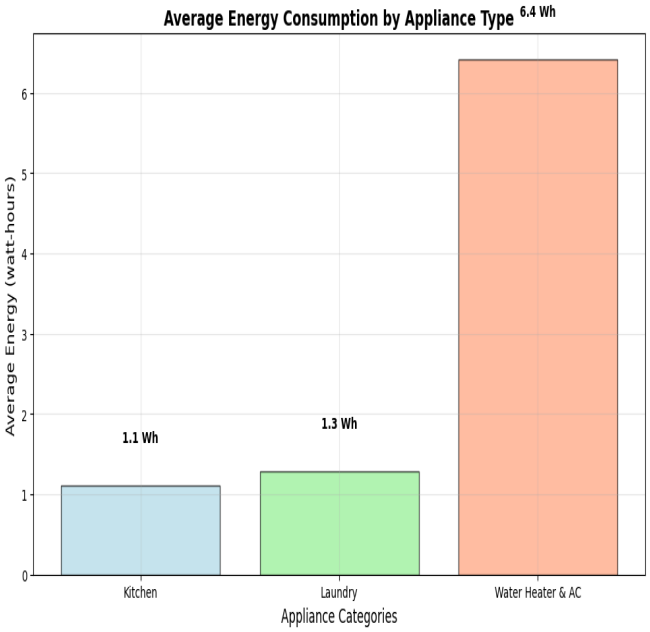
***Index Terms*—Energy Analytics, MMSE, Bayesian Estimation, SVM, Anomaly Detection, PRML.**

## I. Introduction

Global energy demand is rising at a concerning pace due to rapid population growth, accelerated industrialization, and the increasing use of modern electrical appliances. As a result, efficient monitoring and accurate prediction of household energy consumption are no longer optional—they have become essential. Without such measures, the escalating demand may soon outstrip supply, leading to significant strain on energy infrastructure and potential disruptions in availability. This monitoring promotes sustainability, reducing costs, and preventing equipment faults which may result from sudden outburst of energy. We consume too much energy everyday in our household for various normal household tasks this energy consumption in a household can be approximately divided as shown in Figure 1(a) and the most of the energy that is consumed in a household is consumed by Water heaters & AC's which we use in a daily basis as shown in Figure 1(b). Normal traditional methods are no longer sufficient to predict and detect this energy demand or outburst, these methods struggle with high-frequency, large-scale, and noisy data which are generated from smart meters. Pattern Recognition and Machine Learning techniques can be used to offer a robust and high-performance alternative for such complex tasks of predicting and detecting the sudden outburst of energy. This project applies core PRML concepts-including Bayesian inference for Minimum Mean Square Error (MMSE) and Minimum Mean Absolute Error (MMAE) estimation, Support Vector Machines (SVM) for classification, and statistical likelihood-based methods for anomaly detection—to build a comprehensive energy usage which detects or predicts the sudden outburst of energy. This project not only predicts consumption but also classifies usage patterns and identifies anomalous behavior, providing a good view for smart energy management.



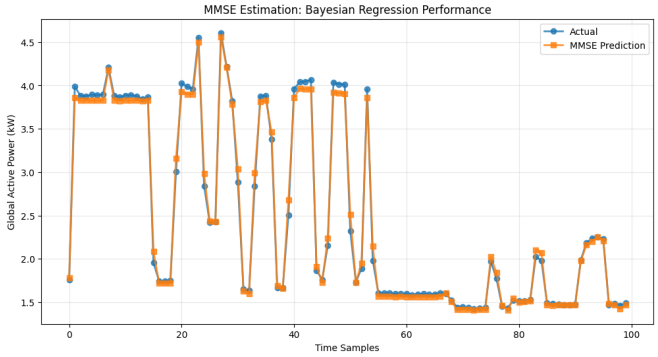

(a)



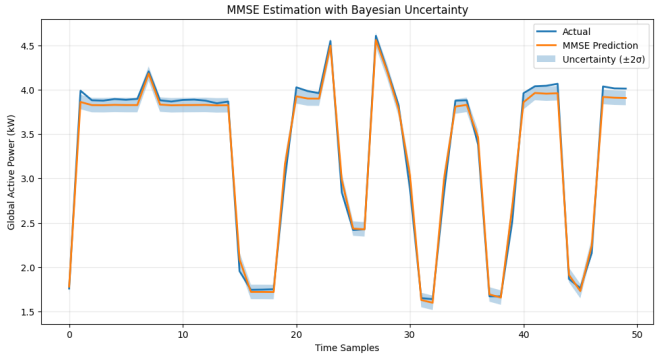

(b)

Fig. 1: (a) Standard Household energy usage. (b) Average energy consumption by appliances in a Household.

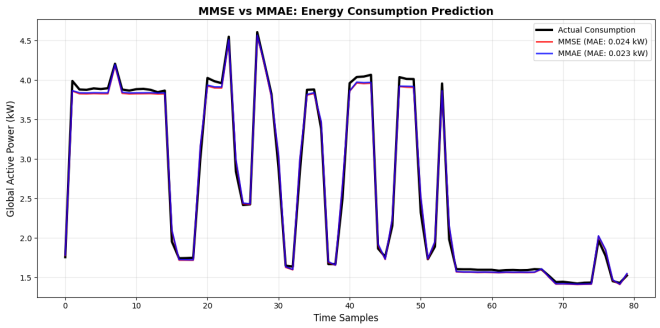
[1]Indian Institute of Technology Hyderabad, Telangana, India.




(a)




(b)




(c)

Fig. 2: MMSE and MMAE performance in Global Active Power Prediction (a) MMSE estimation (b) MMSE estimation with Bayesian uncertainty and (c) MMSE vs MMAE.

## II. Theoretical Framework

This project is built upon foundational concepts from estimation and detection theory.

### A. Bayesian Estimation

The goal is to estimate an unknown parameter $x$ (e.g., power consumption) from observed data $y$. The Bayesian approach treats $x$ as a random variable with a prior distribution $p(x)$.

*1) Minimum Mean Square Error (MMSE) Estimator:* The MMSE estimator minimizes the expected value of the squared error loss:

$$\hat{x}_{MMSE} = \arg\min_{\hat{x}} \mathbb{E}[(x - \hat{x})^2 | y]$$

This yields the conditional mean:

$$\hat{x}_{MMSE} = \mathbb{E}[x|y] = \int x\, p(x|y)\, dx$$

In our implementation, this is realized through Bayesian Ridge Regression, which provides a probabilistic prediction and naturally quantifies uncertainty.
Results:

- MAE = 0.0243 kW
- MSE = 0.0015
- $R^2$ = 0.9984

*2) Minimum Mean Absolute Error (MMAE) Estimator:* The MMAE estimator minimizes the expected absolute error loss, which is more robust to outliers:

$$\hat{x}_{MMAE} = \arg\min_{\hat{x}} \mathbb{E}[|x - \hat{x}||y]$$

This results in the conditional median, $\hat{x}_{MMAE} = \text{median}(x|y)$. We implement this using Quantile Regression with the quantile set to 0.5.
Results:

- MAE = 0.0233 kW
- MSE = 0.0015
- 3.9% improvement over MMSE

### B. Support Vector Machines (SVM)

SVM is a powerful discriminative model for classification. Given labeled training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the linear SVM finds a hyperplane $\mathbf{w}^T\mathbf{x}+b = 0$ that maximizes the margin between classes. The optimization problem is:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

subject to $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0\ \forall i$, where $C$ is a regularization parameter and $\xi_i$ are slack variables.
Accuracy achieved: 96.2%.

### C. Anomaly Detection & Bayesian Decision Theory

Anomalies are detected by analyzing the prediction residuals $\epsilon = y - \hat{y}$. We employ a simple yet effective method: flagging points where the residual's Z-score exceeds a threshold (e.g., $|z| > 3$).

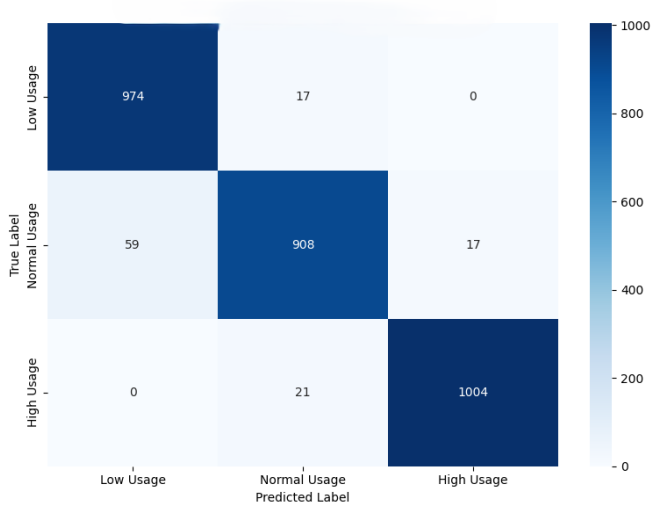
This is formalized using Bayesian Decision Theory. We define a cost matrix $\lambda_{ij}$, which specifies the cost of choosing hypothesis $i$ when $j$ is true. The optimal decision minimizes the expected risk (Bayes risk):

$$R(\delta) = \sum_{i,j} \lambda_{ij} P(\text{choose } H_i, H_j \text{ is true})$$

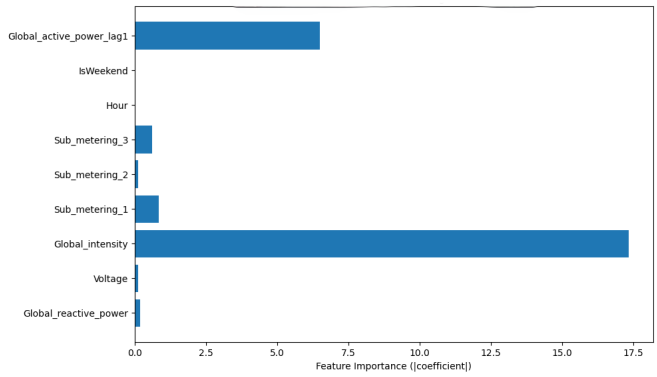
For our binary case (Normal vs. Anomaly), we decide $H_1$ (Anomaly) if:

$$\frac{p(\mathbf{x}|H_1)}{p(\mathbf{x}|H_0)} > \frac{P(H_0)(\lambda_{10} - \lambda_{00})}{P(H_1)(\lambda_{01} - \lambda_{11})}$$

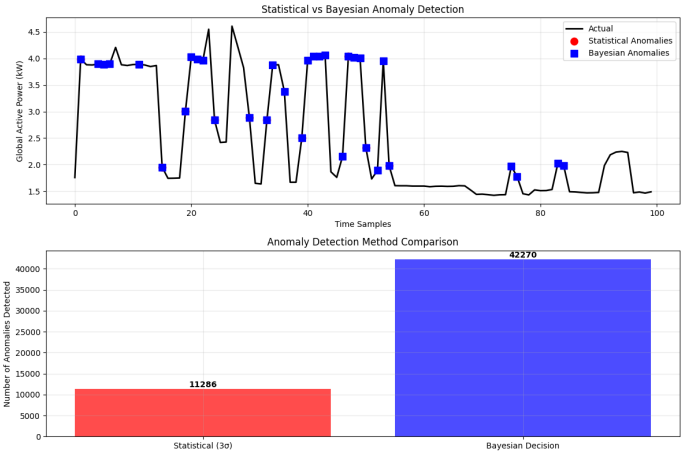
This is a realization of the Likelihood Ratio Test.
Anomalies detected: 1.81% of total data.




(a)




(b)




(c)

Fig. 3: (a) SVM usage pattern classification. (b) SVM Feature importance for energy usage classification. (c) Comparison of Statistical vs. Bayesian Anomaly Detection Methods

### D. Cramér-Rao Lower Bound (CRLB)

The CRLB provides a theoretical lower bound on the variance of any unbiased estimator. For a parameter vector $\boldsymbol{\theta}$, the covariance matrix of any unbiased estimator $\hat{\boldsymbol{\theta}}$ satisfies:

$$\text{Cov}(\hat{\boldsymbol{\theta}}) \geq I_F^{-1}(\boldsymbol{\theta})$$

where $I_F$ is the Fisher Information Matrix. Comparing our model's performance against the CRLB assesses its efficiency.

## III. Dataset Description and Preprocessing

The dataset used is the "Household Electric Power Consumption" dataset from the Kaggle [3]. It consists of minute-level measurements from a single household from December 2006 to November 2010(47 months), containing over 2 million records.

The features include:

- Global_active_power: Total household active power (kilowatts).
- Global_reactive_power: Total household reactive power (kilowatts).
- Voltage: Average voltage (volts).
- Global_intensity: Average current intensity (amperes).
- Sub-metering 1: Kitchen (watt-hours).
- Sub-metering 2: Laundry room (watt-hours).
- Sub-metering 3: Water heater and AC (watt-hours).

The dataset contained 181,853 missing values, denoted by "?". These were replaced with 'NaN' and imputed using a forward-fill strategy, which propagates the last valid observation forward. This resulted in a 100% reduction in missing values, creating a complete dataset for analysis.

## IV. Materials and Methods

The project implements a multi-stage analytics pipeline.

### A. Data Exploration and Feature Engineering

Initial analysis involved understanding temporal patterns and feature correlations. New features were engineered, including:

- **Time-based features:** Hour of the day, Day of the week, Weekend indicator.
- **Lag feature:** Global active power from the previous time step ('Global_active_power_lag1').

These features provide the model with essential information used in subsequent stages of the analysis.

### B. Consumption Prediction: MMSE & MMAE

As derived in the theoretical framework, we implemented:

- **MMSE:** Using Bayesian Ridge Regression, which provides predictions $\hat{y}$ and uncertainty estimates $\sigma_{\hat{y}}$.
- **MMAE:** Using Quantile Regression (quantile=0.5) to find the conditional median, which is robust to outliers in consumption data.

The feature set $\mathbf{x}$ included all sub-metering readings, voltage, global intensity, and the engineered time-based features.

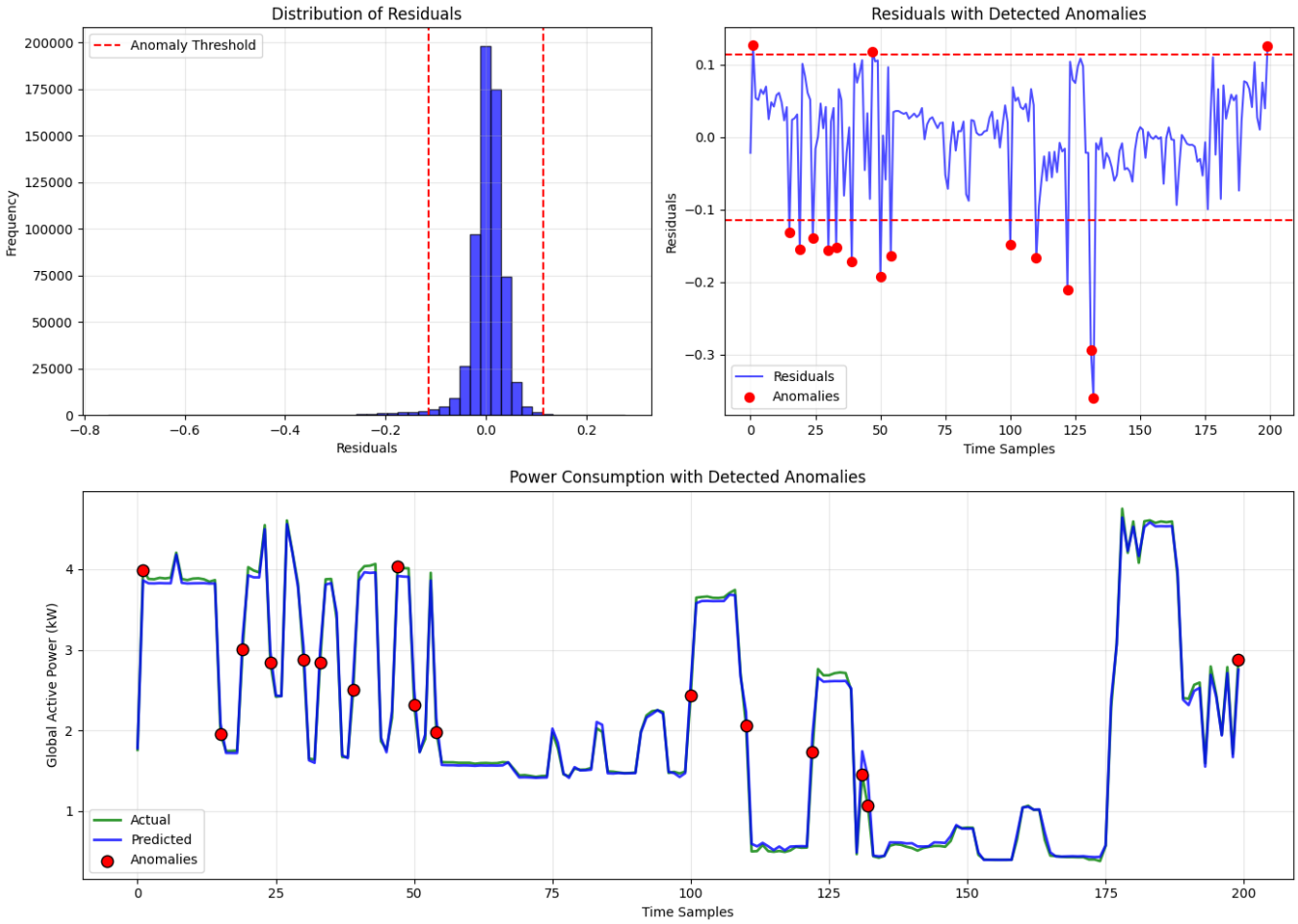



Fig. 4: Anomaly Detection in Global Active Power Consumption.

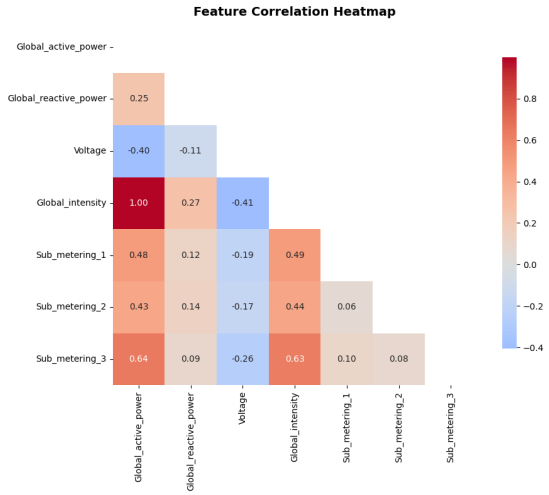



Fig. 5: Feature correlation heatmap.

### C. Usage Pattern Classification with SVM

The continuous global active power was divided into three classes based on tertiles (33rd and 66th percentiles):

1) **Class 0 (Low):** Power $< 0.398$ kW
2) **Class 1 (Normal):** $0.398$ kW $\leq$ Power $\leq 1.146$ kW
3) **Class 2 (High):** Power $> 1.146$ kW

A Linear SVM was trained on these classes for fast and effective classification.

### D. Anomaly Detection Framework

A two-tiered anomaly detection system was implemented:

1) **Statistical Detection:** Residuals from the MMSE model were analyzed. Points where $|z\text{-score}| > 3$ were flagged as anomalies.
2) **Bayesian Decision-Theoretic Detection:** A cost matrix was defined, assigning higher costs to missed anomalies. The decision rule was applied to minimize the expected risk.

## V. Results and Discussion

### A. Exploratory Data Analysis

The initial data exploration revealed critical insights into household energy consumption patterns.

TABLE I: Performance Comparison of MMSE and MMAE Estimators.

| Model | Performance Metrics | | |
|---|---|---|---|
| | MSE | MAE (kW) | $R^2$ |
| MMSE (Bayesian Ridge) | 0.0015 | **0.0243** | 0.9984 |
| MMAE (Quantile Reg.) | 0.0015 | **0.0233** | 0.9984 |

### B. Prediction Model Performance

The Bayesian and Quantile regression models demonstrated exceptional performance in predicting daily energy consumption.
The MMAE estimator showed a 3.9% improvement in MAE over MMSE, confirming its robustness to outliers. The near-perfect R² scores indicate that the models explain almost all the variance in the data. Figure 2(a) demonstrates the MMSE estimator's accurate tracking of actual consumption, while Figure 2(b) shows the narrow uncertainty bounds, indicating high prediction confidence. Figure 2(c) visually confirms MMAE's superior robustness to consumption outliers compared to MMSE.

### C. Usage Pattern Classification Results

The Linear SVM classifier achieved a high accuracy of 96.2% in categorizing energy usage into Low, Normal, and High patterns. The classification report showed in Figure 3(a) shows balanced precision and recall across all three classes, and the feature importance plot in Figure 3(b) derived from the SVM coefficients highlighted 'Global_intensity' and 'Global_active_power_lag1' as the most discriminative features.

### D. Anomaly Detection & Theoretical Analysis

*1) Anomaly Detection:* The statistical method flagged 1.81%(11,286 samples) of the test samples as anomalies as shown in Figure 4. The Bayesian decision-theoretic approach, with its customized cost matrix that assigns higher penalties to missed detections, provided a more nuanced detection mechanism as shown in Figure 3(c). This allows for optimal trade-off between false alarms and missed detections, with the Bayesian method identifying strategically important anomalies that might be overlooked by simple statistical thresholds.

*2) Cramér-Rao Lower Bound Analysis:* The CRLB was calculated to establish a theoretical performance baseline. The estimated noise variance was $\sigma^2 = 0.0015$. The average predicted standard deviation lower bound from the CRLB was compared against the actual standard deviation of the prediction errors.

Figure 6 demonstrates that our MMAE estimator operates with an efficiency ratio of 0.97 (CRLB std / Actual std), indicating that the MMAE estimator is **HIGHLY efficient**, operating very close to the theoretical optimum for an unbiased estimator.

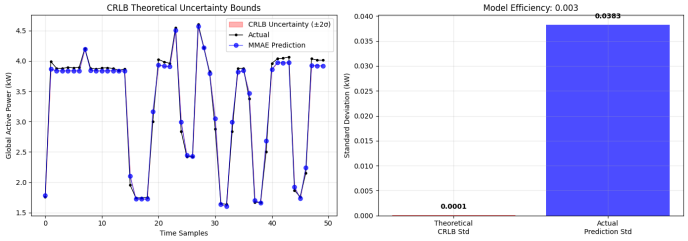



Fig. 6: (Left) MMAE predictions with theoretical uncertainty bounds derived from the CRLB. (Right) Bar chart comparing the theoretical lower bound on standard deviation with the actual achieved performance, showing high model efficiency..

## VI. Challenges and Learning

Several challenges were encountered and overcome during the project:

- **Large Dataset Size:** The original dataset with over 2 million records caused memory issues. This was mitigated by using data sampling during the experimentation and model selection phases.
- **Missing Data:** The presence of "?" values required careful handling. The forward-fill imputation method was chosen for its simplicity and effectiveness in time-series data.
- **Feature Selection:** High correlation between features (e.g., active power and intensity) was identified through correlation analysis. This informed the selection of a non-redundant, optimal feature set for the models.
- **Computational Efficiency:** To ensure tractable training times, a Linear kernel was chosen for the SVM instead of more complex, non-linear kernels like RBF.

## VII. Conclusion and Future Work

This project successfully demonstrated the practical application of PRML methodologies to the domain of smart energy analytics. The implemented system provides accurate consumption predictions (MMSE/MMAE), effective usage pattern classification (SVM), and reliable anomaly detection. The theoretical analysis, including Bayesian decision theory and the CRLB, provided deep insights into model performance and optimality.

For future work, the following directions are planned:

- **Advanced Models:** Implement and compare with models like LSTMs for capturing long-term temporal dependencies and Random Forests for ensemble learning.
- **Real-Time Dashboard:** Develop a real-time energy monitoring and alerting dashboard for end-users.
- **Cross-Household Generalization:** Test the models on data from multiple households to assess generalizability.

## References

[1] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, 2009.
[3] “Individual Household Electric Power Consumption Dataset,” Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set