International Journal of Energy Efficiency Engineering (IJEEE)
Volume 1, Issue 1 (May 2025), Pages 5-26

INTERNATIONAL JOURNAL OF
ENERGY EFFICIENCY ENGINEERING

Predicting Energy Consumption in Buildings Using Various Artificial Intelligence Models

Yazeed E. AbuShanab, Mohamed I Youssef, Omar Shaker, M Abdelaziz Youssef, Ryoichi S Amano[1]

Department of Mechanical Engineering, University of Wisconsin-Milwaukee 115 E. Reindl Way, Glendale, WI 53212

Abstract

Accurately predicting energy consumption in buildings is vital for optimizing energy efficiency, reducing costs, and supporting sustainability efforts. This study uses a dataset that spans and broadcasts hourly energy consumption for a specific building in Spain, using a dataset spanning an entire year. The dataset includes hourly energy usage in kilowatt-hours (kWh) and features representing environmental conditions, including temperature, humidity, and precipitation. alongside time-related variables, including the hour of the day, day of the week, and seasonal markers. These features provide a detailed view of how internal and external conditions influence energy usage patterns. Data preprocessing included handling missing values, feature selection, and engineering temporal variables such as Hour, Day of Year, and Is Weekend, which capture essential behavioral and operational dynamics. The building analyzed is a representative structure with typical heating, ventilation, and air conditioning (HVAC) systems. This model is well-suited for analyzing energy consumption patterns across different environmental and operational conditions. Various regression models were applied, including Linear Regression, Ridge and Lasso Regression, Support Vector Regression (SVR), K-

[1] Corresponding Author: Professor Ryoichi Amano, amano@uwm.edu

Nearest Neighbors (KNN), Random Forest, XGBoost, and Neural Networks. Model performance was assessed using Mean Absolute Error (MAE) and R-squared (R^2) metrics. Random Forest emerged as the best-performing model, achieving an MAE of 8.33 and an R^2 of 0.954, highlighting its strong ability to capture the building's energy consumption patterns. This research highlights the potential of regression models and artificial intelligence in improving energy forecasting, serving as a foundation for advancing building energy management systems.

Keywords: *Energy Consumption in Buildings; Energy Forecasting; Artificial Intelligence; Annual Energy Savings; Linear Regression; Ridge and Lasso Regression; Support Vector Regression (SVR); K-Nearest Neighbors (KNN); Random Forest, XGBoost, and Neural Networks.*

I. INTRODUCTION

Nowadays, it is impossible to overlook the changes and oscillations in the energy industry. Considering the duty of maintaining and optimizing the current sources and emphasizing the significance of energy management throughout the entire system life cycle with all of their methodologies and techniques, it also became clear that there is a significant demand for dependable yet sustainable energy resources. Closing the gap and developing reliable substitutes for such systems are necessary to address this weakness. As mentioned earlier, resource allocation, expenditure optimization, and comprehensive but meticulous design objectively relate to and significantly impact the need. Key elements of any organization's success and operations are the efficiency of resource management, energy optimization, and continuous process improvement procedures, which start with identifying, evaluating, and responding to steps for improving the current procedure or process and conclude with the steps of documenting and recording them for future reference [1] [2].

Driven by increasing energy demands, the need to minimize greenhouse gas emissions, and governmental policies encouraging energy efficiency, smart energy systems have emerged as an effective solution for optimizing energy consumption and promoting sustainability. These

systems offer the potential to mitigate the environmental impact of rising energy use due to urbanization, industrialization, and population growth by reducing pollution and greenhouse gases while conserving resources through intelligent analysis of energy consumption patterns and identification of opportunities for waste reduction [3].

As energy efficiency becomes increasingly important in mitigating climate change and reducing operational costs, developing predictive tools to optimize energy consumption aligns with personal and professional goals. This research study uses machine learning techniques to predict hourly energy consumption in buildings. By analyzing weather and temporal features, such as temperature, humidity, and time of day, the aim is to develop accurate predictive models for optimizing energy management. This research paper compares the performance of various models, including Random Forest, Ridge Regression, and Neural Networks, to identify the most effective approach for this task. In addition to its relevance to sustainability and energy management, it also serves as a valuable application of machine learning techniques in such fields, bridging theory with a real-world problem.

Several models were evaluated to predict energy consumption, each chosen for its unique strengths in handling different aspects of the data. Random Forest (RF) was evaluated due to its ability to handle non-linear relationships and complex data interactions, which are common in energy consumption patterns. As an ensemble method, RF also reduces the risk of overfitting and is robust in dealing with missing data. Linear models, while simpler and more interpretable, were evaluated to provide a baseline for comparison, although they failed to capture the non-linear complexities of the energy consumption data effectively. Support Vector Regression (SVR) was included for its ability to handle high-dimensional spaces and its success in other regression tasks, but it struggled with the non-linearities in the dataset, leading to poorer performance than RF. Other models were considered to benchmark the performance of more sophisticated algorithms. Furthermore, Neural Networks were also evaluated, as they are known for their powerful capability to model complex, non-linear relationships. However, Neural Networks might not be required for the model because the model might not be as complex to handle using Neural Networks.

The main issues in energy consumption analysis are examined in this study, emphasizing anomaly

detection, model performance, and feature relevance. It seeks to pinpoint the most significant variables that affect energy use, such as weather and seasonal trends. It also compares several machine learning models to see which provides the optimum accuracy and interpretability balance. Identifying anomalies or outliers in odd usage patterns is another crucial component of improving energy management. By tackling these issues, the research paper aims to show how machine learning can be used in the real world and offer practical insights for enhancing building operations' energy efficiency.

II. LITERATURE REVIEW

In intelligent energy systems, the quick development of artificial intelligence (AI) algorithms has created new chances to optimize energy use and encourage sustainable habits. To support load balancing, demand side management, and power grid stability optimization calculations, artificial intelligence algorithms can examine user behavior and energy usage trends. In the end, these algorithms can offer energy-saving measures. To solve the increasing issues of energy consumption, environmental sustainability, and resource efficiency, energy efficiency is essential.

The primary goal of univariate regression is to examine the relationship between a dependent variable and a single independent variable, establishing a linear equation that represents the connection between the two models; multilinear Regression is the term for regression models that have one dependent variable and multiple independent variables. Regression analysis is a statistical technique for predicting the relationship between variables with reason and result relationships [4].

Additionally, this can be used as a basis for future studies on the estimation and forecasting of energy use in the food and beverage business, with the results and conclusions being used in a multi-criteria decision-making process, such as the PROMETHEE model, to forecast the behavior of that industry [5]. To classify data (classification) and forecast continuous values (Regression), the Random Forest regression technique first creates several datasets by resampling the original data. As shown in [6], this sampling strategy is commonly used for both small and big population selection. A decision tree is built for these resampled datasets without pruning. One important characteristic of random forests is that, instead of assessing every predictor, a random subset of

the available predictor variables is considered for the optimum split at each decision-making step. After this set of trees is created, the forecasts from each tree are combined to create predictions on fresh data. For regression tasks, this aggregation entails averaging the predictions, and for classification problems, a majority vote. Standard bagging can be seen as a special case of Random Forests where all predictors are considered at each split rather than a random subset [7].

Regression techniques like Lasso and Ridge work well for managing multicollinearity and avoiding overfitting. Multicollinearity and automated model selection benefit from Lasso (L1 regularization), which encourages simpler models with fewer parameters. Ridge regression (L2 regularization), especially well-suited for multicollinearity, lowers estimate bias and variability. The nature of the issue and the intended model attributes determine which of the Lasso and Ridge regression is best. Ridge regression works well when every feature affects performance, but Lasso is best suited for many characteristics with few significant ones. These methods help improve model performance and interpretability, which helps researchers create dependable regression models for various uses. When trying to figure out how best to use each element in a statistical model to predict or comprehend the response variable, multicollinearity can provide skewed or misleading findings [8]. Additionally, it may lead to less dependable results and broader confidence ranges. The particulars of the problem and the intended model attributes determine which of Lasso and Ridge regression is best. Lasso works well when there are many characteristics, but only a few are significant. Ridge regression, on the other hand, works well when multicollinearity is present and every feature affects the model's performance. Both strategies provide strong frameworks for improving models [9].

Support Vector Machines (SVM) is a learning method implemented using SVM, which helps identify minute patterns in large, complicated data sets. The system uses discriminative categorization learning as an example to forecast the classes of previously unknown data. It's based on learning machines that apply the inductive concept of structural risk reduction to achieve good generalization on a small set of learning patterns. Support Vector Regression (SVR) aims to achieve generalized performance by reducing the generalization error bound rather than the observed training error [10]. The concept behind Support Vector Regression (SVR) is to compute a linear regression function in a high-dimensional feature space, where a nonlinear function maps the input data into this space. The approximation of difficult engineering studies, convex

quadratic programming, loss function selection, time series and financial (noisy and risky) prediction, and other areas have all used SVR. In this study, an attempt has been made to discuss the current theory, procedures, new advancements, and applications of SVR [11].

Unlabeled observations are categorized using the K-nearest Neighbors (kNN) classifier by placing them in the same class as the most comparable labeled samples. Both the training and test datasets observational characteristics are gathered [12].
The kNN algorithm's diagnostic performance is greatly impacted by the choice of k. Although a large k lessens the effect of variation by random error, it also increases the possibility of overlooking a subtle but significant pattern. Finding a balance between overfitting and underfitting is crucial when selecting a k value [13].

Neural networks are inspired by the brain's structure for information processing. While they do not perfectly replicate the brain's functions, they are scientifically motivated models. Due to their ability to learn from data, neural networks have proven to be highly effective in various forecasting and business classification applications [14]. To effectively complete a job, the artificial neural network learns by adjusting the weights in the network design. It may automatically learn from examples or input-output relationships, or it can learn from existing training patterns. Despite their continued remarkable performance on well-known machine learning challenges, it has been challenging to prove that neural network models can reason about ideas [15] [16]. Artificial intelligence (AI) techniques are applied in many study domains, including material strength and properties, healthcare, risk assessment and prediction, soil mechanics and characteristics, and building energy consumption [17].

III. METHODOLOGY

A. Study Limitations

This study focuses on a residential facility with a moderate electricity consumption load compared to industrial and commercial buildings. One limitation is its geographical scope, as the research was conducted on a building in Spain, which may affect the generalizability of the findings to other regions with different climatic, economic, or regulatory conditions. Additionally, while expanding the study to cover more years and neighboring buildings could

provide a broader perspective on energy consumption and optimization, data availability poses a significant challenge. Access to long-term and multi-location energy data is often restricted due to privacy concerns, measurement inconsistencies, or institutional constraints. Future research could benefit from improved data-sharing policies and collaborations to enhance the applicability of the findings across diverse locations.

Moreover, while the study evaluated several machine learning models, including Random Forest, linear models, SVR, and Neural Networks, it did not explore time-series models such as Long Short-Term Memory (LSTM) networks or ARIMA. These models, which are specifically designed to capture sequential dependencies in time-dependent data, could potentially improve performance by better modeling the temporal patterns inherent in energy consumption. Future research could explore the application of these models to address this limitation and enhance predictive accuracy.

B. Workflow

Figure 1 shown below illustrates the step-by-step process of predicting building energy consumption using regression models. It begins with data collection, where the hourly energy consumption and environmental conditions (temperature, humidity, precipitation) are gathered for a time span of a year. The next step, data preprocessing, addresses missing values and prepares the data for analysis. Feature engineering follows, creating essential temporal variables like Hour, Day of Year, and Is Weekend to capture energy usage patterns effectively. After the data is processed, various regression models are selected for comparison, including Linear Regression, Ridge and Lasso Regression, SVR, KNN, Random Forest, XGBoost, and Neural Networks. These models are then trained in the model training step using the preprocessed data. Performance is assessed to determine the best model based on evaluation metrics such as MAE and R^2. The final results interpretation and insights stage derives conclusions that help in optimizing building energy consumption.

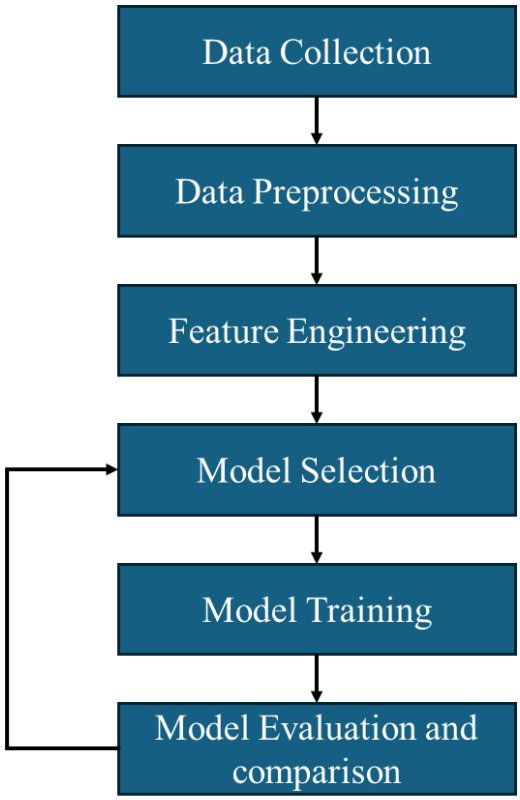


Figure 1: Flowchart for finding the best AI model.

Finally, the results are thoroughly discussed and interpreted to identify the best-performing model based on various evaluation metrics. This involves comparing model performance across different criteria, such as accuracy, precision, recall, and overall prediction reliability. After selecting the best model, feature interpretation is introduced as a crucial step to ensure the model's alignment with reality and its ability to provide meaningful insights. Feature importance techniques, such as SHAP is utilized to assess the contribution of individual features to the model's predictions. This step is critical for validating the model’s decisions and ensuring that the selected features make logical sense in the context of the problem. Additionally, feature interpretation helps in understanding whether the model is capturing the correct relationships between input variables and the target variable, ensuring consistency with domain knowledge and real-world expectations.

C. Dataset

Using hourly energy usage records for two buildings over the span of 3 years, the "Building Energy Consumption Dataset" from Mendeley Data is used in this study [18]. Key features of the dataset include the following: weather variables (minimum, maximum, and average temperature), precipitation (PRECTOT), humidity (RH2M), and energy consumption (the target variable) measured in kWh. Temporal variables include the hour, day of the week, weekend indicator, and day of the year. This dataset, which is openly accessible, provides enough variation to facilitate the creation and assessment of machine learning models.

This study investigates many regression models to forecast energy usage, each with its advantages. Multiple linear regression is a simple method for modeling linear connections. An ensemble technique called Random Forest Regression reduces overfitting and improves generalization by mixing many decision trees to increase accuracy and stability. Ridge and Lasso Regression maintains predictive strength while avoiding overfitting by introducing regularization. Support Vector Regression (SVR) uses support vector machines to deal with nonlinear connections. Using the average of the nearest data points in feature space, K-Nearest Neighbors (KNN) makes value predictions. Furthermore, Neural Networks offer a more sophisticated and adaptable method for identifying intricate patterns in energy usage data.

Additionally, the performance of each model has been assessed using Mean Absolute Error (MAE) and R-squared (R^2) metrics, standard measures for regression performance. The process details have been assessed through data preprocessing, model building, evaluation, and comparison of different machine learning techniques for predicting building energy consumption. The models' effectiveness is measured using predictive accuracy and interpretability to identify the best approach for energy consumption forecasting. The dataset used in this research study consists of hourly energy consumption data from a building, providing insights into how energy usage fluctuates across time in response to various internal and external factors. The target variable, Energy Consumption (ENERGY), represents the energy consumption in kilowatt-hours (kWh) at each hour. Several features related to environmental conditions, including temperature, humidity, and precipitation, are also provided. Specifically, the dataset includes columns like T2M, T2M_MIN, and T2M_MAX, representing

the average, minimum, and maximum temperatures recorded during each hour. These temperature features are crucial as they could impact the building's heating and cooling needs. RH2M represents the relative humidity, which can also influence energy usage, particularly concerning HVAC (heating, ventilation, and air conditioning) operations.

D. Data Preprocessing

In addition to environmental variables, the model includes temporal features such as Hour, DayOfYear, and IsWeekend to capture the effect of time on energy consumption patterns. The Hour feature provides information on the specific time of day, which is essential in understanding daily usage patterns, while DayOfYear tracks seasonal variations. IsWeekend serves as a binary indicator, reflecting whether the observation falls on a weekend, which may have a different energy consumption pattern due to reduced occupancy or different operational schedules in the building. This dataset provides a comprehensive foundation for predicting energy consumption, combining environmental, temporal, and building-specific factors.

The first step in preparing the machine learning data was handling missing values. The dataset was checked for missing or null entries, which are common in real-world data and can cause issues during model training. Any rows containing missing values were removed to ensure that only complete records were used. This decision was made based on the fact that missing data could introduce bias or reduce the quality of predictions. After removing rows with missing values, the dataset became more consistent and ready for further processing. A feature selection was conducted to identify the most relevant features for predicting energy consumption. Certain features, such as PRECTOT (total precipitation) and RH2M (relative humidity), ALLSKY (radiation) were dropped from the dataset after performing feature importance analysis as shown in Figure 2 (this is just a sample from the random forest method, but all other feature importance show similar results).

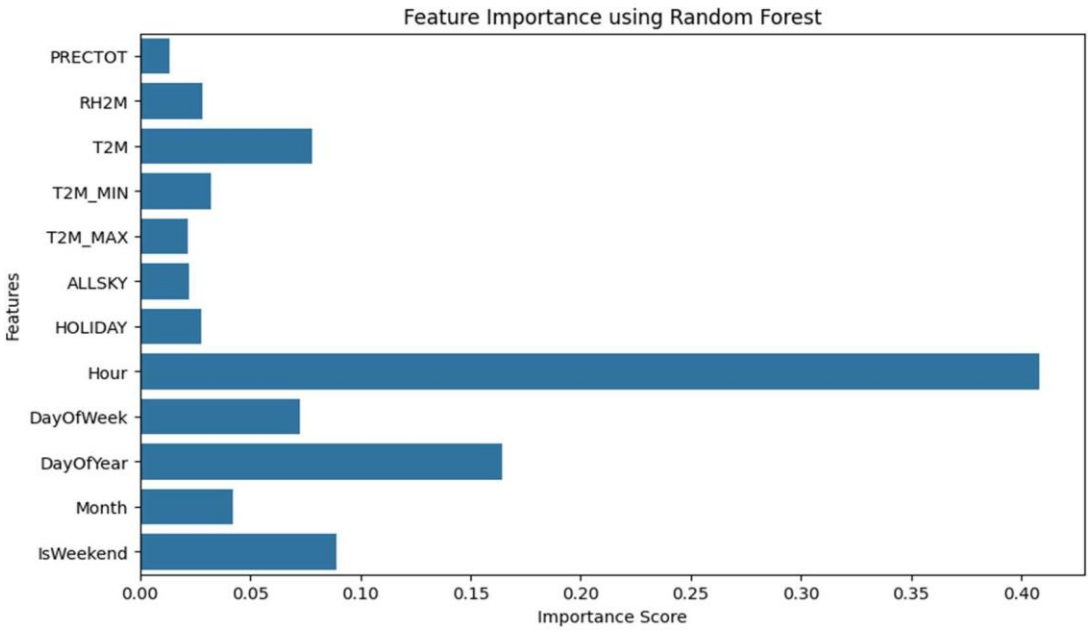


Figure 2 Feature Importance results using Random Forest

This step was crucial, as it reduced dimensionality and removed variables that did not significantly contribute to the prediction model. By eliminating unnecessary features, the model could focus on the variables with the highest predictive power, such as temperature and temporal features. Afterward, feature engineering was utilized, where the DATE column was transformed into useful time-based features. The original DATE column, which included both the date and hour, was split to extract the hour of the day, the day of the year, and a binary feature indicating whether the day was a weekend. These temporal features are important because they help capture daily, weekly, and seasonal patterns in energy consumption, which are crucial for making accurate predictions. Including a weekend indicator, for example, helps the model account for potential differences in energy consumption between weekdays and weekends, which can be influenced by factors such as occupancy and operational schedules.

To ensure that all features were comparable, standard scaling was applied for all methods except the Min-Max Scaler for the Neural Network. Scaling is particularly important for machine learning algorithms sensitive to the magnitude of input features, such as Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). The Min-Max Scaler transforms the data so that all features are scaled to a range between 0 and 1, making the models less biased toward features with larger numeric ranges. This normalization process ensures that the model treats

each feature equally, preventing one feature from dominating the learning process. In addition, a train-test split has been implemented to evaluate the model's performance. The data was divided into training and testing sets, with 80% of the data used for training and the remaining 20% reserved for testing. This approach allows for proper validation and helps prevent overfitting, ensuring the model can generalize well to unseen data.

IV. RESULTS AND DISCUSSION

This study evaluates several machine learning models based on their performance in predicting energy consumption, measured by Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Table 1 summarizes the performance of each method used in this project.

Table 1 Results of the various regression models.

MODEL	MEAN ABSOLUTE ERROR (MAE)	MEAN SQUARED ERROR (MSE)	R^2
LINEAR REGRESSION	50.08	4216.94	0.197
RANDOM FOREST	8.33	242.08	0.954
XGBOOST	17.14	700.26	0.867
RIDGE REGRESSION	49.31	4101.75	0.219
LASSO REGRESSION	49.97	4183.10	0.203
SVR	35.18	2882.00	0.451
KNN	18.38	910.94	0.826
NEURAL NETWORK	**18.64**	**806.03**	**0.85**

The regression model results reveal significant differences in predictive performance for energy consumption. Among the evaluated models, Random Forest emerges as the best performer, achieving the lowest Mean Absolute Error (MAE) of 8.33, a Mean Squared Error (MSE) of 242.08, and the highest R-squared (R^2) of 0.954. This indicates that the model explains approximately 95.4% of the variance in the data, as shown in Figure 3.

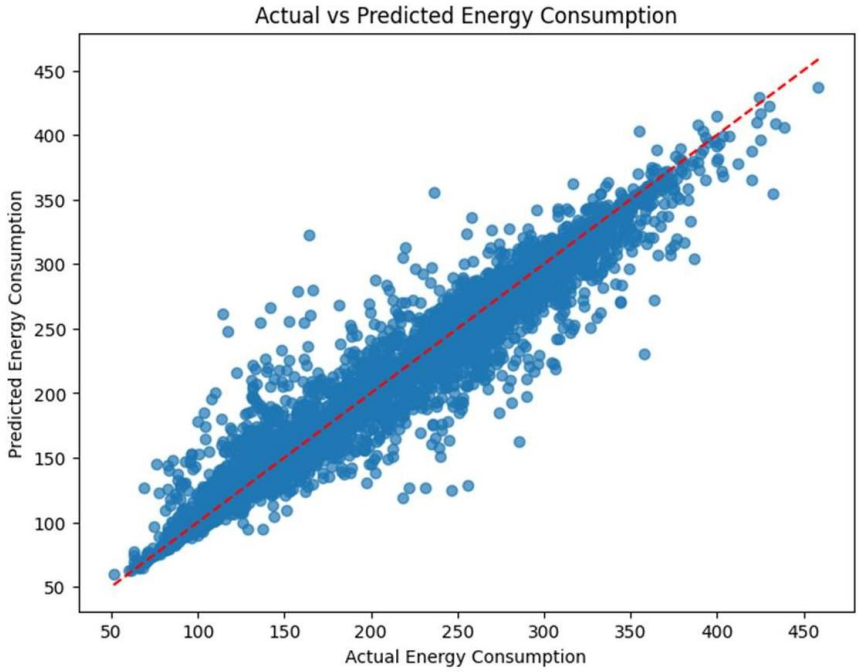


Figure 3 Predicted Energy consumption vs actual

XGBoost also demonstrated strong predictive accuracy, achieving an MAE of 17.14, MSE of 700.26, and an R^2 of 0.867, though it was slightly less effective than Random Forest. K-Nearest Neighbors (KNN) followed with an MAE of 18.38, MSE of 910.94, and an R^2 of 0.826, indicating good performance but higher error rates compared to tree-based models. In contrast, Support Vector Regression (SVR) had a significantly higher MAE of 35.18, MSE of 2882.00, and a relatively low R^2 of 0.451, suggesting it was less effective at capturing underlying patterns in the data. The Neural Network model trained smoothly without overfitting, as shown in Figure 4. It produced decent results with an MAE of 18.64, MSE of 806.03, and an R^2 of 0.85, outperforming SVR but falling behind Random Forest and XGBoost in predictive accuracy.

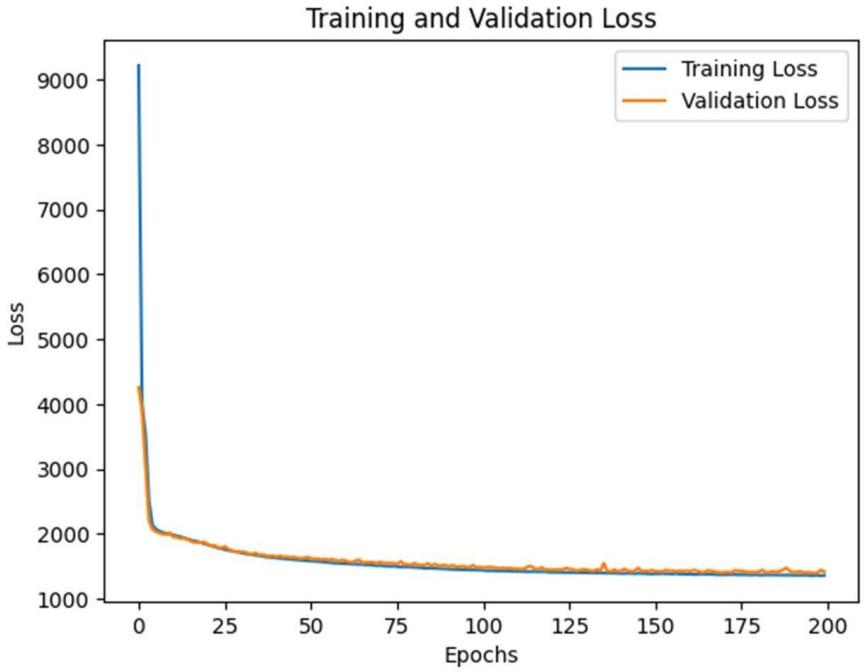


Figure 4 Training process of the neural network

However, further tuning and testing can be done on the Neural Network. As discussed previously a feed forward Neural Network consists of and input and out layer and hidden layers in between, where each layer consists of certain number of neurons. The learning algorithm used was Adaptive Moment Estimation optimizer (Adam) is an optimization algorithm that combines the advantages of momentum and adaptive learning rates. It computes individual learning rates for each parameter using estimates of the first and second moments of the gradients, enabling efficient and robust training of deep learning models. However, the learning rate can be changed and tuned. Different structures of number of layers, neurons and learning rates were tested and evaluated. The results of this tuning can be shown in Table 1.

Table 1 performance results for different Neural Networks

Layers-neurons	**LEARNING RATE**	**MEAN ABSOLUTE ERROR (MAE)**	**MEAN SQUARED ERROR (MSE)**	**R^2**
64	0.0010	27.860467	1739.491082	0.668633
128-64	0.0010	24.186951	1404.341130	0.732478
128-64-32	0.0010	21.331557	1061.582213	0.797773

64	0.0100	27.398675	1590.457467	0.697024
128-64	0.0100	19.685038	903.640517	0.827860
128-64-32	**0.0100**	**18.639559**	**806.029579**	**0.846454**
64	0.0001	35.197151	2326.152578	0.556877
128-64	0.0001	26.667670	1688.658357	0.678317
128-64-32	0.0001	25.715233	1539.650502	0.706702

It can be noted that the maximum achieved R^2 value achieved was approximately 0.84 which enhanced the original Neural Networks performance but still slightly under performing the random forest method.

Ridge and Lasso Regression exhibited poor performance, with high MAE values (49.31 and 49.97, respectively) and low R^2 scores (0.219 for Ridge and 0.203 for Lasso), likely due to their inability to capture the complexities in the data. Similarly, Linear Regression performed poorly, with an MAE of 50.08, MSE of 4216.94, and an R^2 of 0.197, suggesting that the linear assumptions do not hold for this problem. Figure 5 illustrates the performance of the Ridge regression method.

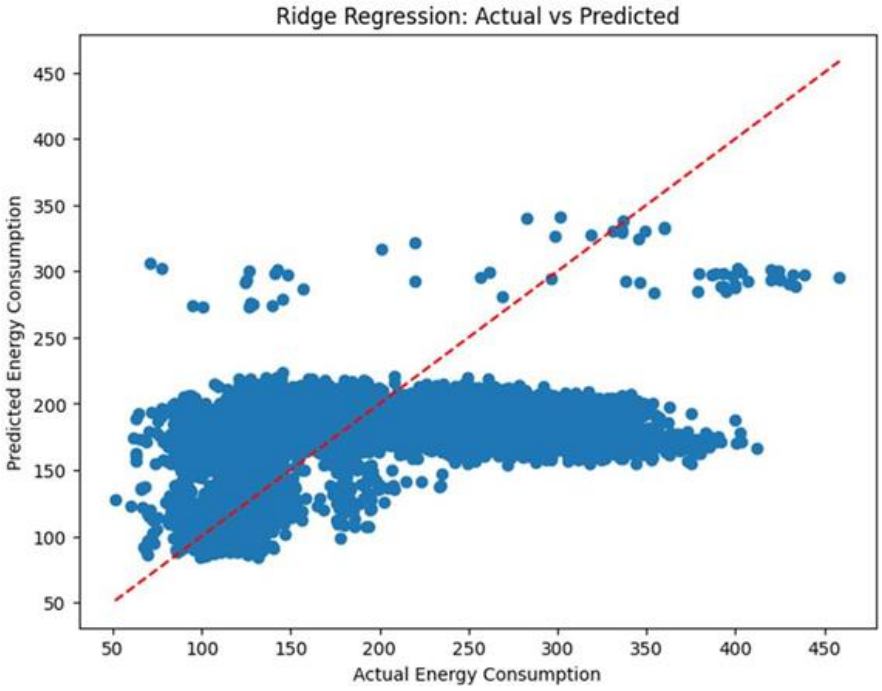


Figure 5 Ridge regression performance

The results suggest that tree-based models, particularly Random Forest, are the most suitable for this energy consumption prediction task. In contrast, linear models and Support Vector Regression (SVR) struggle to capture the data's complexities effectively. For building managers, these findings highlight the potential for implementing more advanced machine learning models, like Random Forest, to improve the accuracy of energy consumption forecasts. Furthermore, feature interpretability is done using SHapley Additive exPlanations (SHAP) [19]. SHAP is a machine learning interpretability method based on game theory, designed to explain individual predictions of machine learning models. It provides a unified approach to understanding model outputs by calculating the contribution of each feature to a particular prediction. The SHAP summary plot (Figure 6) shows how different features impact energy consumption predictions. "Hour" influences energy use significantly, with higher values in the evening leading to increased consumption, while early hours reduce it. "IsWeekend" indicates higher energy use on weekends compared to weekdays. "T2M" (temperature) has a mixed effect, with higher temperatures slightly increasing energy consumption, likely due to cooling needs. "DayOfYear" shows minimal impact on energy use, with no clear trend. Overall, the plot reveals the key factors that drive energy consumption based on time of day, weekends, temperature, and seasonal variations.

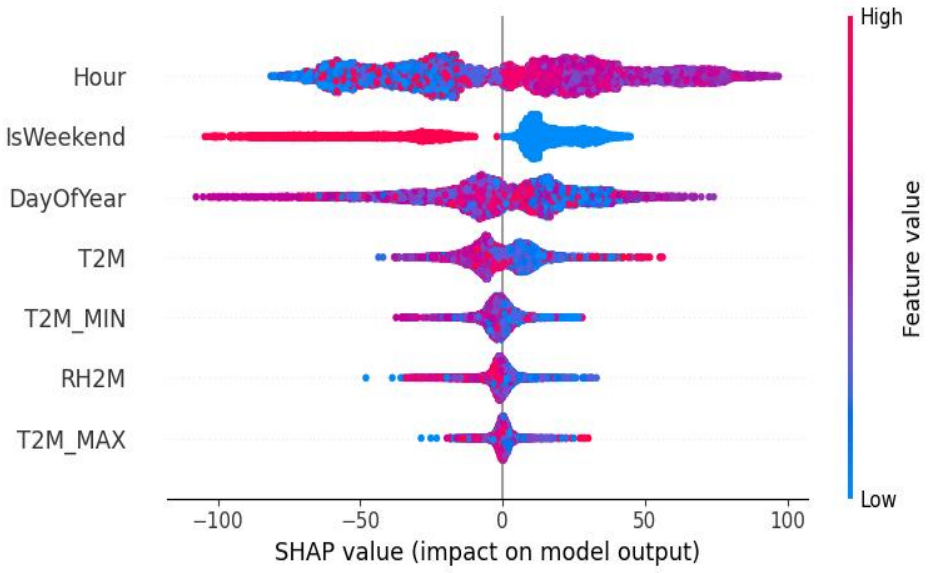


Figure 6 SHAP summary plot

Such predictions can inform better decision-making in energy optimization, leading to cost savings and more efficient resource management. Building managers can leverage these insights to implement strategies that reduce energy consumption, optimize heating and cooling schedules, and ultimately lower operational costs. Future research could explore how these models could be integrated into real-time energy management systems.

V. CONCLUSIONS

This study aimed to predict building energy consumption using various machine learning regression models. The dataset included hourly energy consumption data along with weather and temporal features. After preprocessing—handling missing values, feature selection, and scaling—several regression models were applied, including Linear Regression, Random Forest, Ridge and Lasso Regression, Support Vector Regression (SVR), K-Nearest Neighbors (KNN) and feedforward Artificial Neural Networks (ANN). Model performance was evaluated using Mean Absolute Error (MAE) and R-squared (R^2) to assess accuracy and generalization. Among the models, Random Forest Regression achieved the best performance, with an MAE of 8.33, MSE of 242.08, and an R^2 of 0.954, indicating its strong predictive power. XGBoost followed closely, with an MAE of 17.14 and an R^2 of 0.867. Ridge Regression also showed unacceptable accuracy (MAE: 49.31, R^2: 0.219), while simpler models like Linear Regression (MAE: 50.08, R^2: 0.197) struggled with the dataset's complexity. SVR performed poorly as well, with an MAE of 35.18 and an R^2 of 0.451, demonstrating its limitations in capturing energy consumption patterns. Furthermore, ANN performed really well where originally they performed slightly worse than KNN however after tuning and optimizing hyper parameters the neural network was able to outperform KNN but still slightly underperforms XGBoost and random forest. Overall, the results highlight the effectiveness of ensemble methods and regularized models in handling complex relationships, while traditional linear approaches were less effective for this task.

The findings of this study have significant implications for the field of energy consumption prediction. By comparing different machine learning models, the study highlights the importance of using ensemble methods and regularized models, which are more adept at handling complex, non-linear relationships in energy data. The strong performance of Random

Forest and XGBoost suggests that these models can be valuable tools for energy forecasting, potentially assisting building managers and policymakers in making more accurate energy usage predictions and decisions. Additionally, the study provides insights into model selection, guiding future work in energy efficiency, optimization, and demand forecasting, ultimately contributing to more sustainable energy practices and better resource management.

VI. EXTENSIVE STUDIES

Although this study achieved promising results, several avenues for improvement and further exploration exist. First, including more granular or diverse data, such as building-specific features (e.g., floor area, insulation type, occupancy schedules) and additional weather parameters (e.g., wind speed, cloud cover), could enhance model performance. Moreover, time-series forecasting techniques, such as ARIMA or LSTM (Long Short-Term Memory) networks, could be explored to better capture the temporal dynamics and sequential dependencies in energy consumption.

Additionally, feature engineering can be expanded by exploring advanced techniques such as creating interaction terms between weather and temporal features or incorporating external factors like holidays and special events. Hyperparameter optimization and cross-validation can also be further refined to improve the robustness of the models.

Acknowledgments

The US Department of Energy funds this study under DE-EE0009728.

References

1. R. Amano, M. I. Youssef and M. A. Youssef, "Exploring the Correlation Between Energy Intensity and Specific Energy Consumption in Food And Kindred Industry for the Midwest States," International Journal of Energy for a Clean Environment, Vol. 26, No. 4, 2025.

2. R. Amano, M. A. Youssef, M. I. Youssef, "Employing Linear Regression Analysis: Investigating The Relationship Between Energy Intensity and Specific Consumption in U.S. Midwest Plastic Production Facilities," International Journal of Energy for a Clean

Environment, 2025.

3. J. Dong, J. Gao, J. Yu, L. Kong, N. Jiang and Q. Wu, "Leveraging AI Algorithms for Energy Efficiency: A Smart Energy SystemPerspective," Advances in Artificial Intelligence, Big Data and Algorithms, pp. 57-64, 2023.
4. G. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," Procedia - Social and Behavioral Sciences, vol. 106, p. 234–240, 2013.
5. M. I. Youssef and B. Webster, "A multi-criteria decision making approach to the new product development process in industry," in Reports in Mechanical Engineering, 2022.
6. M. I. Youssef and Y. M. Hausawi, "Utilizing the enterprise architecture model to develop the structure of public sector entities in Saudi Arabia,," Journal of Engineering Management and Systems Engineering, vol. 3, no. 3, pp. 164-174, 2024.
7. S. Cutler, D. Cutler and J. Stevens, "Random Forests," in Ensemble Machine Learning: Methods and Applications, Springer, 2011, pp. 157-176.S.
8. D. Kobak, J. Lomond and B. Sanchez, "The optimal ridge penalty for real dimensional data can be zero or negative due to the implicit ridge regularization," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 6863-6878, 2020.
9. S. K. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman and D. Awwad, "Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19," Migration Letters, vol. 20, no. 6, pp. 139-153, 2023.
10. K. Pelckmans, I. Goethals, J. D. Brabanter and B. De Moor, "Componentwise Least Squares Support Vector Machines," in Support Vector Machines: Theory and Applications, 2005.
11. D. Basak, S. Pal and D. Chandr, "Support Vector Regression," Statistics and Computing, vol. 11, no. 10, 2007.
12. Z. Zhang, "Introduction to machine learning: K-nearest neighbors," Annals of Translational Medicine, vol. 4, no. 11, pp. 218-218, 2016.

13. Z. Zhang, "Too much covariates in a multivariable modelmay cause the problem of overfitting," Journal of Thoracic Disease, vol. 6, no. 9, pp. E196-E197, 2014.

14. O. Farghaly and P. Deshpande, "Fully convolutional neural network-based segmentation of brain metastases: a comprehensive approach for accurate detection and localization," vol. 36, p. 20711–20722, 2024.

15. M. Islam, G. Chen and S. Jin, "An Overview of Neural Network," American Journal of Neural Networks and Applications, vol. 5, no. 1, p. 5, 2019.

16. O. Farghaly and P. Deshpande, "Leveraging Machine Learning to Predict National Basketball Association Player Injuries," 2024 IEEE International Workshop on Sport, Technology and Research (STAR), pp. 216-221, 2024.

17. A. K. Sleiti, S. Gowid, . W. A. Al-Ammari and Y. AbuShanab, "Accurate prediction of dynamic viscosity of polyalpha-olefin boron nitride nanofluids using machine learning," Heliyon, vol. 9, no. el6716, 2023.

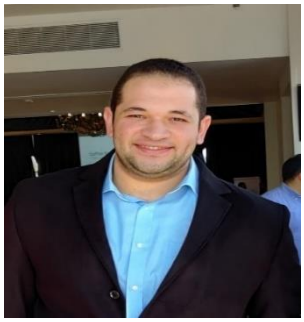
18. Mariano, Deyslen (2024), “Building Energy Consumption Datasets”, Mendeley Data, V1, doi: 10.17632/mzkyh37mtr.1

19. J. Zhang, X. Ma, J. Zhang, D. Sun, X. Zhou, C. Mi, H. Wen, "Insights into geospatial heterogeneity of landslide susceptibility based on the SHAP-XGBoost model," Journal of Environmental Management, vol. 332, p. art. no 117357, 2023.

Author Biographies

Yazeed AbuShanab is a mechanical engineering graduate with a master's degree and a PhD candidate at UWM. Currently an energy engineer at UWM-ITAC, their research focuses on energy systems, pipeline leakage detection, and thermodynamic property approximation using AI, integrating engineering and AI for innovative energy solutions.

Mohamed I. Youssef, a Ph.D. candidate and research assistant at UWM, holds an MSc from Florida Tech and a BSc in Mechanical Power Engineering. With experience in academia and industry, he specializes in project management, energy systems, wind turbines, and renewable energy. He holds CEM®, PMP®, RMP®, C-KPIP®, and CMRP® certifications.



Omar Shaker is a PhD candidate in Mechanical Engineering at the University of Wisconsin-Milwaukee, working under Professor Ryo Amano. As the Team Lead at UWM's Industrial Training and Assessment Center (ITAC), his research focuses on dual-rotor wind turbine performance, airfoil optimization, and energy efficiency. He contributes significantly to publications and training initiatives.

Mohamed Abdelaziz Youssef, M.E., is a Ph.D. candidate in Mechanical Engineering at the University of Wisconsin-Milwaukee and serves as an Energy Auditor at the Industrial Training and Assessment Center. He brings over seven years of professional experience in industrial process analysis, with a focus on enhancing energy efficiency and optimizing system performance.

Prof. Ryo Amano is the Richard & Joanne Grigg Fellow Professor, and Alan D. Kulwicki Fellow Professor, specializes in fluid mechanics, heat transfer, and energy systems. His research covers gas turbines, rocket engines, propulsion, aerodynamics, wind and hydro energy, biomass combustion, and wastewater treatment. He leads funded projects from NSF, DOE, NASA, and others and directs the DOE-funded Industrial Assessment Center.