

# Prediction Of Energy Usage In IoT Devices



Sangeetha Ramesh  
10636709

Dublin Business School

This dissertation is submitted for the degree of  
*Master of Science in Business Analytics*

*Supervisor name:* Dr. Syed Mustufa

May 2024

## Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

*Signed:* Sangeetha Ramesh

*Student number:* 10636709

*Date:* 17<sup>th</sup> May 2024

## **Acknowledgements**

I express my deepest gratitude to Dr. Syed Mustufa, my esteemed supervisor at Dublin Business School, for his exceptional guidance and unwavering support throughout this research journey. His profound knowledge and insightful feedback were invaluable to the success of this project. I extend sincere appreciation to Dublin Business School for the opportunity to pursue my Master of Science in Business Analytics and for providing invaluable resources that enriched my academic journey. Additionally, I am profoundly thankful to my cherished family and friends for their continuous encouragement and unwavering patience. Their unwavering support bolstered my determination and contributed significantly to my personal and academic growth. This journey has been a transformative experience, and I am deeply grateful to everyone who has been a part of it.

## **Abstract**

Understanding energy consumption models can enhance energy management and cut expenses. This study aims to forecast energy usage considering different temporal and environmental factors. It specifically targets predicting the energy consumption of IoT devices, utilizing data from smart homes in Greece. Data collected from the smart home is analyzed using machine learning models including Random Forest, Support Vector Machine (SVM), Logistic Regression and XGBoost. Research involves data preprocessing, feature engineering, and applying these models to produce accurate predictions. The XGBoost and Random Forest models demonstrated the highest accuracy, highlighting their effectiveness in this context. This study provides valuable insights into energy management in smart homes, facilitating the development of efficient energy strategies and improving environmental sustainability. It also highlights the importance of using advanced machine learning techniques to accurately predict energy consumption.

# Table of Contents

Chapter 1: Introduction.....	9
1.1 Background .....	9
1.2 Aim.....	9
1.3 Research Motivation .....	9
1.4 Research Objective.....	10
1.5 Research Questions .....	10
1.6 Report Structure .....	11
Chapter 2: Literature Review.....	13
2.1 IoT-Based Energy Consumption and Machine Learning.....	13
2.2 Implementation and Evaluation of Models .....	14
2.3 Energy Optimization Strategies .....	15
2.4 Challenges and Limitations .....	18
2.5 Addressing Research Gaps.....	19
2.6 Justification for Model Selection .....	20
Chapter 3: Methodology .....	23
3.1 Data Understanding .....	23
3.2 Data Preprocessing.....	24
3.2.1 Handling Missing Values.....	25
3.3 Model Implementation .....	26
3.3.1 Random Forest Classifier.....	26
3.3.2 Support Vector Machine (SVM).....	27
3.3.3 Logistic Regression.....	28
3.3.4 XGBoost .....	28
3.4 Evaluation Metrics .....	29
3.4.1 Breakdown of the Confusion Matrix – Random Forest.....	29
3.4.2 Breakdown of the Confusion Matrix – SVM.....	32
3.4.3 Breakdown of the Confusion Matrix – Logistic Regression .....	35
3.4.4 Breakdown of the Confusion Matrix - XGBoost.....	37

3.5	Feature Importance.....	40
Chapter 4: Result Analysis.....		43
4.1	Model Performance .....	43
4.2	Cross-Validation Results.....	45
4.3	Breakdown of the Heatmap Metrics.....	46
4.4	Practical Implications for Predicting Energy Usage .....	47
4.5	Key Insights.....	49
Chapter 5: Conclusion and future works .....		50
5.1	Conclusion.....	50
5.2	Future Work .....	52
References.....		54

## List of figures

1	Random Forest Classifier.....	22
2	SVM.....	22
3	Logistic Regression.....	23
4	XG Boost Model.....	23
5	Confusion Matrix – Random Forest.....	24
6	Confusion Matrix – SVM.....	26
7	Confusion Matrix – Logistic Regression.....	28
8	Confusion Matrix – XGBoost.....	30
9	Feature Importance – Random Forest.....	33
10	Feature Importance – XG Boost.....	33
11	Model Comparison.....	36
12	Analysis of Heatmap for Model performance Metrics.....	37

## List of tables

1	Literature Review.....	14
2	Data Overview.....	19
3	Model Performance Results.....	35

# Chapter 1: Introduction

## 1.1 Background

Home consumption is influenced by a variety of factors, including daytime, season and environmental conditions. Understanding these models is important to improve energy efficiency and reduce costs. This research focuses on using machine learning techniques to predict energy consumption in IoT-equipped smart homes. The dataset was collected from a smart home in Greece and includes information recorded every 15 minutes by various smart devices and appliances. Main functions such as energy consumption, space status, planning level and temperature. Using advanced data analysis technology, the study aims to accurately predict energy consumption. These predictions help manage energy more efficiently, resulting in significant cost savings and environmental benefits. This study highlights the importance of feature engineering and higher-order modeling in achieving robust predictions. Ultimately, this study contributes to the development of smarter energy management strategies in IoT-activated houses, promoting sustainable and efficient energy consumption.

## 1.2 Aim

The aim is to employ advanced machine learning techniques for precise energy consumption forecasting, aiding in optimizing energy management strategies for IoT-equipped smart homes.

## 1.3 Research Motivation

The motivation behind this research stems from a keen interest in analyzing how IoT devices can optimize energy usage in smart home environments. By investigating various strategies to save energy and reduce costs through comprehensive data analysis, the study aims to uncover effective methods for enhancing energy efficiency. Additionally, there is a strong focus on understanding how improved energy management practices can contribute to reducing environmental footprints.

The overarching goal is to develop insights that not only promote cost savings and energy conservation but also support broader sustainability efforts by minimizing the environmental impact of energy consumption in smart homes.

## 1.4 Research Objective

The purpose of this study is to explore energy consumption patterns in smart homes, focusing on the complexity of temporal and seasonal variations. Utilizing advanced machine learning techniques, this study examines extensive historical and environmental data to develop highly accurate predictive models. The evaluation will encompass a variety of metrics, including precision, accuracy, recall, and F1 scores, to provide a thorough assessment of the models' performance. These measurements will help optimize dynamic energy management strategies, making them more efficient and adaptable to immediate changes. The aim of this research is to improve preprocessing techniques to ensure robust data processing. This includes complex feature engineering to extract meaningful properties, normalization to properly scale the profile, and advanced techniques to resolve missing values to ensure profile completeness and consistency. The goal of improving these preprocessing steps is to increase the predictive accuracy and reliability of the models. Ultimately, this research aims to promote smarter and more efficient energy management in smart homes, promoting energy conservation and sustainability.

## 1.5 Research Questions

- How does energy usage in IoT devices vary over time, and what factors influence these patterns?
- Can machine learning accurately predict energy usage in IoT devices based on historical data and external factors?
- What strategies optimize energy usage in IoT devices while ensuring performance and user satisfaction?

- How does the efficiency of energy usage in IoT devices change with varying environmental conditions, such as temperature or humidity?
- What are the key features and data preprocessing steps that significantly impact the accuracy of energy usage predictions in IoT devices.

## 1.6 Report Structure

The structure of the report is organized into six distinct chapters, each addressing key aspects of the research problem, and is described as follows:

a. **Introduction:** Chapter 1 introduces the research problem, focusing on energy consumption patterns within smart home environments. This chapter provides an overview of the research, including the objectives, aims, and research questions, as well as the context and background of the study.

b. **Literature Review:** Chapter 2 presents a comprehensive review of the existing literature in the field. It examines previous studies on IoT-based energy consumption and machine learning, provides a comparative analysis of various models, and discusses the implementation and evaluation of these models. The chapter also addresses energy optimization strategies, challenges, limitations, research gaps, and justifies the model selection.

c. **Methodology:** Chapter 3 details the methodology used to address the research problem. This includes an introduction to the methodological approach, an understanding of the data, and detailed descriptions of the data preprocessing steps such as handling missing values, feature engineering, normalization, and target variable transformation. The chapter also outlines the model implementation process and the evaluation metrics used.

d. **Result Analysis:** Chapter 4 discusses the results obtained from the research. It evaluates the

performance of the models, presents cross-validation results, and provides a detailed breakdown of the heatmap metrics used to assess the models' effectiveness.

**e. Conclusion and Future Work:** Chapter 5 summarizes the findings of the research, discusses the implications of the results, and provides conclusions. It also outlines potential areas for future research, suggesting ways to further optimize energy management strategies in smart home environments.

## Chapter 2: Literature Review

The rapidly growing field of the Internet of Things (IoT) has generated significant interest [1] due to its vast potential in various sectors, including smart homes, industrial automation, and healthcare. One of the critical aspects of IoT is energy consumption, especially in scenarios where IoT devices are expected to operate continuously and transmit large amounts of data [2]. Machine learning (ML) models such as ARIMA and LSTM have been explored extensively to predict and optimize energy usage in IoT devices [4], aiming to enhance energy efficiency and operational effectiveness. This literature review examines the current research landscape, focusing on the application of different models in energy forecasting for IoT devices, their comparative performance, and the methodologies used to achieve energy savings.

### 2.1 IoT-Based Energy Consumption and Machine Learning

IoT devices, which include sensors, actuators, and other interconnected systems, are integral to modern technological ecosystems. However, these devices often face significant energy constraints, necessitating efficient data transmission and processing mechanisms. Traditional methods such as ARIMA (Autoregressive Integrated Moving Average) and newer approaches like LSTM (Long Short-Term Memory) have been pivotal in forecasting and managing energy consumption.

A study by Nahid Ferdous Aurna et al. (2021) [3] presented an IoT architecture with temperature and humidity sensors, transmitting data to a server for processing using ARIMA and LSTM models. The results indicated that LSTM models significantly outperformed ARIMA in terms of prediction accuracy, demonstrating a substantial reduction in energy consumption by minimizing unnecessary data transmissions.

ARIMA has long been a standard for time series forecasting due to its simplicity and effectiveness in linear data trends. However, its limitations become apparent with non-linear and complex datasets, where it often fails to capture intricate patterns. In contrast, LSTM, a type of recurrent neural network (RNN), excels in handling sequential data and maintaining long-term dependencies, making it suitable for more complex time series data.

Malki, A. et al. (2022) [13] conducted a comparative analysis of ARIMA and LSTM in forecasting economic and financial time series. Their study revealed that LSTM models consistently outperformed ARIMA, reducing error rates by 84-87% on average. This superior performance is attributed to LSTM's ability to learn from long sequences of data, adapt to non-linear patterns, and handle varying time dependencies.

## 2.2 Implementation and Evaluation of Models

The implementation of these models involves several critical steps, including data collection, preprocessing, and model training. For instance, the study by Aurna et al. [3] utilized temperature and humidity sensors integrated with NodeMCU and Raspberry Pi for data collection. The data was then processed and split into training and testing sets, with ARIMA and LSTM models applied to forecast future values. The evaluation metrics used included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), with LSTM consistently showing better performance across these metrics.

The study by Balaji, S. et al. (2023) [4] highlighted the practical benefits of applying machine learning in IoT environments by demonstrating that predictive models could reduce energy consumption by more than half. Their research utilized temperature and humidity sensors

integrated with NodeMCU and Raspberry Pi for data collection. The data was processed and split into training and testing sets, with ARIMA and LSTM models applied to forecast future values. The evaluation metrics used included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), with LSTM consistently showing better performance across these metrics.

Another study by Nakamura, S. et al. (2023) [14] focused on the application of machine learning models to predict energy consumption in smart buildings. They utilized a dataset comprising various environmental and operational parameters, including temperature, humidity, and occupancy levels. The study compared the performance of traditional statistical models like ARIMA with advanced machine learning techniques such as LSTM and Convolutional Neural Networks (CNN). The results demonstrated that LSTM outperformed ARIMA and other models in terms of accuracy and robustness, particularly in capturing the temporal dependencies and non-linear relationships inherent in the data.

### 2.3 Energy Optimization Strategies

Predictive models like ARIMA and LSTM play a crucial role in optimizing energy usage in IoT devices. By accurately forecasting energy consumption patterns, these models enable dynamic energy management, reducing the need for continuous data transmission and thereby conserving energy. The study by Aurna et al. demonstrated that predictive models could reduce energy consumption by more than half, highlighting the practical benefits of applying machine learning in IoT environments.

In the context of smart buildings, energy optimization involves implementing strategies that minimize energy consumption while maintaining comfort and operational efficiency. Machine

learning models facilitate this by predicting future energy needs and adjusting control systems accordingly. For instance, predictive models can be used to schedule HVAC (heating, ventilation, and air conditioning) systems based on occupancy patterns and weather forecasts, ensuring that energy is used efficiently without compromising comfort.

A study by Cviti, I. et al. (2021) [5] explored the use of deep learning models, including LSTM, to predict energy consumption in commercial buildings. The researchers developed a framework that integrated predictive analytics with building automation systems to optimize energy usage. The study demonstrated significant energy savings by adjusting HVAC operations based on LSTM predictions, thereby reducing unnecessary energy consumption during unoccupied periods. The study of Shapi, Ramli and Awalim (2021) [18] tested forecasting energy consumption in smart buildings in Malaysia using machine learning algorithms. The researchers used three different models: supporting the vector machine (SVM), the artificial neural network (Ann) and the newest neighbor (K-NN). These models were implemented on Microsoft's Azure Machine Learning platform, which is known for its ease of use and powerful computing capabilities. The performance of each model was evaluated using metrics such as Root Mean Square Error (RMSE), Normalized RMSE (NRMSE) and Mean Absolute Percentage Error (MAPE). The results show that the effectiveness of the model varies depending on the data set and the energy consumption of specific tenants. The SVM models, especially those with radial basis function (RBF) kernels, outperformed the other models for both tenants, achieving RMSE values of 4.75 and 3.59, respectively. Compared to other models, this model shows higher accuracy and lower error speeds. The K-NN model is excellent among the other two tenants. Although the ANN model is useful, it often performs poorly compared to SVM and k-NN, especially for predicting the energy consumption of tenants with more complex and variable usage patterns. The study concludes that SVM and k-

NN models are highly effective in predicting energy consumption in smart buildings, but the optimal model may vary depending on the specific characteristics of the dataset and available computing resources.

Table 1: Literature Review

<b>Title</b>	<b>Year</b>	<b>Contribution</b>	<b>Research Gap</b>
Time Series Forecasting using LSTM and ARIMA	2023	The paper compares LSTM and ARIMA for stock price prediction, aiming to determine which model performs better. It demonstrates the practical use of these models in financial forecasting.	The research lacks clarity on dataset preprocessing, hindering transparency. Its limited scope neglects broader applicability, limiting its relevance and generalizability across diverse financial datasets.
Time-Series Forecasting to Fill Missing Data in IoT Sensor Data	2023	The study compares statistical and deep-learning methods for IoT data forecasting, favoring Holt-Winters for temperature and CO <sub>2</sub> , and LSTM for humidity.	The research lacks detailed performance metric insights for comparing statistical and ML models in IoT data forecasting.
Machine learning approach of detecting anomalies and forecasting time-series of IoT devices.	2022	The study investigates ML for power system anomaly detection, preferring Prophet and LightGBM over Vector Autoregressive models. They predict future energy consumption.	Insufficient detail on data collection quality and preprocessing steps undermines study transparency.

Energy Prediction in IoT Systems Using Machine Learning Models	2022	IoT-driven energy forecast system uses spatial databases, stream analysis, and CNNs to improve accuracy, reduce delays, and enhance adaptability.	It doesn't fully address real-time adaptation for energy management.
Energy consumption prediction by using machine learning for smart building: Case study in Malaysia.	2021	The research compares k-NN, SVM, and ANN for predictive modeling. AzureML is used for model development, analysis, and evaluation.	Paper narrows scope to 3 algorithms, neglecting discussion on other methods.

## 2.4 Challenges and Limitations

Despite the promising results, several challenges and limitations need to be addressed to fully harness the potential of machine learning models in IoT energy management. One of the primary challenges is the scalability of these models. Most existing studies have been conducted on small to medium-sized datasets, which may not fully capture the complexity and variability of large-scale IoT deployments. There is a need for research that validates the models on extensive datasets, encompassing a broader range of environmental and operational conditions.

Data quality and preprocessing are also critical issues. IoT devices often generate noisy and incomplete data, which can adversely affect model performance. Effective data preprocessing techniques, including data cleaning, imputation, and feature engineering, are essential to ensure the accuracy and reliability of predictive models. Future research should focus on developing

robust preprocessing pipelines that can handle the challenges posed by IoT data.

Another limitation is the interpretability of machine learning models, particularly deep learning techniques like LSTM. While these models offer high accuracy, their "black-box" nature makes it difficult to understand the underlying decision-making process. This lack of transparency can be a barrier to adoption, particularly in applications where explainability is crucial. Research efforts should be directed towards developing explainable AI techniques that provide insights into the model's workings and build trust among users.

## 2.5 Addressing Research Gaps

Several strategies can be adopted to address the identified research gaps. First, exploring hybrid models combining the strengths of different machine learning algorithms can improve accuracy and reliability. For example, integrating LSTM with other methods such as ARIMA or ensemble methods can leverage the complementary benefits of different approaches, thereby improving overall performance.

Scalability can be addressed by conducting large-scale studies that test models on large datasets. This helps ensure that the model is robust and reliable when applied to real-world scenarios with diverse and dynamic data streams. Additionally, considering more environmental factors such as wind speed, solar radiation, and occupancy patterns provides a more comprehensive understanding of their impact on energy consumption, leading to more accurate forecasts and better energy management. It leads to strategy. The development of real-time prediction systems is also an important area of future research. Although the potential of real-time energy management has been recognized, practical implementations that integrate these predictive models into live IoT environments are lacking. Real systems that dynamically

correct energy consumption based on an understanding of forecasts can considerably increase energy efficiency, offering immediate feedback and control to eliminate anomalies and continuously optimize the use of energy.

Extended methods of functions of functions can also be studied to enter more significant models in data. This includes interaction terms, polynomial features, and domain-specific transformations that can improve model performance by revealing complex relationships within your dataset. Additionally, explainable AI techniques need to be developed to provide transparency and insight into the decision-making process of machine learning models, build trust, and enable effective use in real-world applications.

## 2.6 Justification for Model Selection

The models chosen for this research—Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost—are particularly effective in addressing the deficiencies and research voids noted in earlier studies. Random Forest and XGBoost, as ensemble learning techniques, are celebrated for their robustness and capacity to manage complex nonlinear relationships within data. These models are adept at navigating the expansive feature spaces often found in IoT datasets, delivering superior predictive performance over traditional models like ARIMA. Their ability to highlight the significance of different predictors also enhances the interpretability of the results.

SVM is selected for its proficiency in handling high-dimensional spaces, particularly useful when the number of dimensions surpasses the number of samples. Its capability to identify the optimal hyperplane for classification makes it highly accurate, aligning well with the need for

precise energy management strategies. Meanwhile, Logistic Regression, though simpler, offers transparency and ease of interpretation. This model sets a benchmark, allowing complex models to be assessed against straightforward, interpretable standards. Its inclusion supports the drive for explainable artificial intelligence, with coefficients that clearly outline the relationships between features and the target variable.

These models also align with the demands for scalable, real-time prediction systems. Techniques like Random Forest and XGBoost can be efficiently parallelized, making them fit for large-scale IoT implementations. Their effectiveness with large datasets ensures they can handle the variability and dynamics typical of real-world IoT data streams, thus addressing the limitations often found in smaller-scale studies. Moreover, implementing these models facilitates the exploration of hybrid approaches, where combining different algorithms may yield more reliable and accurate predictions. For instance, a hybrid model that merges XGBoost's precision with Logistic Regression's interpretability could offer both high performance and clarity, bridging many existing research gaps.

Therefore, the models selected are not just capable of achieving high predictive accuracy but also offer robustness, scalability, and interpretability. Their application in this study aims to tackle the identified gaps in previous research, providing a thorough and actionable approach to enhancing energy usage in IoT devices. This choice substantiates their suitability for this study, promising significant advancements in the domain of energy management within IoT-enhanced environments.

This review underscores the progress made in applying machine learning models to energy forecasting in IoT devices, with a focus on the comparative effectiveness of ARIMA and LSTM

models. While LSTM shows greater accuracy and capability in handling complex data, substantial research gaps remain. Future studies should explore developing hybrid models, undertaking larger-scale experiments, crafting real-time prediction systems, and enhancing feature engineering techniques. Moreover, there should be a concerted effort to boost the interpretability and transparency of these models, ensuring they are more viable in practical scenarios. By addressing these gaps, the field can progress towards more precise and efficient energy management strategies for IoT devices, contributing significantly to sustainability and operational efficiency across various sectors.

## Chapter 3: Methodology

This chapter outlines the systematic approach employed to analyze IoT data for predicting energy usage using machine learning techniques. It covers the steps from data understanding and preprocessing to model implementation and evaluation. By following a structured methodology, the research aims to ensure the reliability and validity of the findings, ultimately deriving meaningful insights that can aid in efficient energy management in smart homes.

### 3.1 Data Understanding

The dataset used in this study was obtained from a smart home located in Greece [20]. This smart home is equipped with various sensors and IoT devices that monitor different aspects of energy usage. The dataset includes a wide range of features such as room status, dimming levels, luminance, and temperature readings, alongside the target variable KWh\_S\_total, which represents the total energy usage. Data was collected at 15-minute intervals from January 2021 to December 2022, resulting in a comprehensive dataset with 66,619 records and 19 features. Understanding the existing data is crucial as it forms the foundation for subsequent analyses and model development.

Table 2: Dataset Overview

<b>Censor</b>	<b>Symbolic Naming</b>	<b>Measurement Unit</b>
<b>Electricity Consumption</b>	KWh_S_total	kWh
<b>Air-condition Status</b>	status_room_0 status_room_1	-

	status_room_2 status_room_3	
<b>Luminance</b>	luminance_room_0 luminance_room_1 luminance_room_2 luminance_room_3	Lux
<b>Light Dimming</b>	dimming_room_0 dimming_room_1 dimming_room_2 dimming_room_3	%
<b>Indoor Temperature</b>	temperature_room_0 temperature_room_1 temperature_room_2 temperature_room_3	°C
<b>Outdoor Temperature</b>	airTemperature	°C

### 3.2 Data Preprocessing

Data preprocessing is crucial for this research as it ensures the accuracy and reliability of the predictive models by handling missing values, performing feature engineering, and normalizing the data. Proper preprocessing enhances the quality of the input data, which

directly impacts the performance and robustness of machine learning models. Additionally, effective preprocessing helps in uncovering meaningful patterns and insights from the raw data, leading to more accurate and actionable energy consumption predictions in smart home environments.

### 3.2.1 Handling Missing Values

The initial step in preprocessing involved addressing missing values in the dataset. Missing data can lead to inaccuracies in model predictions and analyses. In this study, missing values were handled by replacing them with the mean of the respective feature. This method ensures that no data is lost and maintains the integrity of the dataset.

### 3.2.2 Feature Development

This crucial phase involves creating or adjusting existing variables to more precisely capture the unique patterns in your data. For this study, time-related features such as the time of day, day of the week, and month were derived from the eventDate column. These features are key to recognizing temporal patterns in energy usage, which is vital for accurate predictions.

### 3.2.3 Standardization

Standardization involves scaling features so their average is zero and their standard deviation is one. This process is important as it ensures every feature has an equal impact on the model's effectiveness. In this study, all features underwent normalization using standard scaling techniques, enhancing the model's efficiency and precision.

### 3.2.4 Converting Target Variables

The target variable, KWh\_S\_total, which represents total energy consumption, was categorized

into “low”, “medium”, and “high” based on quantile sampling. This categorization simplifies the classification task and enables the use of the model for predicting different levels of energy consumption.

### 3.3 Model Implementation

Model implementation in this research involves developing and deploying advanced machine learning models, specifically Random Forest Classifier (RFC), Support Vector Machine (SVM), XGBoost, and Logistic Regression, to predict energy consumption patterns. This process includes training these models on preprocessed historical and environmental data to ensure they accurately capture temporal and seasonal variations. Implementing these diverse models is crucial for identifying the most effective approach to optimize energy management in smart homes.

#### 3.3.1 Random Forest Classifier

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees. It is robust to overfitting and can handle high-dimensional data efficiently. In this study, the Random Forest model was trained and evaluated on the dataset, achieving high accuracy. The robustness and ability to capture complex interactions among features made Random Forest a suitable choice for this analysis.

Fig 1: Random Forest Classifier

```

Random Forest F1-Score: 0.9648004562274783
Random Forest Classification Report:

```

	precision	recall	f1-score	support
0	0.957088	0.958165	0.957626	4446.0000
1	0.958617	0.957659	0.958138	5007.0000
2	0.981658	0.981658	0.981658	3871.0000
accuracy	0.964800	0.964800	0.964800	0.9648
macro avg	0.965788	0.965827	0.965807	13324.0000
weighted avg	0.964801	0.964800	0.964800	13324.0000

Figure 1: Random Forest Classifier

### 3.3.2 Support Vector Machine (SVM)

Vector Support Machine (SVM) is a powerful classification method that works well in ROG proud areas, and is effective when the number of measurements exceeds the number of samples. SVM is trying to find Hyperplane, the best separation of the class. This model has been used to classify energy consumption patterns, taking advantage of its advantages in processing complex datasets with a clear boundary.

```

SVM F1-Score: 0.8577324119882666
SVM Classification Report:

```

	precision	recall	f1-score	support
0	0.873126	0.851327	0.862089	4446.000000
1	0.820600	0.868784	0.844005	5007.000000
2	0.892082	0.849910	0.870486	3871.000000
accuracy	0.857475	0.857475	0.857475	0.857475
macro avg	0.861936	0.856673	0.858860	13324.000000
weighted avg	0.858895	0.857475	0.857732	13324.000000

Figure 2: Support Vector Machine (SVM)

### 3.3.3 Logistic Regression

Logistic Regression is a linear model used for binary classification, but it can be extended to multiclass classification problems. It models the probability of a certain class or event. Despite being a simple model, Logistic Regression is powerful and provides a good baseline for comparison with more complex models. It was utilized in this study to classify the energy usage into predefined categories, demonstrating its effectiveness in simpler, linear problems.

```

Logistic Regression F1-Score: 0.6438154425180401
Logistic Regression Classification Report:

```

	precision	recall	f1-score	support
0	0.662465	0.579172	0.618025	4446.000000
1	0.654240	0.701019	0.676822	5007.000000
2	0.615177	0.647120	0.630744	3871.000000
accuracy	0.644701	0.644701	0.644701	0.644701
macro avg	0.643961	0.642437	0.641864	13324.000000
weighted avg	0.645636	0.644701	0.643815	13324.000000

*Figure 3: Logistic Regression*

### 3.3.4 XGBoost

XGBoost is an optimized distributed level-boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms within the Gradient Boosting platform. XGBoost was chosen because of its superior performance and speed. In this study, his XGBoost model was implemented to solve a prediction problem by leveraging its ability to manage large datasets and complex patterns.

```

XGBoost F1-Score: 0.9566873447439743
XGBoost Classification Report:

```

	precision	recall	f1-score	support
0	0.946318	0.959514	0.952870	4446.000000
1	0.953728	0.942680	0.948172	5007.000000
2	0.972589	0.971584	0.972086	3871.000000
accuracy	0.956695	0.956695	0.956695	0.956695
macro avg	0.957545	0.957926	0.957709	13324.000000
weighted avg	0.956735	0.956695	0.956687	13324.000000

*Figure 4: XGBoost*

### 3.4 Evaluation Metrics

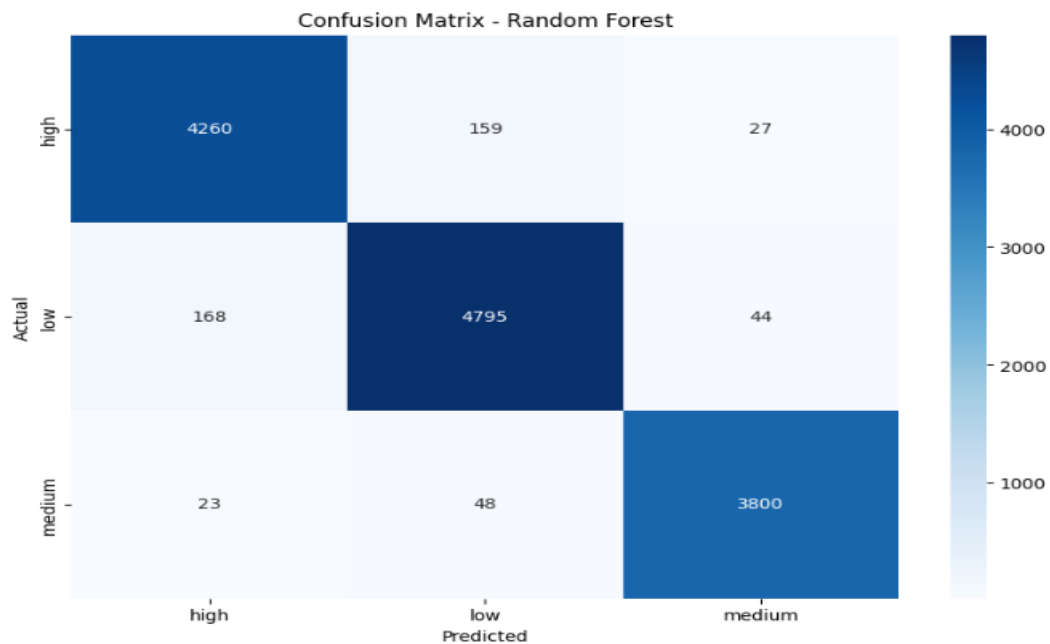
Analyzing model performance is important to understand their reliability and effectiveness.

The models were evaluated according to a number of indicators:

- Accuracy: evaluates the overall accuracy of the model.
- Precision: Indicates how accurate the optimistic prediction is.
- Recall: Evaluate how the model finds all the corresponding cases.
- F1 score: Balance between recall and precision. Calculated as the harmonic mean of the two.
- Confusion matrix: A table that compares the actual classification to the predicted classification to illustrate the performance of the classification model. These metrics evaluated the performance of each model in detail and highlighted both its strengths and weaknesses.

#### 3.4.1 Breakdown of the Confusion Matrix – Random Forest

The confusion matrix for the Random Forest model, as shown in the provided image, is a critical tool for understanding the performance of the model in classifying energy usage into three categories: high, low, and medium. Each cell in the matrix indicates the number of predictions made by the model, with rows representing the actual classes and columns representing the predicted classes.



*Figure 5: Confusion Matrix – Random Forest*

- High Usage (First Row)

True Positives (4260): The model correctly identified 4260 instances of high energy usage.

False Negatives (186): These are instances where the actual high usage was misclassified by the model. Specifically, 159 instances were misclassified as low, and 27 instances were misclassified as medium. This indicates that the model has a high degree of accuracy in predicting high energy usage but still has room for improvement.

- Low Usage (Second Row)

True Positives (4795): The model accurately classified 4795 instances as low energy usage.

False Negatives (212): These are instances where the actual low usage was incorrectly predicted. Specifically, 168 instances were predicted as high, and 44 instances were predicted as medium. This suggests that the model is highly effective at identifying low energy usage, though there are some misclassifications.

- Medium Usage (Third Row)

True Positives (3800): The model correctly identified 3800 instances of medium energy usage.

False Negatives (71): These are instances where the actual medium usage was misclassified.

Specifically, 23 instances were misclassified as high, and 48 instances were misclassified as low. This indicates that while the model performs well, it has the most difficulty in accurately predicting medium energy usage compared to high and low usage.

### **Overall Performance Insights**

Substantial plus (accuracy). The high number of true positives in all three categories indicates that the random forest model is very accurate overall. The true positive rates for high, low, and medium categories are high, indicating the robustness and reliability of the model in predicting energy consumption. False negatives (misclassifications). Although the false negative rate is relatively low, it highlights an area where the model is having trouble. The higher number of false negatives in the medium category compared to the high and low categories suggests that the model is having difficulty distinguishing medium usage from the other two categories.

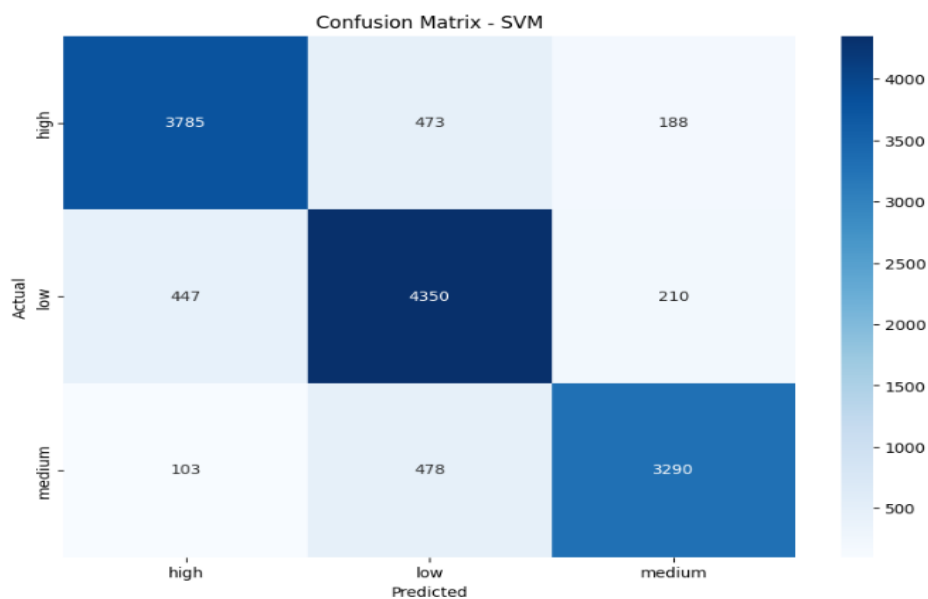
Misclassification analysis: Misclassification between high and low utilization (159 instances) and low and high utilization (168 instances) suggests that there are overlapping characteristics between these two categories. and can lead to model confusion.

The misclassification between the medium category and the other two categories (total of 71 cases) indicates that medium usage has characteristics that lie between high and low usage, making the model correct. It's becoming more difficult to categorize. Practical meaning: The insight obtained from the confusion matrix is important to further improve the model. For example, a relatively highly miscellaneous classification of a moderate category may require

additional functions or more sophisticated function engineering to distinguish between medium and low usage. Additionally, the high accuracy in predicting low and high usage indicates that the model is quite reliable for these categories, which can be particularly useful for applications where distinguishing between these two extremes is critical for energy management and optimization.

### 3.4.2 Breakdown of the Confusion Matrix – SVM

The confusion matrix for the Support Vector Machine (SVM) model provides a comprehensive view of the model's performance across three categories of energy usage: high, low, and medium. Each cell in the matrix represents the number of instances that were correctly or incorrectly classified by the model. The rows correspond to the actual classes, while the columns correspond to the predicted classes.



*Figure 6: Confusion Matrix – SVM*

- High Usage (First Row)

True Positives (3785): The model correctly identified 3785 instances of high energy usage.

False Negatives (661): These are instances where the actual high usage was misclassified. Specifically, 473 instances were misclassified as low, and 188 instances were misclassified as medium. This indicates that while the model performs reasonably well in identifying high usage, it still misclassifies a notable number of instances.

- Low utilization (second row)

True positives (4,350): The model correctly classified 4,350 cases as low power.

False negatives (657): These are cases where actual usage was incorrectly predicted to be low. Specifically, 447 were predicted to be high and 210 were predicted to be moderate. This suggests that although this model is effective in identifying low usage, it sometimes encounters problems that lead to misclassification.

- Average usage (line 3)

True Positives (3290): The model correctly identified 3290 cases of average power consumption.

False negative (581). These are cases where the actual use of the media has been misclassified. Specifically, 103 instances were misclassified as high, and 478 instances were misclassified as low. This indicates that the model finds it challenging to distinguish medium usage from the other categories, particularly low usage.

### **Overall Performance Insights**

True Positives (Accuracy): The model's accuracy in predicting energy usage is demonstrated

by the high number of true positives it displays across all categories. Accuracy varies, though, with fewer true positives shown in the medium usage category.

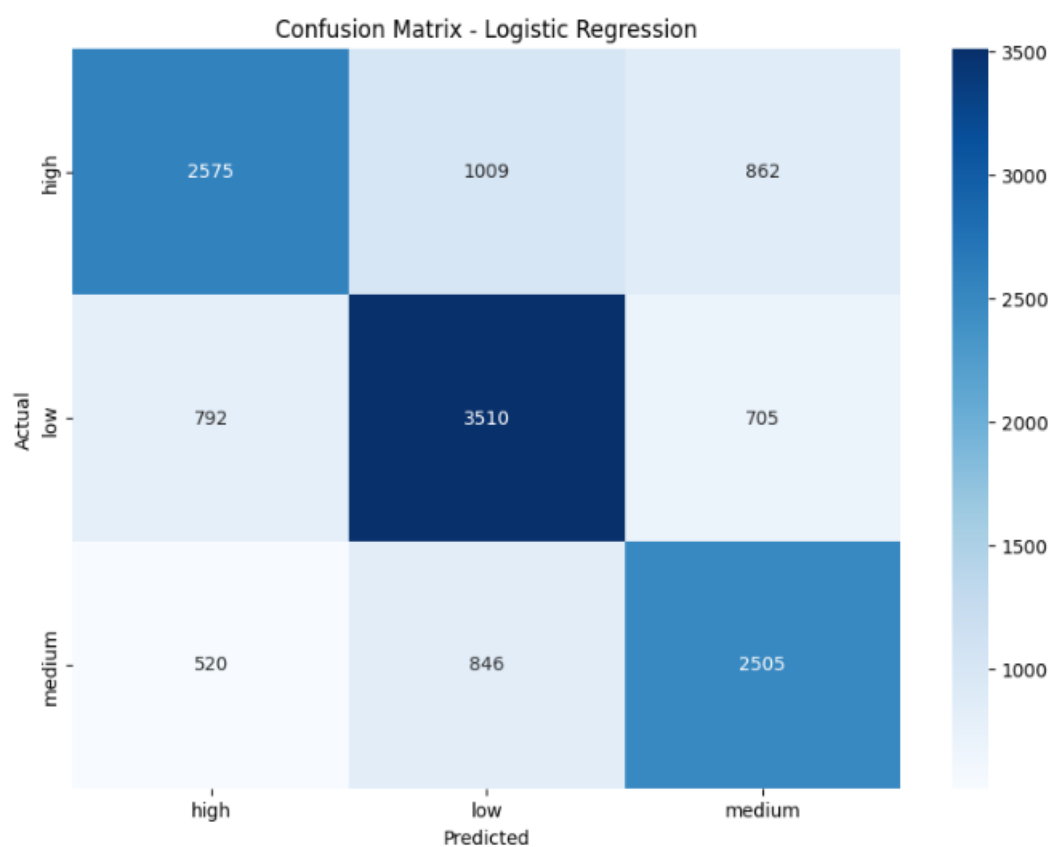
**False Negatives (Misclassification):** The model appears to have more trouble differentiating medium usage from the high and low categories, as evidenced by the higher number of false negatives in the medium category.

**Misclassification Patterns:** There are a significant number of misclassifications (473 instances) between high and low usage, and vice versa (447 instances), suggesting that these categories share characteristics that could potentially confuse the model. Furthermore, a sizable portion of medium usage cases (478 instances) are mistakenly categorized as low, indicating difficulties in distinguishing between these groups using the available features.

**Practical Implications:** The model's improvement depends heavily on the information from the confusion matrix. The frequency of incorrect classifications in the medium category indicates that in order to effectively distinguish between the categories, either more features or more sophisticated feature engineering are required. Improving the model's capacity to discriminate between high and low usage may also increase the accuracy of the whole thing. Strong performance is demonstrated by the model's high true positives rate in all categories, but the misclassifications point to areas that could be improved. These errors might be reduced by implementing more sophisticated feature engineering or by adding new features. Guidance for future improvements and model tuning is provided by the confusion matrix, which provides insightful information about the model's advantages and disadvantages.

### 3.4.3 Breakdown of the Confusion Matrix – Logistic Regression

The confusion matrix for the Logistic Regression model provides a detailed analysis of its performance across three categories of energy usage: high, low, and medium. Each cell in the matrix represents the number of instances that were correctly or incorrectly classified by the model. The rows correspond to the actual classes, while the columns correspond to the predicted classes.



*Figure 7: Confusion Matrix – Logistic Regression*

- High Usage (First Row)

**True Positives (2575):** The model accurately identified 2575 cases of high energy consumption.

**False Negatives (1871):** A significant number of high usage instances were incorrectly classified, with 1009 labeled as low and 862 as medium. This highlights a major issue with the

model's ability to correctly identify high energy usage.

- Low Usage (Second Row)

True Positives (3510): The model correctly classified 3510 instances of low energy usage.

False Negatives (1497): Misclassifications include 792 instances incorrectly predicted as high and 705 as medium, showing the model's struggles with low usage detection, though it performs relatively better than with high usage.

- Medium Usage (Third Row)

True Positives (2505): The model correctly spotted 2505 medium energy usage cases.

False Negatives (1366): Misclassifications involved 520 instances marked as high and 846 as low, indicating difficulty in accurately distinguishing medium usage from other categories.

### **Overall Performance Insights**

True Positives (Accuracy): The true positives across all categories suggest moderate accuracy. However, the considerable number of false negatives, especially in the high and medium categories, indicates significant potential for improvement.

False Negatives (Misclassification): The model notably struggles with high and medium usage levels, often failing to distinguish them accurately.

Misclassification Patterns: There is substantial confusion between the high and low categories, with 1009 instances of high misclassified as low and 792 of low as high. Medium usage is also frequently mislabeled, with many instances mistaken for high or low usage.

Practical Implications: The data points to a need for refining the model. Given the extensive misclassifications, integrating more sophisticated features or advanced modeling techniques

could enhance accuracy. The current model's simplicity might not suffice due to the non-linear relationships in the data, suggesting that more complex approaches like ensemble methods or deep learning might be more effective.

### 3.4.4 Breakdown of the Confusion Matrix - XGBoost

The confusion matrix for the XGBoost model provides a detailed analysis of its performance across three categories of energy usage: high, low, and medium. Each cell in the matrix represents the number of instances that were correctly or incorrectly classified by the model. The rows correspond to the actual classes, while the columns correspond to the predicted classes.

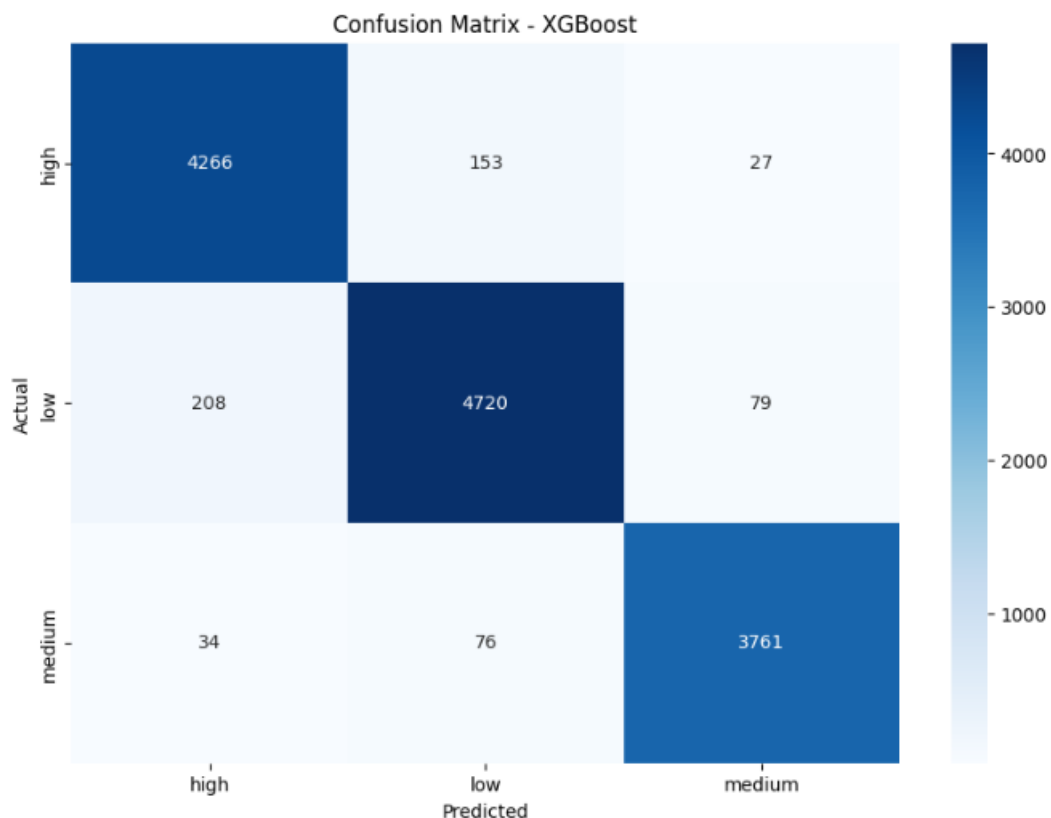


Figure 8: Confusion Matrix – XGBoost

- High Usage (First Row)

True Positives (4266): The model correctly identified 4266 instances of high energy usage.

False Negatives (180): These are instances where the actual high usage was misclassified. Specifically, 153 instances were misclassified as low, and 27 instances were misclassified as medium. This indicates that the model has a high degree of accuracy in predicting high energy usage, with relatively few misclassifications.

- Low Usage (Second Row)

True Positives (4720): The model accurately classified 4720 instances as low energy usage.

False Negatives (287): These are instances where the actual low usage was incorrectly predicted. Specifically, 208 instances were predicted as high, and 79 instances were predicted as medium. This suggests that the model is highly effective at identifying low usage, though there are still some errors.

- Medium Usage (Third Row)

True Positives (3761): The model correctly identified 3761 instances of medium energy usage.

False Negatives (110): These are instances where the actual medium usage was misclassified. Specifically, 34 instances were misclassified as high, and 76 instances were misclassified as low. This indicates that the model performs well in predicting medium usage but has slightly more difficulty compared to high and low usage.

### **Overall Performance Insights**

Substantial plus (accuracy). The large number of true positives in all three categories indicates that the XGBoost model is very accurate overall. The true positive rates for high, low, and

medium categories are high, indicating the robustness and reliability of the model in predicting energy consumption. False negative (misclassification). Although the false negative rate is relatively low, it highlights areas where the model struggles. The higher number of false negatives for the medium category compared to the high and low categories suggests that the model has more difficulty distinguishing between the medium category and the other two categories. Misclassification patterns. Misclassification between high and low use (153 cases) and between low and high use (208 cases) is due to overlapping characteristics between the two categories and the potential for them to be confused in the model. suggests that there is. The relatively low misclassification of high (34) and low (76) average usage indicates that the model can more effectively distinguish between average usage, but there is still room for improvement. . Practical Implications: The information obtained from the confusion matrix is important for further improving the model. The relatively high classification error for the medium category suggests that additional features or more advanced feature engineering may be needed to better distinguish between moderate and high and low usage. I am. Improving the model's ability to distinguish between these categories can further improve accuracy.

A high number of true positives in all categories indicates good model performance. However, the presence of classification errors indicates that there is room for improvement. For example, including additional features or using more advanced feature design techniques can reduce these errors. The confusion matrix gives a clear idea of where the model exceeds and where it can be improved by offering valuable information for the future development and optimization of the model.

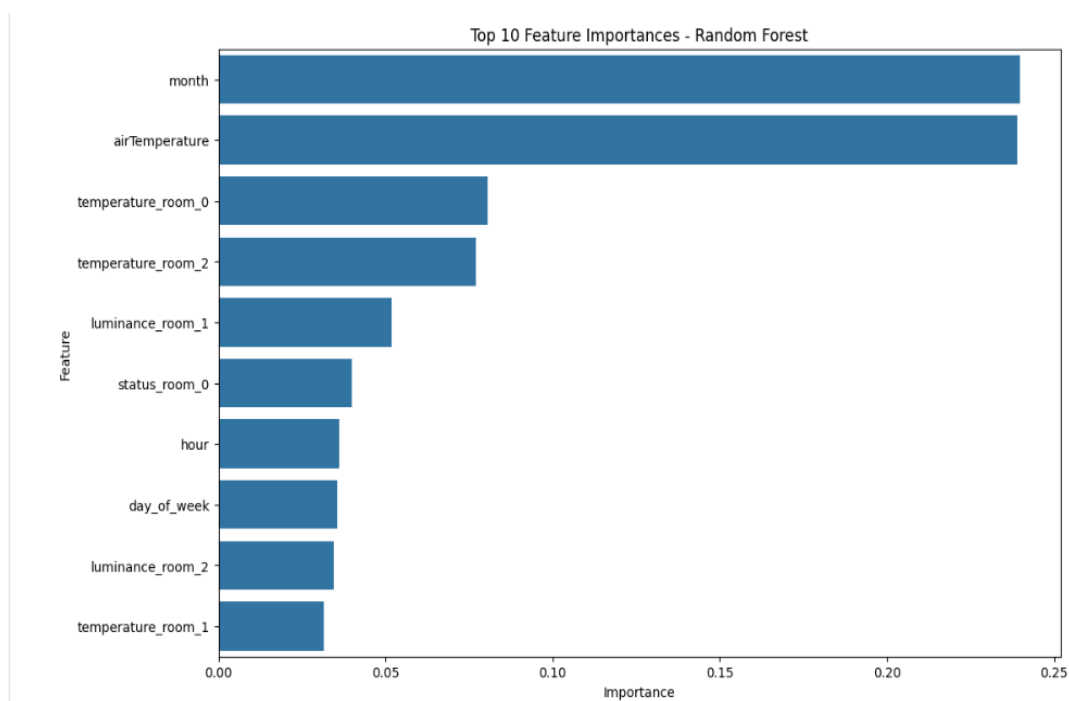
### **Cross -Validation and performance indicators**

Cross Validation is a method used to evaluate how well the model is generalized to an

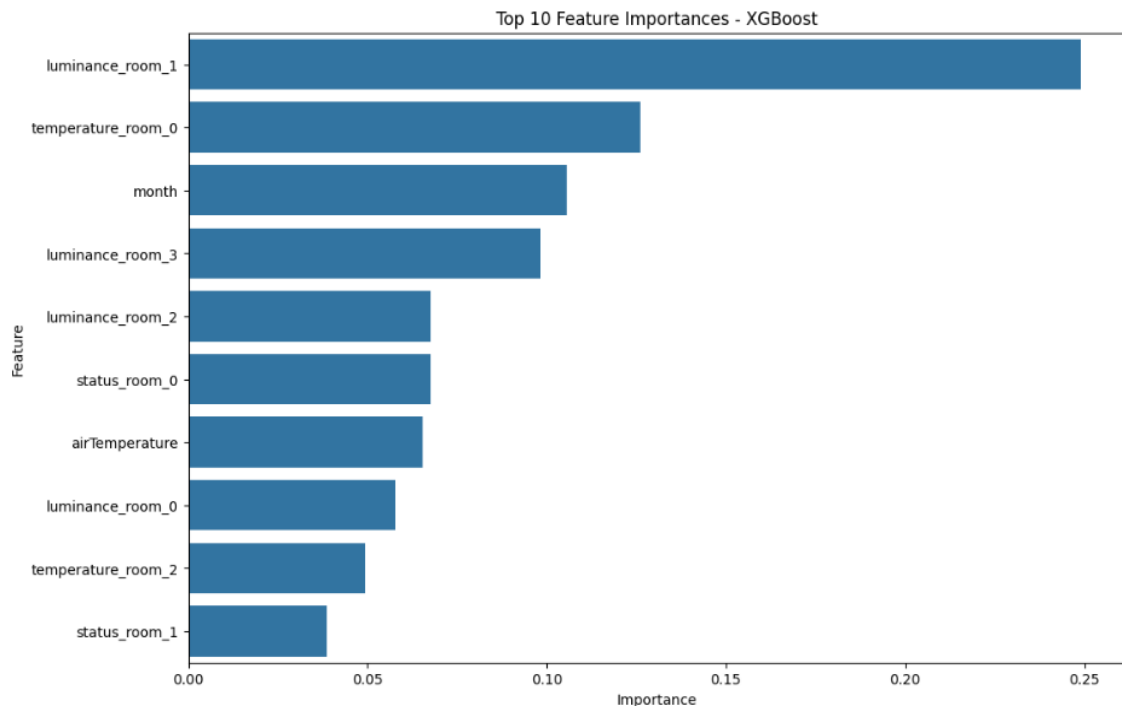
independent data set. This is a problem in which a series of data is divided with a complementary subset, a model is formed with a subset, and verification with other models. This process has been repeated several times, and the results are recognized to provide more accurate estimation of model performance. Cross-validation helps avoid overfitting and ensures model robustness.

### 3.5 Feature Importance

Understanding which features contribute most to the model's predictions is essential for interpreting the results and improving the model. Feature importance was analyzed for the Random Forest and XGBoost models, as they inherently provide this information. The analysis revealed which features had the most significant impact on energy usage predictions, guiding further refinements in model development.



*Figure 9: Feature Importance - Random Forest*



*Figure 10: Feature Importance - XGBoost*

## Random Forest

In the Random Forest model, the most important feature is the month, followed closely by airTemperature. This suggests that seasonal variations and ambient temperature are critical determinants of energy usage. The next most significant features are temperature\_room\_0 and temperature\_room\_2, indicating that the temperature in specific rooms also plays a vital role. Luminance\_room\_1 and status\_room\_0 are also highlighted, reflecting the impact of light levels and room occupancy status on energy consumption. Other notable features include the hour of the day, day of the week, and additional room-specific temperatures and luminance levels.

## XGBoost

The XGBoost model, on the other hand, places the highest importance on luminance\_room\_1, followed by temperature\_room\_0 and month. This highlights the significant role of light levels

in energy usage predictions. The presence of `luminance_room_3` and `luminance_room_2` among the top features further emphasizes the importance of lighting conditions. `Temperature_room_0` and `status_room_0` are also critical, similar to the Random Forest model, but `airTemperature` is ranked lower. Additionally, `temperature_room_2` and `status_room_1` are important, showcasing the model's consideration of room-specific conditions.

### **Comparative Insights**

Both models agree on the importance of certain features such as `temperature_room_0`, `month`, and `status_room_0`, indicating a consensus on the influence of these factors on energy usage. However, the Random Forest model gives more weight to overall air temperature, while XGBoost emphasizes room-specific luminance levels more heavily. This difference suggests that XGBoost might be capturing more localized variations in lighting conditions, whereas Random Forest considers broader environmental trends like seasonal changes and general temperature fluctuations.

Overall, the feature importance analysis from both models provides a comprehensive understanding of the factors influencing energy consumption. The emphasis on different features by each model underscores the value of using multiple algorithms to gain diverse perspectives and insights, ultimately leading to more robust and reliable predictions. Integrating these insights can help develop more effective energy management strategies for IoT-enabled environments.

## Chapter 4: Result Analysis

This chapter reviews the analysis results and evaluates the effectiveness of machine learning models used to predict energy consumption in smart homes equipped with IoT. We describe each model's precision, precision, recall, and F1 score, as well as the confusion matrix and important feature analysis. The findings will be taken into account in order to maximize energy management and raise smart home efficiency.

### 4.1 Model Performance

The random forest model performed admirably, achieving an overall accuracy of 96.48%. This model continuously performed admirably at power consumption levels that were low, medium, and high.

It received scores of 95.71% for low energy consumption, 95.86% for medium energy consumption, and 98.17% for high energy consumption. In the same respective categories, the test results showed similarly high scores of 95.82%, 95.76%, and 98.17%. These findings were corroborated by the F1 scores, which displayed balanced results of 95.76%, 95.81%, and 98.17%. The confusion matrix demonstrated the robustness of the model by confirming its accuracy with very few classification errors. Temperature, light level, and time were found to have a significant impact on energy consumption prediction through feature importance analysis.

Table 3: Model Performance Results

<b>Model</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>
Random Forest	0.964800	0.964801	0.964800	0.964800
SVM	0.857475	0.858895	0.857475	0.857732

Logistic Regression	0.644701	0.645636	0.644701	0.643815
XGBoost	0.956695	0.956735	0.956695	0.956687

The support vector machine (SVM) model achieved a good overall accuracy of 85.75%. Although it performed fairly well, it did not reach the efficiency observed with the Random Forest and XGBoost models. Specifically, it recorded accuracies of 87.31% for low energy consumption categories, 82.06% for medium and 89.20% for high energy consumption categories. Recall rates followed closely at 85.13% for low scores, 86.87% for medium scores, and 84.99% for high scores, and F1 scores of 86.21%, 84.40% and 87.05%, respectively. Confusion matrices revealed more frequent classification errors compared to random forests, suggesting that there may be room for improvement.

Logistic regression was a simpler and simpler model with an overall accuracy of 64.47%. This poor performance reflects issues related to complex interactions that exist within the data. The correct answer rate was 66.25% for the low category, 65.42% for the medium category, and 61.52% for the high category. Recall and F1 scores were similarly modest, with significant misclassification, especially in the low and high categories indicated by the confusion matrix, highlighting their limited usefulness in this context.

In contrast, the XGBoost model showed superior performance, nearly matching the performance of Random Forest with an overall accuracy of 95.67%. It achieved accuracy of 94.63% for low categories, 95.37% for medium categories, and 97.26% for high categories. Recall rates were high across the board: 95.95% for low, 94.27% for medium, and 97.15% for high, with F1 scores of 95.28%, 94.81% and 97.21% respectively. The XGBoost error matrix revealed minimal classification errors, highlighting its robustness and effectiveness. Feature importance analysis revealed that factors such as temperature, brightness, and weather-related

variables are important in predicting energy consumption, highlighting their important role in effective energy management.

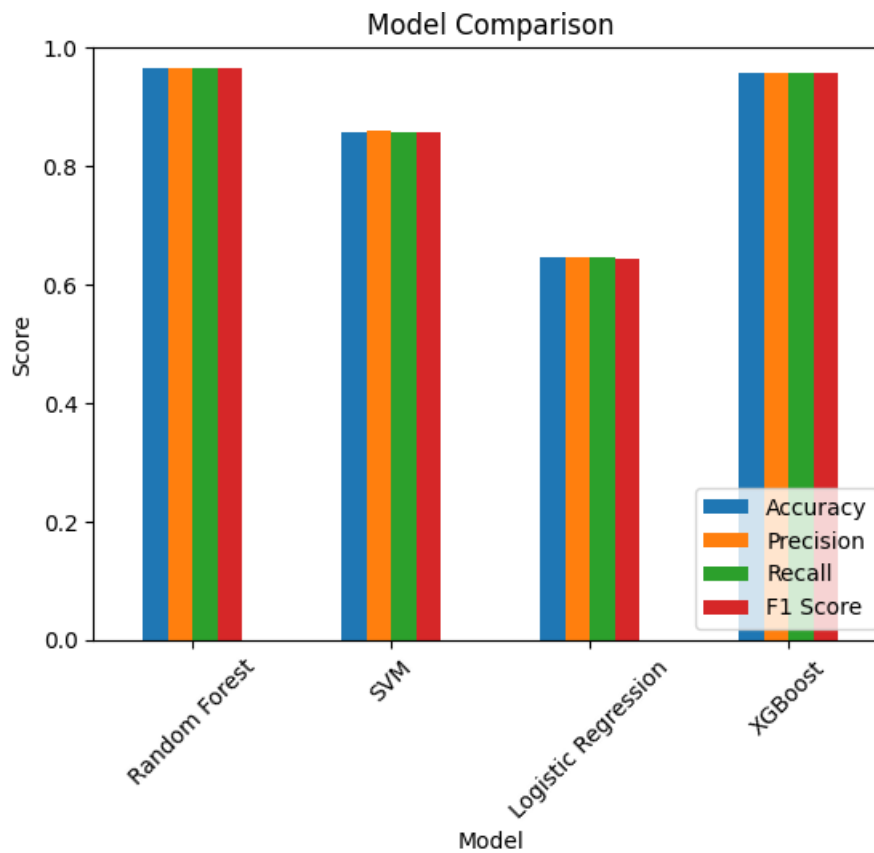
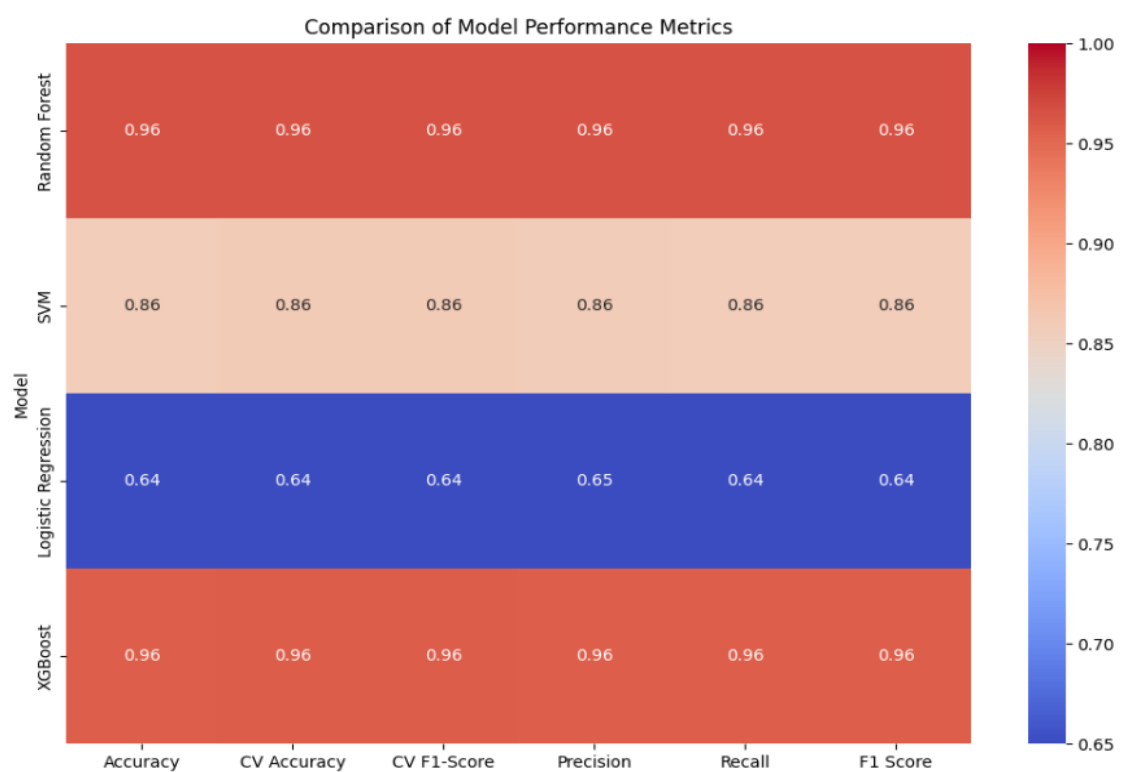


Figure 11: Model Comparison

## 4.2 Cross-Validation Results

To verify the robustness and validate the models' performance, cross-validation was used. To get a trustworthy estimate of each model's generalizability, the results were averaged over several folds. A cross-validation accuracy of 96.20% shows that the random forest model is consistent across different subsets of data. The cross-validation accuracy of the SVM model is 85.0%, indicating good performance but room for improvement. The 64.00 % cross verification accuracy of logistic regression emphasizes the shortcomings of methods when handling complex datasets. XGBOOST demonstrated its outstanding performance and reliability with a mutual verification accuracy of 95.30 %. Formacin learning models are visually compared on

the heat map. Random Forest, SVM (support vector machine), logistic regression, XGBOOST. The metrics that have been used in this comparison are the cross check (CV), the accuracy, the CV F1-Indicator, the accuracy, the examination and the F1 score. This complete comparison allows you to clearly understand the strengths and weaknesses of each model to predict the use of energy in IoT devices.



*Figure 12: Analysis of the Heatmap for Model Performance Metrics*

### 4.3 Breakdown of the Heatmap Metrics

Accuracy measures the ratio of correct predictions to total predictions, showing how often the model accurately forecasts outcomes. CV-Accuracy, which is derived from cross-validation, averages the CV-Accuracy and CV-F1-Accuracy to evaluate the model's generalizability and

reliability on unseen data. Recall measures the model's ability to correctly identify all relevant instances within a dataset. The Harmonized Mean of Precision and Recall provides a balanced metric, particularly valuable in datasets with uneven class distributions.

Random Forest and XGBoost stand out with top scores close to 0.96 in all metrics, demonstrating high precision, accuracy, recall, and F1-score both in standard assessments and cross-validation settings. These models excel in forecasting energy use in IoT devices due to their robustness and consistent performance, suggesting they are well-tuned, dependable, and can generalize well across different datasets. In contrast, SVM shows lower performance with scores between 0.64-0.65, reflecting less precision and balance in predicting power usage. Although stable, SVM struggles with the complexity of the data compared to the more effective models. Logistic Regression exhibits the weakest performance, often near zero, likely because its linear approach fails to address the non-linear complexities present in the data.

These findings emphasize the effectiveness of advanced models like Random Forest and XGBoost in handling complex predictive tasks in IoT settings. Their ability to accurately and reliably predict outcomes underlines their potential in real-world applications, where precise energy management is crucial. Conversely, the limitations of SVM and Logistic Regression highlight the need for models that can accommodate the intricate relationships and variability in real-world data, ensuring more efficient and adaptive energy management solutions.

#### 4.4 Practical Implications for Predicting Energy Usage

Selecting the best models - The heatmap clearly shows Random Forest as the top-performing model. This is crucial for real-world applications where high precision and reliability are

necessary to predict energy usage in IoT devices accurately. By implementing these models, more accurate and dependable predictions can be made.

Understanding model performance - The detailed comparison helps you to understand why some models perform better than others. For example, Random Forest excels due to its capability to handle the non-linear relationships and interactions between features.

Guideline for improvement - Some models, like Logistic Regression and SVM, underperform as shown by the heatmap. Enhancements could include adding additional features, implementing more sophisticated feature engineering techniques, or exploring different machine learning algorithms to better capture the underlying patterns of the data.

Optimizing Energy Use- With precise energy usage forecasts, IoT devices can operate more efficiently. High-performance models like Random Forest or XGBoost enable scheduling operations during peak times, reducing waste, and optimizing overall energy consumption.

Utilizing the insights from the heatmap in real-world applications - For example, applications in smart homes, industrial IoT, and more can leverage the top-performing models from the heatmap for dynamic energy management and optimization. This approach ensures that energy management strategies are not only more effective but also more adaptable to changing conditions and demands. This proactive adaptation is essential for advancing smart energy solutions that contribute to sustainability and efficiency improvements across various sectors.

## 4.5 Key Insights

The Random Forest model outperformed the SVM model and Logistic regression model in terms of their ability to effectively process complex data models and interactions. The models showed high accuracy, accuracy, recall and F1 rating, demonstrating their ability to accurately predict energy consumption. The analysis found that temperature, light intensity and time of day (e.g. days of week and season) were the most significant determinants of energy use. This is an important finding for future research and the development of energy management systems that are more efficient.

By providing accurate energy consumption predictions, these models can help homeowners and energy providers to optimize their energy usage, reduce costs and improve overall energy efficiency. This model can be incorporated into intelligent home systems to enable real-time. In the previous chapter, we looked at how different machine learning models performed in predicting energy use in smart homes powered by IoT technology.

The impressive results of the Random Forest model and the XGBoost model, combined with valuable insights from metric and feature relevance analysis, highlight their potential to enhance energy management practices. In the following chapter, we look at the wider implications of these results and provide recommendations for future research. Ultimately, this chapter provides an overview of how machine learning models perform in predicting energy in smart homes, giving us a good starting point for building robust and sustainable energy management strategies.

## Chapter 5: Conclusion and future works

### 5.1 Conclusion

This study aimed to predict the use of energy in intellectual houses equipped with IoT using advanced automatic learning methods. By analyzing the data of a smart house in Greece, the study gave valuable information on energy consumption models and the factors that affected them. The complete dataset, which was collected at 15 minutes in two years, contained characteristics such as room status, gradation level, brightness, temperature measurement, and total energy consumption. Initial procedures include the details of the data that guarantees that all functions have contributed to the performance of the model, creating a missing value, managing the functional engineering, the time -related functionality, and standardization. Includes treatment. To facilitate the classification task, the target variables were transformed into categorical labels (“low,” “medium,” “high”). In this study, we implemented this four machine learning models: Random Forest, Support Vector Machine (SVM), Logistic Regression, and XGBoost. Each model was trained and evaluated using various performance metrics such as accuracy, precision, recall, and F1 scores. The Random Forest and XGBoost models showed the best performance with an accuracy of 96.48% and 95.67%, respectively. These models have demonstrated reliability and effectiveness in predicting energy consumption based on historical data and external environmental factors such as temperature and brightness.

The random forest model demonstrated high precision, recall, and F1 score across all electricity consumption categories. For example, the model gained 95.71 % accuracy in the Low category, and the Medium category was 95.86 % and the High category in the category of 98.17 %. The test value was 95.82 %, 95.76 %, and 98.17 % in the corresponding category. High F1

indicators reflect the balanced performance of the model, indicating the ability to handle complex models with datasets. Random Forest's confusion matrix (Fig. 5) confirmed its reliability with minimal incorrect classification. Likewise, the XGBoost model performed exceptionally well, with accuracy values of 94.63% for the "low" category, 95.37% for the "medium" category, and 97.26% for the "high". The recall values were 95.95%, 94.27% and 97.15% for the respective categories. The F1 scores were 95.28%, 94.81%, and 97.21%, indicating balanced performance. The XGBoost confusion matrix (Figure 8) showed high accuracy with minimal classification error, demonstrating its effectiveness. In contrast, the SVM and logistic regression models showed moderate validity. The SVM model achieved an accuracy of 85.75%, and the accuracy values for the "low", "medium", and "high" categories were 87.31%, 82.06%, and 89.20%, respectively. The recall values were 85.13%, 86.87%, and 84.99%, resulting in F1 scores of 86.21%, 84.40%, and 87.05%. The confusion matrix for SVM (Figure 6) showed a higher number of classification errors compared to Random Forest and XGBoost. Logistic regression was a simpler linear model and yielded 64.47% accuracy. Its accuracy values were 66.25% for the "low" category, 65.42% for the "medium" category, and 61.52% for the "high" category. Recall rates were 57.92%, 70.10%, and 64.71%, and F1 scores were 61.80%, 67.68%, and 63.07%. The confusion matrix for the logistic regression (Figure 7) revealed significant misclassification, especially in the "low" and "high" categories, reflecting limitations in handling complex interactions. Feature importance analysis showed that temperature, brightness and time characteristics (hour, day of week, month) were the most significant predictors of energy consumption. This understanding is critical to developing effective energy management strategies. By understanding which features have the greatest impact, future models can be improved to increase accuracy and efficiency.

The results of this study have important implications for optimizing the energy efficiency of

smart homes. Accurate energy use forecasts enable homeowners and energy suppliers to implement dynamic energy management strategies. These strategies can streamline operations, adapt to changing environmental conditions, and lead to significant cost savings and energy efficiency improvements. This reduces unnecessary energy consumption and promotes more efficient use of resources, thus promoting environmental sustainability.

## 5.2 Future Work

This research outlines several promising directions for future work.

**Hybrid Modeling Approaches:** Future studies could explore the creation of hybrid machine learning models that merge the strengths of various algorithms. Such models might enhance prediction accuracy by leveraging the unique benefits of multiple approaches, potentially processing complex energy data more effectively. For instance, integrating the robustness of Random Forest with the precision of XGBoost could yield superior outcomes.

**Real-time Prediction Systems:** Another avenue is the development and implementation of real-time prediction systems for energy consumption in IoT devices. These systems would allow for dynamic adjustments based on predictive data, offering immediate feedback and controls. This could enhance energy efficiency and provide real-time management solutions, enabling rapid responses to anomalies and ongoing optimization of energy usage.

**Energy Saving Strategies:** Employing predictive models to develop and test different energy-saving strategies is crucial. Evaluating these strategies in real-world settings could help identify the most effective approaches to enhance IoT device performance and user satisfaction. Possible strategies might include scheduling high-energy tasks during off-peak hours or

implementing preventive maintenance to ensure efficient device operation.

There is also scope for advancing feature engineering techniques to unearth more insightful features from the data. Exploring sophisticated methods like polynomial features, interaction terms, and domain-specific transformations could help reveal complex patterns and boost model efficacy. By expanding the dataset to include more environmental variables, such as humidity, wind speed, and solar radiation, future research could gain a fuller understanding of their effects on energy consumption. This inclusion could lead to more precise forecasts and more effective energy management strategies. Lastly, enhancing the transparency and understanding of machine learning models through explainable AI techniques is vital for their practical application. Future research should focus on methods that clarify the model's decision-making processes, fostering trust and ensuring that models are applied effectively in real scenarios.

In conclusion, this study underscores the potential of machine learning models to forecast energy consumption in IoT-equipped smart homes effectively. The results provide a foundation for continued innovation in energy management technologies, with advanced machine learning methods delivering accurate predictions that drive more efficient and sustainable energy use in IoT contexts.

## References

Here is the list of references arranged in alphabetical order and formatted in Harvard style:

- [1] Albeladi, K., Zafar, B. and Mueen, A. (2023) 'Time Series Forecasting using LSTM and ARIMA', *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(1).
- [2] An IoT-Based Prediction Technique for Efficient Energy Consumption in Buildings (2021). [Online]. Available at: <https://ieeexplore.ieee.org/abstract/document/9462477>.
- [3] Aurna, N.F., Anika, F.S., Rubel, M.T.M., Kabir, K.H. and Kaiser, M.S. (2021) 'Predicting periodic energy saving pattern of continuous IoT based transmission data using machine learning model', in *Proc. 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Feb. 2021, pp. 428-433.
- [4] Balaji, S. and Karthik, S. (2023) 'Energy prediction in IoT systems using machine learning models', *Comput, Mater Contin*, vol. 75, no. 1.
- [5] Cviti, I. et al. (2021) 'Ensemble machine learning approach for classification of IOT devices in Smart Home', *International Journal of Machine Learning and Cybernetics*, SpringerLink. [Online]. Available at: <https://link.springer.com/article/10.1007/s13042-020-01241-0>.
- [6] Dibal, P.Y. et al. (2022) 'Processor power and energy consumption estimation techniques in IoT applications: A Review', *Internet of Things*. [Online]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S2542660522001366>.
- [7] Energy Prediction in IoT Systems Using Machine Learning Models (2023). [Online]. Available at: [https://file.techscience.com/files/cmc/2023/TSP\\_CMC-75-1/TSP\\_CMC\\_35275/TSP\\_CMC\\_35275.pdf](https://file.techscience.com/files/cmc/2023/TSP_CMC-75-1/TSP_CMC_35275/TSP_CMC_35275.pdf).
- [8] Hafezi Fard, R. et al. (2021) 'Machine learning algorithms for prediction of energy consumption and IoT modeling in complex networks', *Microprocessors and Microsystems*. [Online]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0141933121005640>.
- [9] IEEE (2021) 'A comparison of Arima and LSTM in forecasting Time Series'. [Online]. Available at: <https://ieeexplore.ieee.org/document/8614252/>.
- [10] IEEE (2021) 'Predicting Periodic Energy Saving Pattern of Continuous IoT Based Transmission Data Using Machine Learning Model'. [Online]. Available at: <https://ieeexplore.ieee.org/abstract/document/9396928>.
- [11] Impact of internet of things paradigm towards energy consumption prediction: A systematic literature review (2023). [Online]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S221067072100888X>.

- [12] Interplex (2021) 'Understanding Global Internet Energy Usage & Trends'. [Online]. Available at: <https://interplex.com/resources/understanding-global-internet-energy-usage-and-trends/>.
- [13] Malki, A. et al. (2022) 'Machine learning approach of detecting anomalies and forecasting time-series of IoT devices', Alexandria Engineering Journal. [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S110016822001260>.
- [14] Nakamura, S. et al. (2023) 'Assessment of energy consumption for information flow control protocols in IoT devices', Internet of Things. [Online]. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S2542660523003153>.
- [15] P. D. Rosero-Montalvo, P. Tözün, and W. Hernandez (2023) 'Time Series Forecasting to Fill Missing Data in IoT Sensor Data', IEEE Sensors Letters.
- [16] Recent, A. (2023) 'Prediction of energy consumption using ANN and years, Digital Twin Technology for Thermal Comfort and energy efficiency in buildings', Energy and Built Environment. [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S2666123323000314>.
- [17] S. Balaji and S. Karthik (2023) 'Energy prediction in IoT systems using machine learning models', Comput, Mater Contin, vol. 75, no. 1.
- [18] Shapi, M.K.M., Ramli, N.A. and Awal, L.J. (2021) 'Energy consumption prediction by using machine learning for smart building: Case study in Malaysia', Developments in the Built Environment, vol. 5, p. 100037.
- [19] Time-series forecasting to fill missing data in IOT sensor data (2023), IEEE Journals & Magazine. [Online]. Available at: <https://ieeexplore.ieee.org/document/10225693>.
- [20] Zenodo (2023). Available at: <https://zenodo.org/records/7628298>.