Mediterranean Smart Cities Conference (MSCC 2024)
2-4 May, 2024, Martil, Morocco

Intelligent system-based Energy Consumption Monitoring using IoT and Machine Learning

Mohamed hamza Boulaich
National High School for Computer Science and Analysis,
Mohamed V University in Rabat
Rabat, Morocco
Mohamedhamza_boulaich@um5.ac.ma

Said Ohamouddou
National High School for Computer Science and Analysis,
Mohamed V University in Rabat
Rabat,Morocco
said_ohamouddou@um5.ac.ma

Hibatallah Meliani
National High School for Computer Science and Analysis,
Mohamed V University in Rabat
Rabat,Morocco
hibatallah_meliani@um5.ac.ma

Mohammed Ali Ennasar
National High School for Computer Science and Analysis,
Mohamed V University in Rabat
Rabat, Morocco
mohammed-ali_ennasar@um5.ac.ma

Abdellatif El Afia
National High School for Computer Science and Analysis,
Mohamed V University in Rabat
Rabat, Morocco
abdellatif.elafia@ensias.um5.ac.ma

***Abstract*—This work presents a new approach of intelligent systems to efficiently control and enhance energy consumption in intelligent buildings. We utilize a Raspberry Pi model that is equipped with sensors to establish an Internet of Things (IoT) system for the purpose of detecting and quantifying the temperature within the building. Then transport these data into the server using the efficient communication technology MQTT. In order to calculate energy consumption, we utilize a range of machine learning methods such as Random Forest, Gradient Boosting, MLP Regressor, K-neighbors, Decision Tree, Support Vector Regression, Ridge, Linear Regression, Lasso, and ElasticNet. To determine who is the best model of ML, we employ evaluation measures such as MAE, MSE, RMSE, and R-squared. The RF offers the best prediction performance with high accuracy.**

Keywords—Internet of Things (IoT); raspberry Pi; Wireless Sensor Network; Machine Learning (ML) Algorithm.

I. Introduction

Energy consumption in Morocco increases progressively due to the demography growth. By 2022, the population had reached a remarkable 37 million individuals and is projected to extend its growth in the next years [1]. As a result, the growth of the population leads to a corresponding rise in the country's energy consumption. It has produced 41.15 billion kWh of power and will consume 35.39 billion kWh by the year 2021 [1]. To save energy in this country, it should employ smart home and industry systems to regulate and minimize energy usage as much as possible in order to increase energy efficiency. Smart buildings are integrated with advanced technology and interconnected systems to enhance their overall efficiency, functionality, and sustainability.

The digitized built environment plays a crucial role in this time and age, since it is enabled and interpreted through digital language by sensors that gather and analyze data. Data on building energy efficiency is captured by sensors that are embedded both indoors and outside. This data allows for the automatic control of lighting, temperature, heating, ventilation, and air conditioning (HVAC) systems. These systems can alter automatically based on criteria such as occupancy and weather conditions, effectively minimizing energy wastage [2], [3].

The concept of smart buildings relies on Internet of Things (IoT) technology, enabling the integration of household gadgets over a sophisticated network [4]. The Internet of Things (IoT) supplies users with enough information by facilitating connectivity with numerous electronic gadgets through a wireless sensor network, for instance, a smart thermostat that autonomously changes room temperature or a smart light that glows upon detecting movement gives convenience to users while improving energy efficiency [5]. With the development of Artificial Intelligence (AI) and the Internet of Things (IoT), there is a potential to comprehend building energy usage trends using this approach. Particularly, the ML algorithm has embraced this study to anticipate short-term energy consumption in buildings because of its wide use in numerous sectors. Model performance was tested using statistical methods that consist of mean square errors (MSE), mean absolute errors (MAE), root mean square errors (RMSE), and R^2.

The rest of this paper is as follows: Section 2 explores essential literature, while Section 3 introduces several machine learning algorithms, and Section 4 presents the methodology. Section 5 exposes the evaluation results, and

Section 6 provides the concluding thoughts and a framework for future work.

II. RELATED WORK AND CONTRIBUTION

There are several ways to assessing and estimating energy usage that have been employed in prior studies. Energy prediction based on IoT devices is the most critical to enhance the energy usage rate, to forecast anomalies and potential equipment failures for preventive maintenance, and allow the building's energy management system conducts the most effective evaluation [6]. Many scholars have developed their studies using machine learning approaches and IoT devices to predict energy consumption.

Hanane Allioui et al. [10]. Presented a comprehensive state-of-the-art application of the Internet of Things (IoT) connected with numerous devices. IoT has been applied as a solution in different industries and applications. IoT can be used to connect and communicate between equipments and systems, which has led to new solutions that improve efficiency.

In their work, Nazli Tekin et al. [7] examined security and privacy concerns for users of smart home systems (SHS) and conducted a comparative analysis of on-device machine learning (ML) related energy usage for IoT intrusion detection applications. They independently investigated both the training and inference stages, comparing cloud computing-based ML, edge computing-based ML, and IoT device-based ML approaches. The suggested method comprises a prediction model based on a decision tree (DT), which is contrasted with logistic regression (LR), K-Nearest Neighbor (K-NN), Naive-Bayes (NB), Random Forest (RF), and an artificial neural network (ANN). The DT algorithm built on-device delivers improved results in terms of training time, inference time, and power consumption.

Zeqing Wu et al. [8] constructed support vector regression (SVR), artificial neural network (ANN), and random forest (RF) models to estimate energy consumption in smart buildings, assessing the performance of each model. The data includes temperature, humidity, weather station, and electricity usage statistics for electric lights. The evaluation criteria utilized were mean square error (MSE) and mean absolute error (MAE). Their investigation indicated that the Random Forest (RF) model exhibited the highest prediction accuracy.

Nursyura Maili Mazlan et al. [9] proposed a K-nearest neighbor (K-NN) method to assess and predict energy usage, aiming to boost energy management efficiency in a commercial smart building. They separate the dataset into training and testing sets with ratios of 75% and 25%, respectively. Time series data from selected tenants in the commercial smart building, equipped with the Internet of Things (IoT), underwent statistical analysis. The performance of the projected data was evaluated using the root mean square error (RMSE) measure. Their findings revealed that the K-nearest neighbor with k = 5 exhibited the most accuracy, producing the lowest average RMSE values of 5.73, 8.54, and 0.35 for each respective renter in the building.

Soualihou Ngnamise Njimbouom et al. [11] conducted experiments employing four machine learning techniques: random forest (RF), gradient boost decision tree (GBDT), support vector regressor (SVR), and decision tree for regression (DT). The study examined different models to determine the best-performing strategy, measuring their performance using root mean square error (RMSE) and mean absolute error (MAE) criteria. The results suggested that the random forest (RF) model outperformed the others, reaching MAE and RMSE values of 14.91 and 20.84, respectively.

Forecasting power demand is crucial for smart cities, as it allows for accurate assessment of the effects of many variables, Malek Sarhani and abdellatif el afia [16]-[18], investigated using the partical swarm optimization for both feature selection and model selection problems of support vector regression for electric lead forcasting. The experimental results can achieve better performance when compared with the classical SVR model while using feature selection and without using it.

Main contribution: In this study, our attention lies on the dataset offered by the Women in Data Science (WiDS) project in 2022 [12], and its application in predicting energy usage using machine learning algorithms. We apply several learning methods such as Random Forest (RF), Gradient Boosting (GB), MLP Regressor, K-neighbors, Decision Tree (DT), Support Vector Regression (SVR), Ridge, Linear Regression (LR), Lasso, and ElasticNet to estimate energy usage. Table 1 shows comparing various machine learning regressors:

Table1: comparative study of machine learning Regressors

Algorithm	Description	Strengths	Weaknesses	When to Use
Linear Regression	Finds a linear relationship between features and target variable	Simple, interpretable, fast	Assumes linearity, sensitive to outliers	Good baseline, works well for linear relationships
Ridge Regression	Regularized linear regression to reduce overfitting	Reduces variance, handles collinearity	Less interpretable than linear regression	When dealing with multicollinearity or high dimensional data
Lasso Regression	L1 regularized linear regression for feature selection	Performs feature selection, good for sparse data	Less interpretable than ridge, may not capture complex relationships	When interpretability is desired and dimensionality reduction is a benefit
Elastic Net	Combines L1 and L2 regularization of Lasso and Ridge	Incorporates feature selection and reduces variance	Tuning hyperparameters can be complex	When dealing with high dimensional data with potentially sparse features
K-Nearest Neighbors (KNN)	Predicts based on the k nearest neighbors in the training data	Non-parametric, works well for high dimensional data	Sensitive to irrelevant features, curse of dimensionality	For small datasets or when the relationship between features and target is complex

Decision Tree	Tree-like structure for splitting data based on features	Interpretable, handles non-linear relationships	Prone to overfitting, sensitive to small changes in data	Good for initial exploration and understanding feature importance
Random Forest	Ensemble of decision trees, improves on single decision tree	More robust to overfitting, handles non-linear relationships	Less interpretable than decision trees, can be computationally expensive	General purpose regression for a variety of datasets
Gradient Boosting	Sequential ensemble method that builds on previous trees	Highly accurate, can handle complex relationships	Black box model, computationally expensive	When high accuracy is needed and interpretability is less important
Support Vector Regression (SVR)	Finds a hyperplane with the largest margin between data points	Handles high dimensional data, good for small datasets	Sensitive to parameter tuning, can be slow for large datasets	When dealing with limited data and potentially non-linear relationships
MLPRegressor (Multi-Layer Perceptron)	Artificial neural network for non-linear regression	Highly flexible, can model complex relationships	Black box model, prone to overfitting, computationally expensive	For complex, non-linear problems where interpretability is less important

III. The Machine Learning Algorithm

In this study, seven machine learning methods have been utilized to undertake energy efficiency prediction. We introduce some of them, but not all, because they are dependent on each other.

A. *Random Forest*

The Random Forest algorithm [13] is an ensemble approach constructed of decision trees, where each tree is built from a bootstrapped iteration of the training dataset. Each tree grows by recursively partitioning nodes, starting from the root node, using the same splitting criterion until specified stopping criteria are satisfied. In our approach, we leverage the CART technique from the sklearn package to generate binary trees. The Random Forest algorithm can be codified as indicated in "Eq. (1)".

$$G\left(x_i, v_{i,j}\right) = \frac{n_{left}}{N_s} H\left(X_{left}\right) + \frac{n_{right}}{N_s}\left(X_{right}\right) \qquad (1)$$

x_i: is a categorical variable.
$v_{i,j}$: is the split value of the categorical variables.
N_S: is the number of all training samples of the current node
n_{left} and n_{right}: are the number of training samples of the left and right child nodes after segmentation.
X_{left} and X_{right}: are the training sample sets of the left and right child nodes respectively.

B. *Gradient Boosting*

Gradient boosting [14] is an ensemble learning strategy that combines the predictions of numerous weak learners, often decision trees, to generate a robust predictive model. The Gradient Boosting algorithm can be codified as indicated in "Eq. (2)"

$$g_t(x) = E_y\left[\frac{\partial\Psi(y, f(x))}{\partial f(x)}|x\right]_{f(x)=\hat{f}^{t-1}(x)} \qquad (2)$$

$g_t(x)$: gradient
$\Psi(y, f(x))$: loss function
$f(x)$: function estimate
E_y: expected y-loss function

C. *MLP Regressor*

MLPRegressor, short for Multi-Layer Perceptron Regressor, is a form of artificial neural network (ANN) built specifically for regression problems. Its objective is to anticipate a continuous target variable.

IV. The Proposed Method

This study intends to present a system based on intelligent IoT with the support of sensors in smart buildings, and to apply machine learning (ML) in sensor-related disciplines.

A. *Methodology*

The major goal of this work is to operate a smart building system that integrates energy consumption management utilizing IoT sensors, machine learning, and MQTT, which entails a systematic method. Figure. 1 depicts a simplified system approach. Here's a proposed procedure:

- First, we applied an IoT-based Raspberry Pi in order to receive the information acquired by the temperature sensor.
- Second, transfer the data to the server via MQTT communication protocols.
- Third, these data gathered by the sensor move on to be addressed under three stages: data collecting, data preparation, and prediction.
- Finally, convey the advice from Module GSM Sim 900 to the user (turn off superfluous lights, unhook unwanted equipment, unplug unused electronic, etc.).

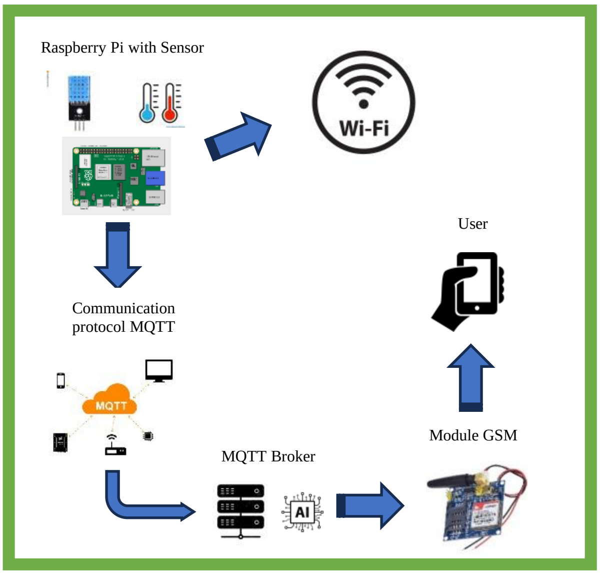

Figure 1: The architecture of smart home system

The procedure of machine learning (ML) in sensor-related disciplines demonstrates three stages of data collecting: data collection, data preparation, and prediction. Figure. 2 is an illustration of the overall concept model.

- Data collection is the process of acquiring useful information or data points from numerous sources utilizing a sensor-based method. The purpose here is to construct a dataset that has the necessary features and labels for training a machine learning model.
- Data preparation comprises cleaning, removing extraneous features, and arranging raw data into a format appropriate for machine learning algorithms. The purpose here is to handle missing values and repair errors.
- Prediction is the process of utilizing a trained machine learning model to make predictions or classifications on new, unseen data. It entails examining various ML algorithms evaluated using criteria such as accuracy, precision, recall, or others.

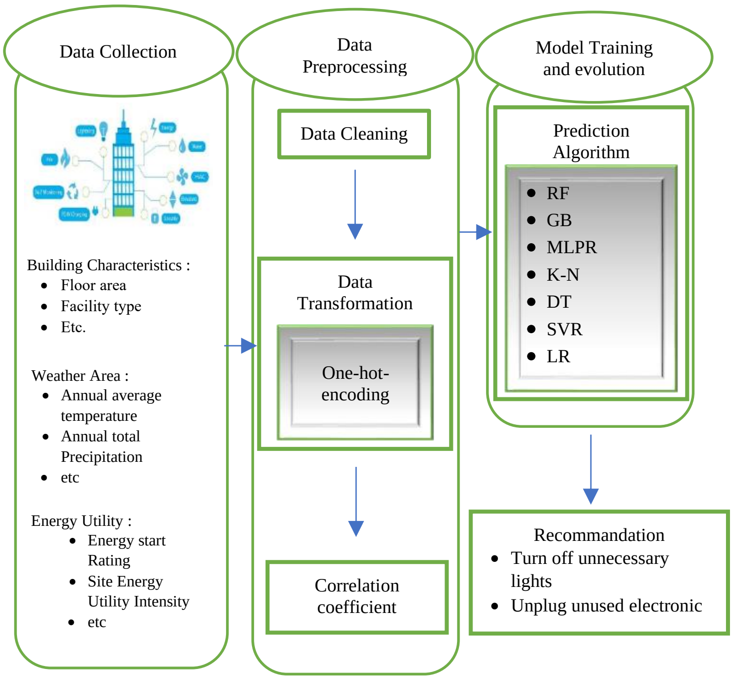

Figure 2: The process of using machine learning in a smart home

B. Development envirnment

The testing was conducted on a workstation mentioned in the table, equipped with 64 GB (2 x 32) of RAM, an AMD Ryzen Serie 7 7700X (4.5 GHz) CPU, and an NVIDIA GPU RTX 3080, running a 64-bit Ubuntu 23.10 operating system. The machine learning methods were implemented using CUDA version 11.8.

Table 2: Workstation Environmental Configuration

Components	Description
GPU	NVIDIA RTX 3080
CPU	AMD Ryzen Serie 7 7700X
RAM	64 GB (32 GB X 2)
OS	Ubuntu 23.10 64 bit
CUDA	11.8
Python	3.8.0

C. Dataset

The dataset offered by the Women in Data Science WiDS Datathon 2022 [12] focused on a prediction challenge including about 100,000 observations of building energy usage statistics spanning seven years and numerous states across the United States. This dataset includes building characteristics (e.g., floor area, facility type, etc.), weather data for the building's location (e.g., annual average temperature, annual total precipitation, etc.), and energy usage for the building and given year, quantified as Site Energy Usage Intensity (Site EUI).

D. Evaluation Metrics

During this work, we applied four evaluation metrics to assure the trustworthiness of our anticipated outcomes. We evaluated the root mean square error (RMSE), the mean absolute error (MAE), R-squared, and Mean Squared Error (MSE) four typical metrics used in model evaluation, as critical measurements to assess the accuracy and precision of our predictive models.

- MAE: Mean Absolute Error (MAE) [15] is a statistical measure applied in regression analysis to quantify the average amount of errors between anticipated and actual values. It examines the absolute discrepancies between the observed real values and the model-predicted values for each data point, subsequently determining the average of these absolute differences. The formula for the mean absolute error can be represented as given in“Eq. (3).”

$$MAE = \frac{\sum_{i=1}^{m}|y_i - \hat{y}_i|}{m} \quad (3)$$

m: is the number of data points.
y_i: is the i-th predicted value.
$\hat{y}_i$: is the i-th actual value.

- RMSE: Root Mean Square Error (RMSE) [15] is a statistic often applied in regression analysis to assess the average amount of errors between anticipated and actual values. It quantifies the root mean squared difference between the observed real values and the model's projected values. The formula for root mean square error is represented as illustrated in “Eq. (4).”

$$RMSE = \frac{\sqrt{\sum_{i=1}^{m}(|y_i - \hat{y}_i|)^2}}{m} \quad (4)$$

- R-squared: R-squared [15] is a statistical measure that represents the proportion of variability in the dependent variable explained by the independent variables inside a regression model. The formula for computing R-squared can be stated as given in“Eq. (5).”

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(\bar{y} - \hat{y}_i)^2} \quad (5)$$

y_i: is the i-th predicted value.
$\hat{y}_i$: is the i-th actual value.
$\bar{y}$: is the average value of the dependent variable for the i-th data point according to the linear regression.

- MSE: Mean Squared Error (MSE) [15] is particularly valuable when outliers need to be recognized. It is a commonly used statistic in machine learning to assess the effectiveness of a regression model. The formula for mean squared error is given as stated in "Eq. (6)."

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \qquad (6)$$

V. EXPERIMENTAL RESULTS

In this study, the dataset is structured in a tabular format using CSV files, where rows and columns are utilized to represent individual data points. (e.g., information on a specific building), and each column indicates a characteristic or attribute related with the data points (e.g., year built, floor area, site EUI).

Before creating machine learning models, We investigated the correlation coefficient between each independent variable and the dependent variable, simplifying the identification of potential linkages between them. Strong positive correlations suggest that higher values of the independent variable can be related with higher values of the dependent variable (or vice versa). Strong negative correlations suggest an opposite trend, where higher values of one variable can be related with lower values of the other. Weak correlations (closer to 0) suggest little or no linear relationship. Fig. 3 illustrates the distribution of each correlation coefficient.

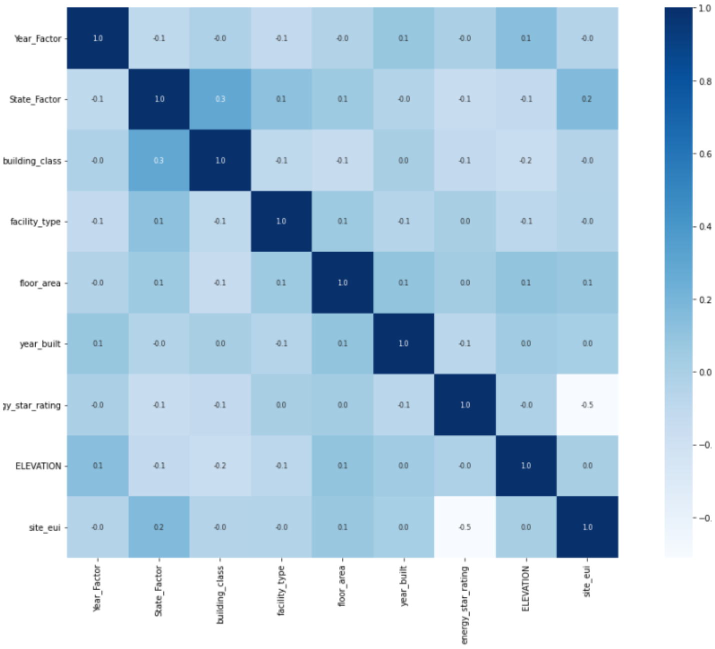

Figure 3: Heatmap of correlation coefficient matrix

After the training, we choose the independent variable- these are the qualities or attributes we believe might impact the result variable and the dependent variable this is the goal variable we are trying to predict or understand. By limiting the threshold of the correlation coefficient to be 1, we selected year factor, building class, floor area, energy star rating, and year factor from the database as independent variables. The total number of independent variables is 5. By choosing energy usage intensity (site-eui) as a dependent variable to train the machine learning method.

MAE, MSE, RMSE, and R-squared were utilized as indications of the robustness of the various evaluated machine learning ML approaches. The error rate produced by the optimal model, Random Forest (RF), was consistently lower than that of Gradient Boosting (GB), MLPR, K-N, DT, and SVR approaches. The random forest model provides the best overall performance, with the lowest MAE, MSE, and RMSE scores and the highest R-squared score. This suggests that the Random Forest model is able to create predictions that are closer to the actual values on average. Table 3 illustrates the statistical performance characteristics of the applied machine learning models, including MAE, MSE, RMSE, and R^2

Table 3: Performance Comparison of Different Machine Learning Algorithms

Model	MAE	MSE	RMSE	R^2
Random Forest	**19.157504**	**1565.858384**	**39.570929**	**0.549126**
Gradient Boosting	22.903550	2107.477478	45.907270	0.393171
MLPRegressor	24.097960	2276.127861	47.708782	0.344610
KNeighbors	25.167562	2473.228792	49.731567	0.287857
Decision Tree	24.132982	2592.079182	50.912466	0.253635
SVR	24.259827	2773.962804	52.668423	0.201263
Ridge	28.614974	2851.941114	53.403568	0.178810
Linear Regression	28.615034	2851.941213	53.403568	0.178810
Lasso	28.317999	2857.149737	53.452313	0.177310
ElasticNet	28.570712	2935.829368	54.183249	0.154655

Table 4 displays the comparison between our results and the results obtained from the three-machine learning algorithms.

Table 4: comparative results of the three-machine learning algorithm.

Reference	Model	MAE	RMSE
[11]	Random Forest	14.63	20.54
	Decision Tree	20.61	26.87
	SVM	17.25	23.52
Our proposed	Random Forest	19.15	39.75
	Decision Tree	24.13	50.91
	SVM	24.25	52.66

The machine learning model described in [11] outperforms other models in a three-algorithm comparison. This was achieved by tuning the hyper-parameters of the models and applying feature selection. The analytic results were then compared using 10-fold cross-validation, focusing on the metrics of mean absolute error (MAE) and root mean square error (RMSE).

VI. CONCLUSION AND FUTURE WORK

In this research, we suggested a new approach intelligent system to monitor and manage energy consumption in smart building employing hardware equipment such as IoT devices with sensor of temperature to measure the physical qualities in the environment. thus, are interested about datasets, that creating models to anticipate building energy usage. Prediction in this research is done by employing Random Forest (RF), Gradient Boosting (GB), MLP-Regressor, K-neighbors, Decision Tree (DT), Support Vector Regression (SVR), and four different types of regression techniques used in machine learning. The projected data is submitted to training and testing in a 75 and 25 ratio, respectively to prevent overestimation. the measures of the model's performance are MAE, MSE, RMSE, R^2 are applied on the predicted data to calculate the accuracy value. After suitable data preparation, the Random Forest (RF) method displayed greater performance compared to other approaches, providing the following results: 19.157504 for MAE, 1565.858384 for MSE, 39.570929 for RMSE, and 0.549126 for R^2. Our model has low error rates allowing us to produce pretty accurate forecasting of the energy usage intensity (site-eui). In subsequent investigations, we will concentrate on the prototype's features. Various sensors, such as temperature sensors, humidity sensors, occupancy sensors, and smart meters, can be installed in a smart building to collect data on energy consumption. This data is subsequently transported to a central hub for processing and analysis.

References

[1] https://globaledge.msu.edu/countries/morocco/statistics

[2] King, Jennifer, and Christopher Perry. Smart buildings: Using smart technology to save energy in existing buildings. Washington, DC, USA: Amercian Council for an Energy-Efficient Economy, 2017.

[3] Dong, Bing, et al. "A review of smart building sensing system for better indoor environment control." Energy and Buildings 199 (2019): 29-46.

[4] Verma, Anurag, et al. "Sensing, controlling, and IoT infrastructure in smart building: A review." IEEE Sensors Journal 19.20 (2019): 9036-9046.

[5] Lin, Huichen, and Neil W. Bergmann. "IoT privacy and security challenges for smart home environments." Information 7.3 (2016): 44.

[6] Zhou, Ruijin, et al. "Building energy use prediction using time series analysis." 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications. IEEE, 2013.

[7] Tekin, Nazli, et al. "Energy consumption of on-device machine learning models for IoT intrusion detection." Inte rnet of Things 21 (2023): 100670.

[8] Wu, Zeqing, and Weishen Chu. "Sampling strategy analysis of machine learning models for energy consumption prediction." 2021 IEEE 9th International Conference on Smart Energy Grid Engineering (SEGE). IEEE, 2021.

[9] Mazlan, N., et al. "A smart building energy management using internet of things (IoT) and machine learning." Test. Eng. Manag 83 (2020): 8083-8090.

[10] Allioui, Hanane, and Youssef Mourdi. "Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey." Sensors 23.19 (2023): 8015.

[11] Ngnamsie Njimbouom, Soualihou, et al. "Predicting Site Energy Usage Intensity Using Machine Learning. Models." Sensors 23.1 (2022): 82.

[12] WiDS Datathon 2022 Available onilne:https://www.kaggle.com/competitions/widsdatatho n 022 (accessed on 4 july 2022)

[13] Hu, Jianchang, and Silke Szymczak. "A review on longitudinal data analysis with random forest." Briefings in Bioinformatics 24.2 (2023): bbad002.

[14] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.

[15] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021): e623.

[16] Sarhani, Malek, and Abdellatif El Afia. "Feature selection and parameter optimization of support vector regression for electric load forecasting." 2016 International Conference on Electrical and Information Technologies (ICEIT). IEEE, 2016.

[17] Sarhani, Malek, and Abdellatif El Afia. "Electric Load Forecasting Using Hybrid Machine Learning Approach Incorporating Feature Selection." BDCA. 2015.

[18] Sarhani, Malek, Abdellatif El Afia, and Rdouan Faizi. "Hybrid approach-based support vector machine for electric load forecasting incorporating feature selection."