Chapter 3

Data Acquisition and Processing

This chapter focuses on the DSs used in the thesis Section. Section 3.1: We provide an introduction on the EM 3.1. Section 3.2: A smart building dataset description type -1 located in Vienna 3.2. Section 3.3: Dataset description of smart building located in Berkley, California 3.3. Section 3.3.1: Features present in the dataset type -2 3.3.1. Section 3.3.2: The pre-processing of DSs before feeding into ML 3.3.2. Section 3.3.3: Describe the feature extraction from the DSs 3.3.3. Section 3.4: This section explain the performance measures and simulation working environment 3.4.

Buildings account for 40% of global energy, which can grow to 50% by 2050. Increasing energy prices call for more intelligent solutions, prompting studies into AI and IoT for energy prediction. The current research introduces a hybrid TCN-GRU DL model for forecasting hourly energy consumption in SBs. Blending temporal neural networks (TCN) for spatial feature extraction with GRUs for temporal processing, the model attains 98% accuracy a better performance than current alternatives. This innovation allows for accurate EM, lowering costs and waste and advancing sustainability objectives. The outcomes illustrate the revolutionary impact of AI in maximizing building energy efficiency in the face of increasing global demand.

3.1 Introduction

In EM research, the quality and preparation of the dataset are crucial for developing accurate and reliable predictive models. High-quality data capturing the temporal dynamics of EC allows for better understanding and modeling of energy usage patterns in SBs. This chapter provides a detailed overview of the dataset used in this study, including its characteristics and the preprocessing steps undertaken to ensure data suitability for forecasting models.

3.2 Dataset Description -1

The dataset used in this study encompasses hourly measurements of EC collected over the course of a year, specifically from January 1, 2013, to December 31, 2013 [7]. The measurements were recorded in a 200-square-meter office space located in Vienna. The data includes timestamps in the format "DD-MM-YYYY HH:MM" (e.g., "01-01-2013 00:00"), indicating the exact date and time of each energy reading. This temporal information is critical for analyzing daily, weekly, and seasonal consumption patterns, as well as distinguishing between weekdays and weekends.

The energy consumption measurement data captures various zones within the office, such as open-plan areas, semi-closed offices, and fully enclosed rooms, allowing detailed analysis of intra-building energy dynamics as shown in table 3.1 . For example, the dataset features energy usage are "EC in open plan office area [o1_1]", "EC in semi-closed office [o2] and [o4]" and "EC in closed office [o3]" at each timestamp, providing comprehensive insights into occupant behavior and environmental influences shown in figure 3.1.

3.3 Dataset Description-2

The **LBNL-ETA dataset**, originally referenced by Luo et al. (2022) [21], which offers a rich compilation of whole-building and end-use ECdata. The dataset includes

TABLE 3.1: Sample Characteristics and Measurement Details

Measurement(s)	Occupancy • Room temperature ambient air • Humidity • Radiation • Temperature of air • Atmospheric wind speed • Atmospheric wind direction • Electrical energy
Technology Type(s)	Sensor • Gauge or Meter Device
Sample Characteristic - Organism	Homo sapiens
Sample Characteristic - Environment	Office building
Sample Characteristic - Location	Vienna

detailed records of **occupancy levels, HVAC system operational states**, and both **indoor and outdoor environmental conditions**. Data was collected over a span of **36 months** from two office floors covering approximately **2,325 square meters**, using nearly **300 sensors** strategically deployed throughout the building. This comprehensive dataset provides deep insights into how energy is consumed across various building systems, particularly focusing on **HVAC**, **lighting**, and **electrical infrastructure**.

1) HVAC System:

The office spaces are equipped with an **Underfloor Air Distribution (UFAD)** system, which delivers heating and cooling through four **rooftop units (RTUs)** that condition air for the underfloor plenum. These RTUs serve different sections of both the ground and second floors. During the 2018–2020 data collection period, two HVAC control strategies were in place: a **traditional rule-based approach** and a **model-predictive control (MPC)** system. The rule-based method followed predefined zone temperature schedules—for instance, setting specific damper positions and temperature setpoints for unoccupied periods (e.g., Saturdays). In contrast, the MPC strategy dynamically optimized fan speed and supply air temperature based on real-time system states and anticipated disturbances, enhancing both energy efficiency and responsiveness.

2) Lighting System:

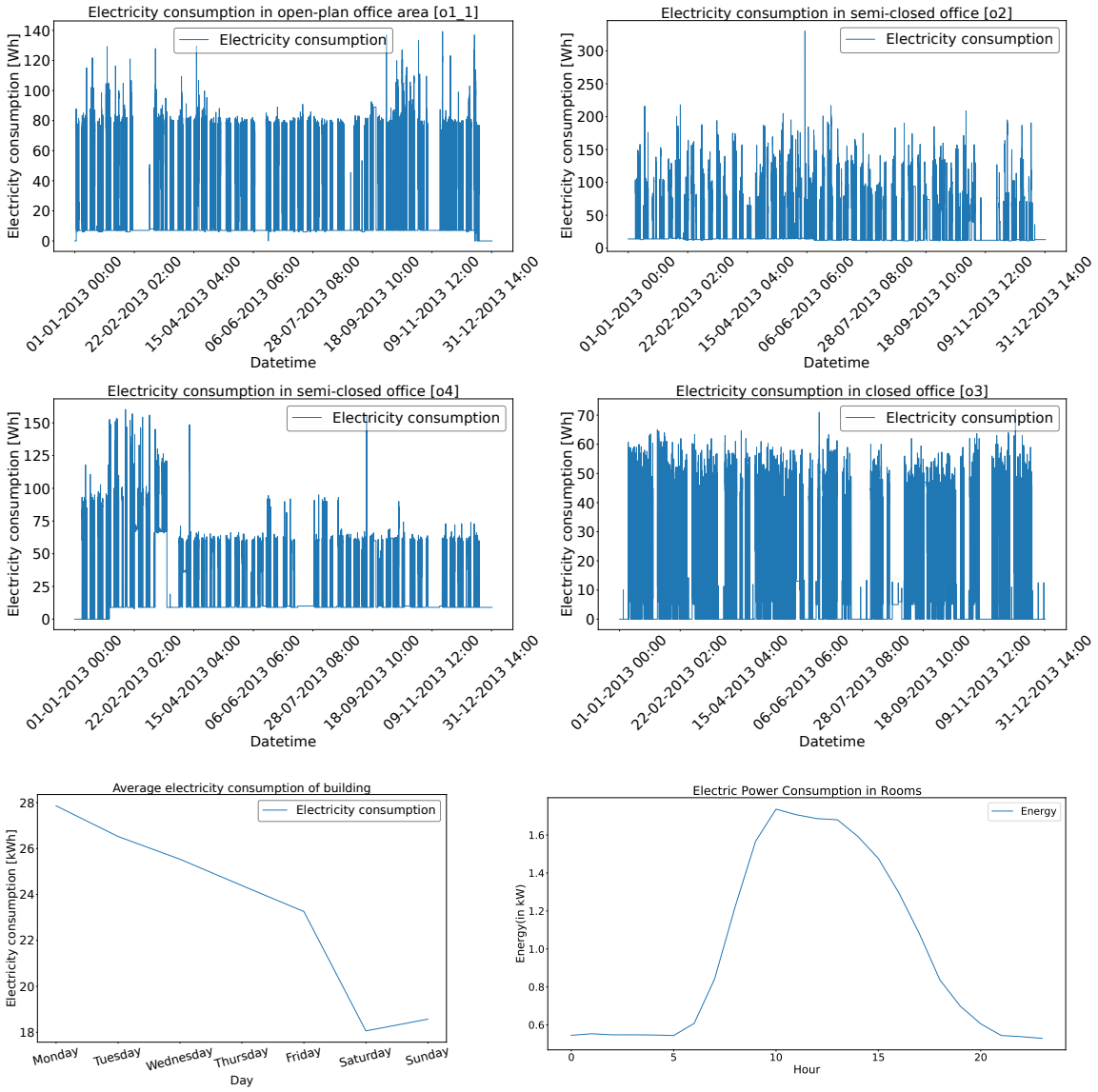


FIGURE 3.1: EC of office rooms: an open-plan (a) single-occupancy semi-closed (b and c) and closed (d). EC in (e) day and (f) hour wise.

The building features an intelligent lighting infrastructure comprising **Philips lighting fixtures**, a **dedicated management server**, and a **user workstation**. Occupancy sensors (photocells) are placed across workplace zones to automate the switching of lights based on presence detection. **Manual roller shades** installed on windows further support natural lighting control in offices and conference rooms. Energy consumption from the lighting system is monitored and managed through a browser-based **Graphical User Interface (GUI)** using **Lutron Quantum Vue**

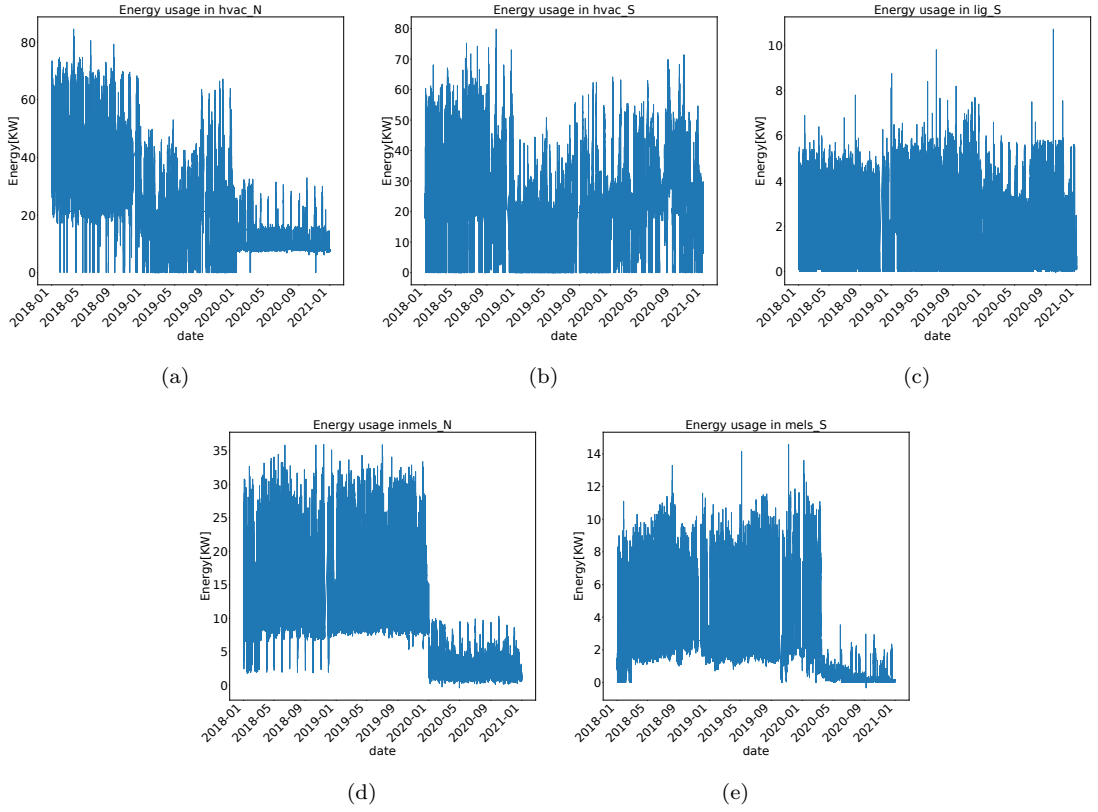


FIGURE 3.2: Individuals EC consumption for: (a) HVAC_N and (b) HVAC_S load, (c) lig_S load, (d) mels_N, and (e) mels_N.

(Lutron, 2017), enabling precise zone-based lighting control.

3) Electrical System:

Electricity is distributed through two main transformers—one supplying power to office spaces and the other to HVAC systems. The office switchboards deliver electricity to common areas, rooms, and individual panels, while the HVAC-related switchboard powers mechanical systems like air conditioning units and geysers. To ensure accurate monitoring and benchmarking, **panel-level energy measurements** are conducted by the **National Energy Research Scientific Computing Center (NERSC)**. Additionally, **General Electric trip units** are employed to monitor consumption in both HVAC and lighting systems.

3.3.1 Preliminary Analysis

The multimodal conditions of an office building of Berkely refer to the diverse environmental and operational factors that influence its energy consumption and occupant comfort. These conditions are typically monitored through a network of sensors that collect time-series data, allowing for a comprehensive analysis of the building's performance. Key multimodal conditions are shown in fig 3.2 and below:

1. **Solar Radiation:** Sunlight entering the building plays a significant role in indoor temperature and lighting demand. By analyzing solar radiation patterns, building systems can optimize the use of blinds and adjust HVAC operations to reduce unnecessary energy use.

2. **Occupant Count:** The number of people in a space directly influences how much energy is needed for heating, cooling, and lighting. Real-time occupancy data enables smart systems to dynamically adjust environmental controls, ensuring comfort while avoiding energy waste when areas are unoccupied.

3. **Indoor Temperature:** Maintaining a consistent and comfortable indoor temperature is essential for productivity and wellbeing. Sensors placed throughout the building help monitor fluctuations and inform real-time HVAC adjustments.

4. CO_2 **Concentration:** High levels of carbon dioxide can signal poor ventilation, which affects air quality and comfort. Monitoring CO_2 helps maintain a healthy indoor environment by triggering adjustments in airflow and ventilation systems when needed.

5. **Outdoor Temperature:** External weather conditions heavily influence how hard HVAC systems need to work. By factoring in outdoor temperature, buildings can anticipate changes and adjust heating or cooling efforts proactively, improving energy efficiency.

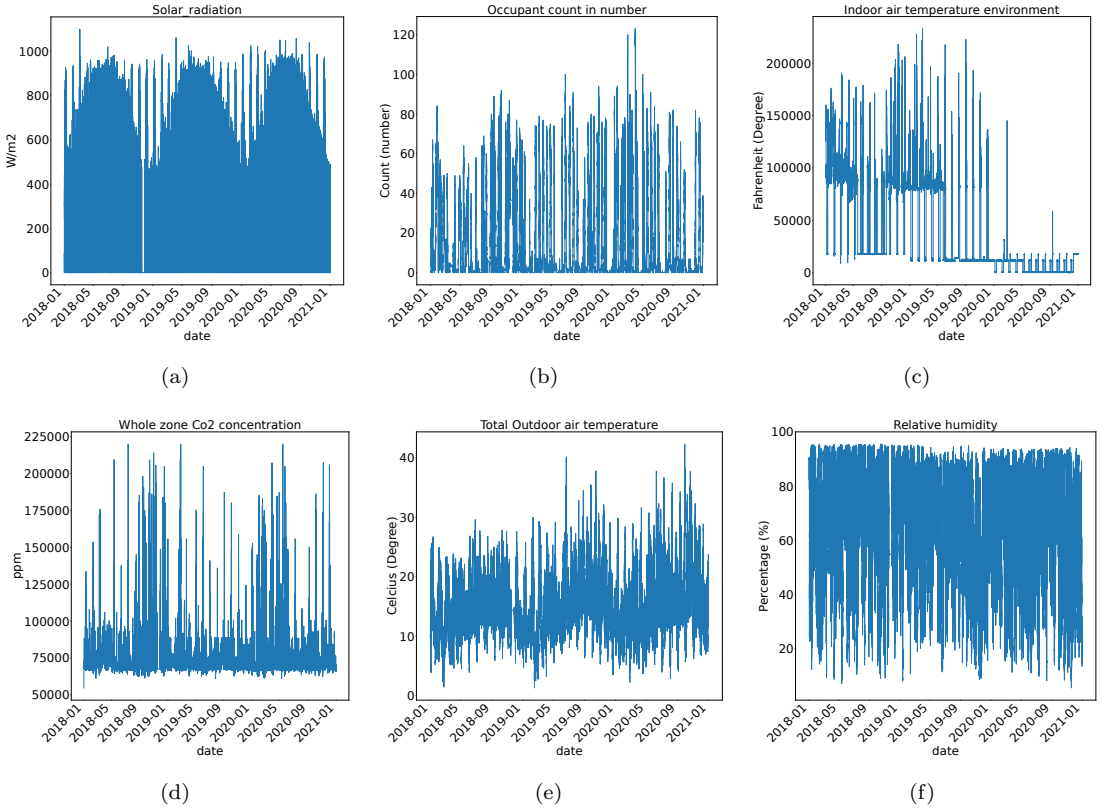


FIGURE 3.3: Mutlimodal conditions of office building: (a) Solar radiation, (b) Occupant count, (c) Indoor temperature, (d) CO_2, (f) Outdoor temperature, and (e) Relative humidity.

6. **Relative Humidity:** Both too much and too little humidity can impact comfort and building health. Sensors monitor humidity to help maintain appropriate levels, preventing issues like mold growth or dry air discomfort.

By combining these multimodal data streams, modern building management systems can take a holistic approach to EF. Advanced predictive models, such as bi-directional LSTM networks, can capture the complex relationships between these variables. This enables more accurate energy usage predictions, allowing buildings to operate more efficiently while maintaining a comfortable environment for occupants.

Before building the forecasting framework, an in-depth preliminary analysis was conducted to understand the characteristics of the EC data and to identify key patterns, inconsistencies, and relevant features.

The dataset spans a three-year period, during which EC values were recorded at 15-minute intervals. To ensure consistency and facilitate analysis, the data was aggregated in multiple ways—daily, weekly, and hourly. This multi-level resampling helped uncover underlying trends and anomalies.

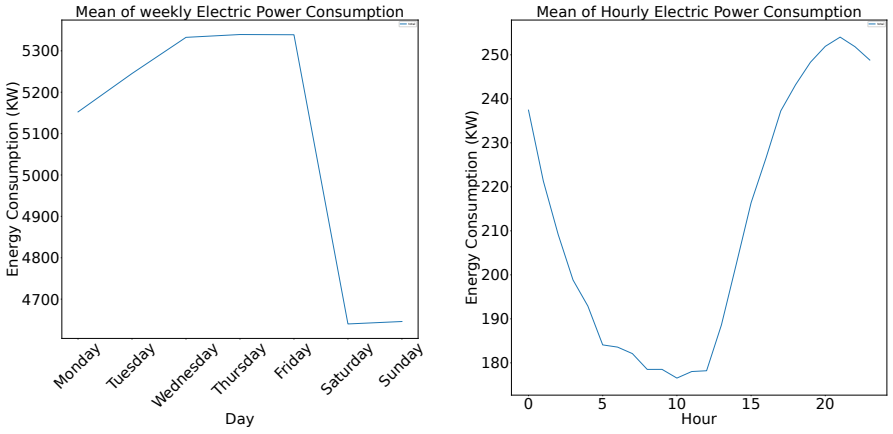
Figure 3.2 shows the overall EC trend over the dataset period, compiled by combining all available load types. Additionally, individual systems such as HVAC (North and South), MELs (North and South), and lighting (South) were visualized separately by resampling their 15-minute intervals into daily aggregates. These visualizations helped identify specific consumption behaviors across different subsystems.

One notable observation was the presence of sudden drops in energy usage, which were attributed to gaps or missing values in the dataset. These missing periods required pre-processing to clean the data and avoid skewed predictions. The missing values were handled through appropriate imputation techniques and resampling methods.

Figure 3.4 (a) illustrates the average daily EC pattern across weeks, highlighting consistent reductions in usage during weekends. This clearly indicated that the type of day (weekday vs. weekend) plays a significant role in influencing energy demand. Similarly, Figure 3.4(b) presents the average hourly consumption across days, where sharp increases during business hours were observed, particularly on weekdays. This insight led to the inclusion of hour of the day as a key feature in the model.

To prepare the data for prediction tasks, the following temporal and contextual features were engineered:

Hourly Aggregation: The original 15-minute data was aggregated to hourly intervals to align with lag requirements and reduce noise.



(a) (b)

FIGURE 3.4: EC on real values in a **??** day and (b) hour wise.

Minute of the Day (Hour Index): Electricity demand follows a strong daily pattern; each hour was encoded from 0 (12 AM) to 23 (11 PM) to capture these cycles.

Day of the Week: A clear difference between weekday and weekend consumption prompted encoding days as integers (0 for Monday through 6 for Sunday) to capture weekly periodicity.

This preliminary analysis helped shape the feature selection process and revealed the importance of temporal regularity and external contextual variables in energy consumption behavior. These insights guided the design of the forecasting model, ensuring it was aligned with real-world patterns and operational dynamics of SBs.

3.3.2 Data Preprocessing

Before building an accurate forecasting model, it was important to clean and prepare the dataset to ensure consistency and reliability. While most of the EC data was well structured, a few challenges emerged specifically, the presence of missing (null) and negative values that needed to be addressed.

Handling Missing Values: The dataset included a total of 3,180 missing entries across various energy load categories from both the North (N) and South (S) wings of the building. Specifically, there were:

- 38 nulls in miscellaneous electric loads (*mels_S*),
- 24 nulls in *mels_N*,
- 34 nulls in lighting loads (*lig_S*),
- and a significant number 1,542 each in *HVAC_N* and *HVAC_S*.

For the categories with fewer missing values like mels_S, mels_N, and lig_S, simple mean imputation was sufficient due to the relatively small gaps and limited impact on the overall data trend.

However, the large number of missing values in HVAC N and S required a more refined approach. A supervised learning method was used to predict these missing values. Features like working day, month, and loads from lig_S, mels_S, and mels_N were considered. A Chi-squared test helped identify the most informative features, and a gradient boosting regressor was used for prediction. The model achieved an R^2 score of 0.84 for HVAC_N and 0.72 for HVAC_S, reflecting strong predictive accuracy and validating the use of this approach for data imputation.

Handling Negative Values: Negative values, which are not physically meaningful in the context of energy consumption, were also found in the dataset. To handle these, each negative entry was replaced with the most recent valid (positive) value preceding it. This simple yet effective technique resulted in an R^2 score of approximately 0.86, supporting its appropriateness for the study's goals.

After preprocessing, the focus shifted toward predicting the building's total energy consumption rather than modeling individual components. All load types were aggregated, resulting in a unified time series that represents the total power usage of

the building, sampled at 15-minute intervals. This cleaned and consolidated dataset formed the foundation for accurate and reliable electricity forecasting.

3.3.3 Feature Extraction

Feature extraction aims to transform raw data into relevant, compact, and informative features that improve the performance of predictive models. It helps in capturing important patterns and relationships within the data, reducing the complexity for the model training.

Types of Features in Energy Consumption Data

- **Temporal Features:** Time-based indicators such as hour of the day, day of the week, weekends, holidays, and seasonal variations.
- **Occupancy Features:** Presence, action, and behavior of occupants, which directly impact energy usage.
- Environmental Features: Indoor/outdoor temperature, humidity, daylight hours, and other environmental factors affecting energy consumption.

Techniques Used for Feature Extraction

  - **Statistical Methods:** To ensure data integrity and model performance, basic statistical techniques are first employed to clean and preprocess the dataset. For instance, any negative energy readings—often artifacts from sensor errors are corrected by replacing them with the most recent valid value. This helps maintain consistency in the time series. Furthermore, patterns such as weekly consumption averages are calculated to smooth out fluctuations and better capture typical usage behavior. These statistical insights not only enhance data quality but also provide a more stable foundation for model training.

- **Frequency Analysis:** The forecasting framework incorporates frequency-based analysis to uncover both ST and LT consumption trends. By leveraging a temporal learning approach, it extracts low-frequency (LT) and high-frequency (ST) features from multivariate time series data. This allows the system to detect subtle variations and periodic spikes in energy usage that occur at different time scales insights critical for generating accurate and robust forecasts.

- **Dimensionality Reduction:** While the process isn't explicitly termed as dimensionality reduction, the approach achieves a similar goal. By transforming and combining frequency-based features into a refined set of representative inputs, the complexity of the original multivariate data is reduced. This enables the next stage of the system to focus on learning meaningful temporal patterns without being overwhelmed by noisy or redundant inputs. In doing so, the model achieves a more effective and streamlined understanding of energy usage behaviors across time.

Multi-Scale Feature Identification

Extracting features at multiple time scales allows capturing different patterns:

- ST patterns (e.g., minute/hourly fluctuations): ST patterns in electricity usage signify the swift and frequent changes that take place over short durations, such as every few minutes or hours. These variations are generally affected by immediate and dynamic elements like shifts in occupancy, device utilization, lighting, HVAC functions, and user actions. For instance, energy demand may surge when employees arrive at the office in the morning, utilize electrical devices, or turn on heating or cooling systems. Likewise, energy consumption may significantly decrease during lunchtime or in periods of vacancy. Grasping ST patterns is crucial for applications that necessitate real-time EM, anomaly detection, or rapid-response control strategies in intelligent building settings.

- LT trends (e.g., daily, weekly, seasonal changes): LT trends refer to more gradual and consistent changes in energy consumption observed over prolonged periods, such as days, weeks, months, or seasons. These trends are influenced by recurring patterns and external factors, including work schedules, weather variations, occupancy habits, holidays, and seasonal influences. For example, buildings may exhibit higher EC on weekdays compared to weekends, or increased heating requirements during the winter months. Identifying these LT trends is beneficial for creating robust EF models, planning infrastructure enhancements, establishing energy budgets, and executing strategic energy-saving measures. They offer essential context for optimizing system operations and ensuring the sustainability of EM practices in buildings.

Feature Selection After extraction, relevant features are selected based on their importance to prediction performance. Techniques such as correlation analysis, mutual information, or built-in model importance metrics are employed to retain the most impactful features, reducing noise and computational overhead.

Impact of Feature Extraction Effective feature extraction enhances model accuracy by providing meaningful inputs that reflect underlying energy consumption drivers, including occupant behavior and environmental conditions, enabling better forecasting and management strategies.

3.4 Experimental and Simulation setup

3.4.1 Performance Measures

Different types of ML, DL, and statistical models are used. Three metrics have been used to compare their performance, which are listed below.

1.MAE

The MAE measures the average absolute difference between predicted and actual values. It is robust to outliers but does not penalize large errors heavily.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Properties:

- Scale-dependent (same units as the target variable).
- Easier to interpret than RMSE.
- Less sensitive to extreme errors compared to RMSE.

2.RMSE

The RMSE computes the square root of the average squared differences between predictions and actual values. It penalizes larger errors more severely.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Properties:

- Scale-dependent (same units as the target variable).
- More sensitive to outliers than MAE.
- Popular in optimization due to differentiability.

3. Coefficient of Determination (R^2 Score)

The R^2 score indicates the proportion of variance in the dependent variable explained by the model. It ranges from $-\infty$ to 1 (higher is better).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Properties:

- Scale-independent (values can be compared across DSs).
- $R^2 = 1$: Perfect fit.
- $R^2 \leq 0$: Model performs worse than a horizontal line.

3.4.2 Simulation setup

All models were trained on a Google Colab with Python3 and backend GPU provided by Google Compute Engine. The configuration includes a single-core Intel(R) Xeon(R) CPU @ 2.20 GHz.