**CIV E 779 Project Report**

Alexander Jordan (ID: 1577190)

Bowen Yang (ID: 1606931)

**Real-Time and Future Building Energy Prediction Using Machine Learning Techniques**

# Introduction

## Background

Building sectors account for 30%-40% of global energy consumption, contributing approximately 19% of greenhouse gas emissions [1]. Heating, ventilation, and air conditioning (HVAC), domestic hot water, lighting, and appliances are the primary services that consume energy in buildings. Globally, HVAC systems account for approximately 40% of total energy consumption [2], and a growing population and rapid urbanization are responsible for the increase in energy consumption. One of the essential parts of building energy management is energy consumption prediction, which could inspire energy policy to reduce energy consumption [3]. Furthermore, daily energy management relies on the energy demand forecast to control the appropriate energy-related equipment. For example, predicted model can be fed to a control system to pre-heat or pre-cool a building before tenants arrives or to storage generated electricity from solar panels if it predicts a peak in electrical usage. Thus, the energy prediction is a key to enable into smart and sustainable buildings design for both new and existing buildings. Much efforts have been made to accurately predict building energy consumption and design an optimal energy efficiency control system to reduce greenhouse gas emissions by managing energy emissions and conserving energy in buildings.

## Problem description

Numerous variables, such as weather, time, building occupancy, etc., make the development of a credible energy prediction model challenging. In building energy prediction systems, machine learning (ML) models, such as tree algorithms, support vector machine (SVM), random forest (RF), and artificial neural networks (ANN), which can describe the relationship between model input and output without requiring complex domain knowledge, have become increasingly important. However, the aforementioned traditional ML approaches have limited ability to cope with time-series data for predicting future energy usage (i.e., next hour, next day and next month). Even though recurrent neural networks (RNN) are effective at handling sequentially data, they are unable to capture long-term dependencies in sequential data [4]. As a variation of the RNN, long short-term memory (LSTM) models may learn the long-term dependent information to produce accurate predictions.

## Question to be addressed

To address the concerns that mentioned above, the following research questions are proposed:

* Are the data from a single year sufficient to estimate real-time and future energy consumption?
* Which machine learning technique is optimal in terms of prediction accuracy and computing cost for predicting energy usage in a smart home?

## Data description and visualization

The main goal of this project is to estimate the real-time and future total energy consumption of a smart building using ML approaches.

The dataset (https://www.kaggle.com/code/malekzadeh/smart-home-data-processing-weather-vs-energy/data) is fetched from Kaggle. The data collection period for a smart home was from 2016-01-01 to 2016-12-16. Nineteen features, including solar generation, appliances (dishwasher, furnace, home office, refrigerator, garage door, kitchen, barn, microwave, and living room), and weather data (temperature, humidity, visibility, pressure, wind speed, and dew point) are extracted to predict real-time and future (next hour) total energy consumption. The information on the factors that were measured is presented in Table 1. Both Fig. 1 and 2 illustrate the total amount of energy that is consumed on a daily and hourly basis, respectively.

**Table 1** Information of variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Categories | Variables | Type | The minimum | The maximum | Unit |
| Energy generation | Generation | Numerical | 0 | 613.88 | W |
| Energy consumption | Dishwasher | Numerical | 0 | 1401.77 | W |
| Furnace | Numerical | 0.33 | 2472.63 | W |
| Home office | Numerical | 0.08 | 971.75 | W |
| Fridge | Numerical | 0.06 | 851.27 | W |
| Wine cellar | Numerical | 0.02 | 1273.94 | W |
| Garage door | Numerical | 0.02 | 1088.93 | W |
| Kitchen | Numerical | 0 | 2265.87 | W |
| Barn | Numerical | 0 | 7027.90 | W |
| Well | Numerical | 0 | 1633.02 | W |
| Microwave | Numerical | 0 | 1929.80 | W |
| Living room | Numerical | 0 | 465.22 | W |
| Weather condition | Temperature | Numerical | -12.64 | 93.72 | °F |
| Humidity | Numerical | 0.13 | 0.98 | % |
| Visibility | Numerical | 0.27 | 10 | Kilometers |
| Apparent temperature | Numerical | -32.08 | 101.12 | °F |
| Pressure | Numerical | 986.4 | 1042.46 | Millibar |
| Windspeed | Numerical | 0 | 22.91 | Km/h |
| Dewpoint | Numerical | -27.24 | 75.49 | °F |
| House overall | Numerical | 0 | 14714.57 | W |

***Chart

Description automatically generated***

**Fig. 1.** Daily total energy consumption.

Chart

Description automatically generated

**Fig. 2.** Hourly total energy consumption.

# Formulation of the mathematical model

## 2.1 LSTM model

Diagram

Description automatically generatedLSTM is specifically designed to prevent the problem of long-term dependency. The cell state and gate structure are the core concept of LSTM. Fig. 3 (a), (b), (c), and (d) depict the calculation process of the forget gate, input gate, cell state, and output gate, respectively. Using the gradient descent approach, the initial values of the weight matrices (i.e., ) and bias (i.e., ) are updated. The information from the current input  and previous hidden state is transmitted through the sigmoid function , and forget gate determines whether the gate should retain the information or not, according to the Eq. (1). The input gate is utilized to update the cell status, which includes two tasks. First, the precious cell output and current state are supplied to a sigmoid function to determine what relevant information needs be updated, as shown in Eq. (2). Second, the previous cell output and current input are again sent to the tangent function in order to generate a new candidate value Eq. (3). The cell state ( is the path of information transmission, so the information can be passed in a serial connection, Eq. (4). Finally, the current output vector is determined by linear transformation through sigmoid function of previous cell state and current input vector and passes the output through tangent function, and LSTM is used to predict the next-hour building energy consumption at time, as shown in Eq. (5) and (6).

**Fig. 3**. Procedure of forward propagation of LSTM.

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  | |  |  |  | | () |
|  | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  | |  |  |  | | () |
|  | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  | |  |  |  | | () |
|  | |  |  |  | | --- | --- | --- | |  |  |  | |  |  |  | |  |  |  | | () |
|  |  | |  |  |  | | --- | --- | --- | |  |  |  | |  | () |  | |  |  |  | |

## 2.2 GRU model

Gated recurrent unit (GRU) is another variant model to overcome traditional RNN problems for the time-series analysis. The three gates of the LSTM are reduced to two: reset gate and update gate. The reset gate determines how much past information we wish to keep, whereas the update gate determines how much new states are copied from the old state. Fig. 4 illustrates a typical structure of GRU, the calculation process is defined in Eq. (7), Eq. (8) and Eq. (9).

|  |  |  |
| --- | --- | --- |
|  |  | () |
|  |  | () |
|  |  | () |

A picture containing text, clock

Description automatically generatedwhere is the input at time step , , and are the hidden states at the time step and , respectively. is the output of the reset gate at time step . and , , and are the weight matrixes.

**Fig. 4.** Basic architecture of a GRU cell.

## RNN model

RNN is a class of ANN with a loop that repeats the same task for each element in a sequence [5], and has a memory for capturing information, whereas traditional neural networks presume that all inputs are independent of one another and hence lack the ability to process sequential data. Fig. 5 shows the general structure of estimating the building’s energy using the RNN , and are the hidden states at the time step and , respectively. and are the weight matrix, as shown in Eq. (10).

|  |  |  |
| --- | --- | --- |
|  |  | () |

***Diagram

Description automatically generated***

**Fig. 5.** A recurrent neural network.

## 2.4 Traditional artificial neural networks

Diagram

Description automatically generatedANN has been developed to imitate human nervous system and used into mathematical models. In its simplest form, as demonstrated in Fig.6. , through represent the input used to train and test the ANN model, along with their corresponding weights , through , bias  and activation function (which can be the activation functions of linear, sigmoid, Tanh, and rectified linear units) applied to the weighted sum of the inputs to compute the output .

**Fig. 6.** Example of a neural network.

|  |  |  |
| --- | --- | --- |
|  |  | () |

**Table 2** Pros and cons of models.

|  |  |  |
| --- | --- | --- |
| Model | Advantages | Disadvantages |
| ANN | * Can be applied to complex non-linear problems. | * Fails to model sequential data. |
| RNN | * Can handle time-series data. | * Issue of gradient vanishing. * Cannot capture ling-term dependencies. |
| GRU | * Can maintain a long-term dependency. | * Slow convergence. * Low learning efficiency. |
| LSTM | * No gradient vanishing issues. * Can learn long-term dependencies. | * Long training time. |

## 2.5 Performance criteria

To evaluate the ML models, R square (R2) and Mean Absolute Percentage Error (MAPE) were used in this study. The formula for R2 MAPE are:

|  |  |  |
| --- | --- | --- |
|  |  | () |

where is the actual value of the data point, and is the predicted value of the data point.

|  |  |  |
| --- | --- | --- |
|  |  | () |

where represents the actual values, represents the prediction, and represents the mean of all the values.

# Solution of the problem

In this study, LSTM and GRU will be used to forecast future energy consumption, while ANN and RNN will be utilized to predict energy consumption in real-time. The code is available in Appendix.

# Interpretation of results

## 4.1 Result of data preprocessing

## Before constructing a future prediction model, data processing is a prerequisite. To obtain a clear view of the time-series data, we first changed the timestamp to the data time. Second, we determined the total for the submeter-sized furnace and kitchen characteristics (i.e., furnace 1, and furnace 2 and kitchen 12 and kitchen 14). Third, the input data should be standardized to account for the varying ranges of each attribute. In this investigation, min-max normalization was utilized. The dataset was separated into training and testing datasets for training and testing purposes. In this study, 80% of the data are used to train the models, whereas 20% are chosen for testing.

## Chart, waterfall chart Description automatically generated4.2 Result of exploratory data analysis

**Fig. 7.** Correlation between features and output.

Fig. 7 describes the correlation between the inputs and output for one year. A dark color indicates the strong correlation while a light color denotes the weak relationship. Obviously, there is a strong correlation between the energy consumption of the furnace and the total energy consumption of the house (coefficient: 0.5), which indicates that the furnace consumes more energy than other appliances because it is a multipurpose device for heating, cooling, and ventilation. Consequently, furnace may become an important factor for future energy forecast. Low correlation between inputs and output does not imply that they are useless; when paired with additional features, the prediction accuracy normally can be enhanced.

In addition, solar and generation have an extraordinarily high correlation, indicating that they have identical values and that one of them should be abolished to eliminate the redundant feature. Thus, solar was eliminated to improve the accuracy of predictions and reduce computing costs.

## Performance comparison among machine learning algorithms for future and real-time energy prediction (could you add more contents and put figures that we created?)

### Prediction accuracy comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Prediction type | Algorithms | Evaluation metrics | |
| R square | MAPE |
| Real-time | ANN | 0.738 | 0.303 |
| RNN or GRU (need to be changed) |  |  |
| Future | LSTM | 0.680 | 0.396 |
| RNN or GRU (need to be changed) |  |  |

### Computational comparison

|  |  |  |
| --- | --- | --- |
| Algorithms | Training (s) | Testing time (s) |
| ANN | 77.125 | 7.529 |
| RNN or GRU (need to be changed) |  |  |
| LSTM | 43.005 | 1.050 |
| RNN or GRU (need to be changed) |  |  |

# Critique of the model (future works)

Overall, the models satisfy the answer the question posted. However, certain constraints must be addressed in the future.

1. Implement feature selection/feature importance to choose the best feature combination to increase the prediction accuracy and reduce computational cost.
2. Investigate the capability of model robustness and generalizability for long-term predictions and the energy estimation of different buildings.
3. Integrate data-driven energy prediction model into further applications, such as model predictive and reinforcement learning control.

# Appendices with supporting materials (code)

**Reference**

[1] G. Pinto, D. Deltetto, A. Capozzoli, Data-driven district energy management with surrogate models and deep reinforcement learning, Applied Energy. 304 (2021) 117642. https://doi.org/10.1016/j.apenergy.2021.117642.

[2] Y. Peng, A. Rysanek, Z. Nagy, A. Schlüter, Occupancy learning-based demand-driven cooling control for office spaces, Building and Environment. 122 (2017) 145–160. https://doi.org/10.1016/j.buildenv.2017.06.010.

[3] X.J. Luo, L.O. Oyedele, Forecasting building energy consumption: Adaptive long-short term memory neural networks driven by genetic algorithm, Advanced Engineering Informatics. 50 (2021) 101357. https://doi.org/10.1016/j.aei.2021.101357.

[4] I. Karijadi, S.-Y. Chou, A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction, Energy and Buildings. 259 (2022) 111908. https://doi.org/10.1016/j.enbuild.2022.111908.

[5] Z. Pang, F. Niu, Z. O’Neill, Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons, Renewable Energy. 156 (2020) 279–289. https://doi.org/10.1016/j.renene.2020.04.042.