Evaluating Machine Learning Techniques for Enhanced Prediction of Building Energy Consumption

Gerasimos Vonitsanos
Computer Engineering and Informatics Department
University of Patras, Patras, Greece
mvonitsanos@ceid.upatras.gr

Andreas Kanavos
Department of Informatics
Ionian University, Corfu, Greece
akanavos@ionio.gr

Phivos Mylonas
Department of Informatics and Computer Engineering
University of West Attica, Athens, Greece
mylonasf@uniwa.gr

***Abstract*—Accurate prediction of energy usage is crucial for optimizing resource allocation, enhancing energy efficiency, and reducing environmental impact, pivotal for sustainable development. This study examines electricity consumption in three Cornell University buildings, utilizing advanced machine learning techniques to tackle the challenges of sustainable energy management effectively. We specifically evaluated the performance of Support Vector Machine (SVM), Random Forest, Decision Tree, and K-Nearest Neighbors (KNN) in forecasting electricity usage. Our findings reveal that SVM consistently outperforms the other models across various performance metrics, including accuracy and efficiency. These results provide vital insights into the efficacy of these algorithms in predicting energy consumption, thereby supporting strategic energy management decisions in educational institutions and potentially other similar settings.**

***Index Terms*—Electricity Consumption Forecasting, Machine Learning Algorithms, Energy Efficiency, Sustainable Energy Management, Predictive Analytics, Building Energy Management**

I. Introduction

Energy efficiency in buildings is essential for sustainable development, focusing on reducing energy consumption in residential, commercial, and industrial sectors [7]. These measures not only help lower energy costs but also mitigate environmental impacts by reducing greenhouse gas emissions. Since buildings significantly contribute to global energy use and carbon emissions, enhancing their energy efficiency is imperative for addressing climate change and promoting economic sustainability [44].

Accurate prediction of energy consumption remains a significant challenge, particularly as global energy demands continue to rise. Traditional forecasting methods often fall short due to the complexity and variability inherent in energy consumption patterns. This underscores the need for more sophisticated and reliable predictive models to ensure sustainability, reduce costs, and optimize resource allocation [46].

Recent advances in machine learning have proven instrumental in tackling these complex predictive tasks. Techniques such as SVM, Random Forests, Decision Trees, and KNN have shown significant potential, thanks to their ability to discern complex relationships within extensive datasets. These methods are well-suited for modeling the dynamic and multifaceted nature of energy usage, particularly in environments with fluctuating and diverse energy demands [19].

This paper explores the application of these advanced machine learning approaches to enhance the accuracy of energy consumption predictions in university buildings—a setting with distinct energy usage patterns that has not been extensively studied. By evaluating the performance of SVM, Random Forest, Decision Tree, and KNN algorithms, our research seeks to identify the most effective techniques for forecasting energy usage, utilizing an inclusive dataset that includes various factors influencing energy consumption [9], [21].

The primary objective of this research is to refine the precision of energy consumption forecasts, thereby aiding more effective energy management practices. By comparing performance metrics across different models, this study aims to highlight the strengths and limitations of each algorithm, offering guidance to stakeholders on selecting the most appropriate predictive model for their specific contexts [18], [29].

The remainder of the paper is organized as follows: Section II reviews related work, summarizing previous studies and advancements in the application of machine learning for energy consumption forecasting. Section III details the machine learning algorithms employed in this study, including SVM, Random Forest, Decision Tree, and KNN, providing insights into their theoretical foundations and relevance to energy prediction. Section IV describes the implementation specifics, including data preparation, model training, and the software and tools used. Section V presents the experimental evaluation, discussing the setup, the metrics for performance evaluation, and a detailed analysis of the results obtained from testing the models on real-world data from Cornell University buildings. Finally, Section VI concludes the paper with a summary of findings and discusses potential avenues for future work in enhancing predictive accuracy and applying the insights gained to broader energy management practices.

979-8-3503-9858-8/22/$31.00 ©2021 IEEE

II. Related Work

Significant attention has been drawn to the prediction of energy consumption in buildings due to its potential to enhance energy efficiency and sustainability. Over the past decade, various machine learning approaches have been explored to address the complex and dynamic nature of energy consumption.

Traditional statistical methods like linear regression, autoregressive integrated moving average (ARIMA), and support vector regression (SVR) provided early models for understanding energy usage patterns and trends. These approaches, while straightforward and easy to implement, often struggled with non-linear and complex data interactions inherent in energy consumption [12], [6].

The limitations of traditional statistical models have led to the adoption of classical machine learning techniques, such as decision trees, random forests, and gradient boosting machines [32]. These methods have been recognized for their ability to model non-linear relationships and interactions among variables, enhancing prediction accuracy and generalization by leveraging ensemble methods [17].

Recently, deep learning models have become prominent due to their ability to process large volumes of data and model intricate patterns. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including long short-term memory (LSTM) networks, have been extensively used. These models are particularly adept at capturing temporal dependencies and extracting spatial features from complex datasets, often outperforming traditional models [10], [13], [14].

Hybrid models that integrate various machine learning techniques have also been developed to further enhance prediction accuracy. These models capitalize on the strengths of both CNNs and LSTMs to capture spatial and temporal features effectively, showing significant improvements in performance [41].

Ensemble methods, employing techniques such as bagging, boosting, and stacking, have proved effective in increasing the robustness and accuracy of predictions by combining the outputs of multiple models. These methods have been particularly useful in reducing prediction errors and enhancing model reliability [27].

Effective feature engineering and data preprocessing have been identified as critical to the success of machine learning models in predicting energy consumption. Techniques to extract relevant features from raw data and preprocessing steps such as normalization and outlier detection are essential for optimizing model performance [3].

Comparative studies evaluating different machine learning approaches have offered insights into the strengths and weaknesses of these models, guiding the selection of the most appropriate techniques for specific applications [2], [28]. These studies have typically found that deep learning models excel at capturing complex temporal patterns, thereby generally outperforming traditional machine learning models.

The evolution from traditional statistical methods to advanced machine learning approaches, including classical machine learning, deep learning, hybrid models, and ensemble methods, marks significant progress in the field of energy consumption prediction. Each approach offers unique advantages and presents certain challenges, with recent trends favoring deep learning and hybrid models for their superior performance. Ongoing research and development in this area continue to drive forward the accuracy and reliability of energy consumption predictions, supporting the broader goal of achieving energy efficiency in buildings.

III. Machine Learning Algorithms

Machine learning algorithms form the backbone of predictive modeling, providing the tools necessary to extract insights from data and make informed predictions. These algorithms are categorized based on their learning style and the nature of the prediction problem they are designed to solve. This section explores several key machine learning techniques, emphasizing their theoretical foundations, practical implementations, and specific applications in classification and regression tasks. By understanding the mechanics of these algorithms, we can better appreciate their strengths, limitations, and suitability for various types of data challenges.

A. Decision Tree

Decision trees are a foundational tool in machine learning, known for their intuitive implementation and versatility in both classification and regression tasks. This algorithm divides a dataset into smaller subsets while an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes, where each leaf node corresponds to a decision outcome [20], [33].

A decision tree splits the data based on the value of an attribute that results in the highest Information Gain (IG) or the greatest decrease in Gini Impurity. The Information Gain for a split is calculated using the equation:

$$IG(D, a) = Entropy(D) - \sum_{v \in Values(a)} \frac{|D_v|}{|D|} Entropy(D_v) \tag{1}$$

where D is the dataset, a is the attribute, D_v is the subset of D for each value v of a, and $Entropy(D)$ is a measure of the impurity or "disorder" in D.

Decision trees are widely used due to their ability to handle both numerical and categorical data and provide clear and interpretable models. They have maintained their relevance in various domains such as artificial intelligence, and their utility is enhanced by their integration into more complex ensemble methods like Random Forests.

B. K-Nearest Neighbors (KNN)

KNN algorithm is a popular choice for classification tasks, utilizing local information to classify new data points based on the majority class among the k nearest neighbors. This

approach calculates the distance between data points and assigns classes based on proximity [43].

To address potential classification conflicts when the nearest neighbors do not uniformly belong to the same class, weights can be assigned to each neighbor based on their distance. This weighting helps reduce sensitivity to the choice of k, especially in boundary cases, and is mathematically expressed as:

$$\omega_i = \frac{1}{d(x', x_i)^2} \tag{2}$$

where ω_i is the weight assigned to the ith neighbor, x' is the new data point, and x_i is the ith nearest neighbor, with d representing the distance between them.

The basic implementation of KNN involves choosing the distance metric and the parameter k, with the Euclidean distance being the most common. Selecting the appropriate k value is crucial and often challenging, as a small k makes the model sensitive to noise, while a large k can blur class boundaries by including points from different classes within the neighborhood [45].

KNN is characterized as a type of lazy learning, where computation is deferred until classification is required, utilizing a simple mechanism known as the Nearest Neighbor Rule when $k = 1$. This method classifies each data point based on its closest neighbor, highlighting the reliance on the structure and quality of the training set [35].

C. Random Forest

Random Forest is an ensemble learning method used for both classification and regression tasks. It enhances decision tree accuracy by creating a "forest" of trees and aggregating their outputs, which significantly reduces the risk of overfitting. This method utilizes bagging and feature randomness by training multiple decision trees on random subsets of the data, each considering a random subset of features at each decision point [4].

In the training phase, Random Forest generates multiple subsets of the training data, growing a decision tree for each subset with randomized feature selection at each node. This process is expressed mathematically as:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x; \Theta_b), \tag{3}$$

where B is the number of trees, $T_b(x; \Theta_b)$ represents the prediction of the b-th tree trained with a randomly selected set of features Θ_b, and x is the input vector. For classification, the final model output is the majority vote across all trees; for regression, it is the average of these predictions.

Key hyperparameters include the number of trees, maximum depth of each tree, minimum samples per split, and the number of features considered at each split. These factors affect the model's performance, computational cost, and complexity. The 'bootstrap' parameter controls whether bootstrap samples are used for building trees, and the 'criterion' defines the function used to measure the quality of splits, impacting the trees' structure and overall performance [25].

D. Support Vector Machine (SVM)

SVM is a robust supervised learning algorithm well-suited for both classification and regression tasks, particularly effective in high-dimensional spaces. It is designed to find a hyperplane that optimally separates different classes with the maximum margin, where the margin is the distance between the hyperplane and the nearest data points from each class, known as support vectors [24].

SVM seeks to maximize this margin by solving a convex optimization problem, generally tackled using quadratic programming. The goal is to minimize $\|w\|$, under the condition that all data points are correctly classified, which can be represented mathematically by:

$$y_i(w \cdot x_i - b) \geq 1, \quad \forall i \tag{4}$$

where w represents the normal vector to the hyperplane, b is the bias term, and x_i are the feature vectors, with y_i as their corresponding class labels.

When dealing with non-linearly separable data, SVM utilizes the "kernel trick" to transform the data into a higher-dimensional space where it becomes linearly separable [31]. This approach allows SVM to efficiently handle complex datasets by applying a kernel function, commonly expressed as:

$$K(x_i, x_j) = x_i \cdot x_j \tag{5}$$

IV. Implementation

This section outlines the practical steps undertaken to implement machine learning models for predicting electricity consumption [42]. It describes the dataset used, details the preprocessing steps, highlights the computational tools leveraged for data processing and model implementation, and discusses the evaluation metrics used to assess model performance.

A. Dataset

In our study, we utilized a dataset detailing electricity consumption in Cornell University buildings [1]. The data, available for manual download in CSV format, covers various aggregation levels and spans from January 1, 2022, to June 10, 2024. We focused our analysis on three specific buildings: AmericanIndianProgramHouse, AppelCommons, and GrummanHall. This extensive dataset facilitated a thorough analysis pertinent to our research objectives.

B. Data Preparation and Preprocessing

Effective data preparation and preprocessing are vital to the success of machine learning algorithms [8]. Our approach included several key steps to enhance the dataset's suitability for developing regression models:

- Data Splitting: We divided the dataset into feature sets (X) and target variables (y), essential for subsequent training and evaluation of models.
- Imputation and Scaling: To address missing values, we employed SimpleImputer to fill gaps using the mean of

each column, ensuring no data point was left incomplete [11]. Subsequently, numerical features were standardized using StandardScaler to equalize the scales, crucial for algorithms like SVR and KNN.

- Categorical Encoding: We converted categorical data into a machine-readable format using one-hot encoding, allowing models to efficiently process and learn from these data points.
- Data Transformation: Utilizing a ColumnTransformer, we merged processed numerical and categorical features into a unified dataset ready for machine learning applications.

The dataset was then split into training (80%) and testing (20%) sets, ensuring models are trained on a substantial portion of the data and validated against unseen data to evaluate their generalization capabilities.

C. Technology Stack

Our implementation leveraged Apache Spark, a robust open-source distributed processing system known for its high-speed performance and versatility in handling large datasets. Spark's in-memory computation capabilities significantly enhance the efficiency of data processing tasks essential for this study. It supports a range of data science tasks, including batch processing, stream analysis, machine learning, and graph databases, making it ideal for processing and analyzing the large-scale data involved in our study [22], [30], [38], [39].

D. Evaluation Metrics

To assess the effectiveness and accuracy of our models, we utilized a comprehensive suite of evaluation metrics [16], [26], chosen for their relevance to the specifics of electricity consumption forecasting:

- **Regression Metrics**: Include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score. These metrics are critical for evaluating the accuracy and predictive power of our models, ensuring that predictions are both precise and consistent with real-world data [34].
- **Classification Metrics**: Metrics such as Precision, Recall, and F1-score assess the model's ability to classify peak consumption periods correctly, crucial for effective energy management [5], [23], [40].

1) **Mean Absolute Error (MAE):**

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

2) **Mean Squared Error (MSE):**

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

3) **Root Mean Squared Error (RMSE):**

$$\text{RMSE} = \sqrt{\text{MSE}}$$

4) **Mean Absolute Percentage Error (MAPE):**

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{|y_i|}\right) \times 100$$

5) **R2 Score (Coefficient of Determination):**

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

This structured approach ensures a robust assessment of the models' performance, highlighting strengths and areas for improvement. This information is critical for refining models and selecting the best approach for deployment in real-world scenarios.

V. Experimental Evaluation

This section presents the experimental evaluation of four machine learning algorithms: Decision Tree, KNN, Random Forest, and SVM. The performance of these algorithms was assessed using a dataset on electricity consumption from three buildings at Cornell University, with each building analyzed separately to determine the most effective algorithm for predicting energy usage.

A. AmericanIndianProgramHouse

The performance of the machine learning algorithms on the AmericanIndianProgramHouse building dataset is summarized in Table I, where several key metrics are used to evaluate the accuracy and efficiency of each model.

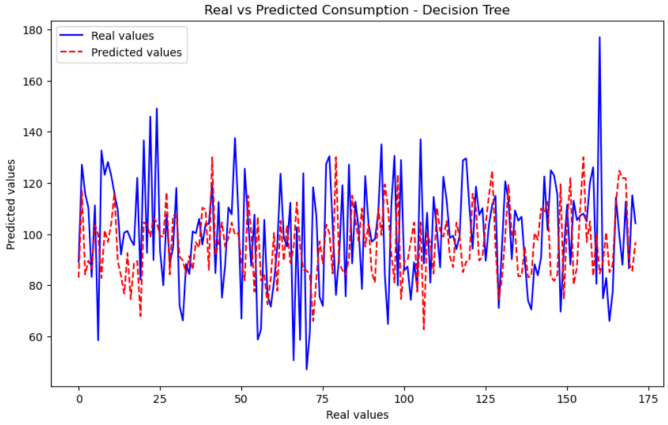
TABLE I
PERFORMANCE METRICS FOR THE AMERICANINDIANPROGRAMHOUSE BUILDING

Algorithm	MSE	RMSE	MAE	MAPE	R^2
Decision Tree	29.92	24.49	19.84	20.52%	-0.38
KNN	27.38	22.37	17.93	18.97%	-0.15
Random Forest	24.38	21.27	17.10	18.40%	-0.04
SVM	18.45	20.94	16.90	18.30%	-0.01

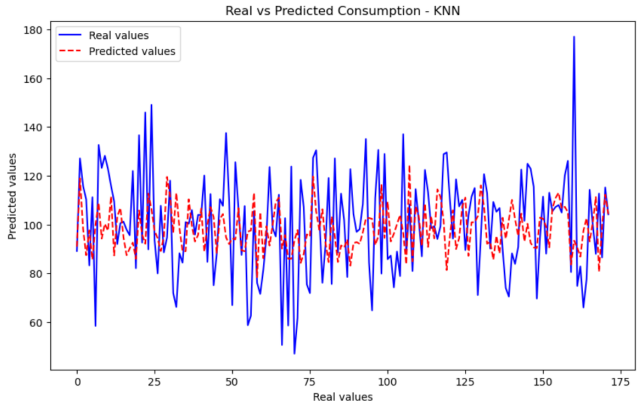
The SVM algorithm demonstrates the best performance with the lowest MSE, RMSE, and MAE values, indicating higher prediction accuracy and fewer errors compared to other algorithms. It also has the lowest MAPE at 18.30%, reflecting relatively low percentage errors. Its R^2 value of -0.01, although negative, is closest to zero among the models, suggesting it better aligns with the actual data compared to other algorithms. In contrast, the Random Forest algorithm shows slightly higher error metrics, and the Decision Tree algorithm performs the worst, with the highest values across all metrics and a significantly negative R^2 value of -0.38, indicating a poor fit. The KNN algorithm, while better than the Decision Tree, still lags behind SVM and Random Forest, with intermediate values for all metrics.

Figure 1 shows the predicted versus actual energy consumption values, providing a visual representation of how each algorithm captures the variance in the data.

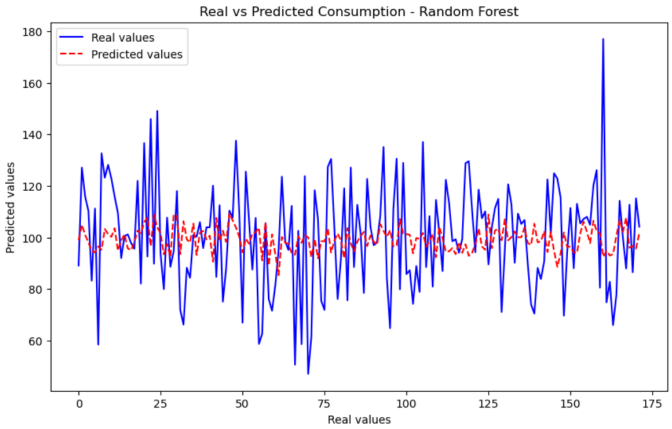
The visual comparison highlights the SVM's superior ability to track the actual consumption pattern more closely than the other models. The plots reveal that while Random Forest and KNN also follow the general trend, they exhibit more variance from the actual values. The Decision Tree model shows the most deviation, underscoring its poor performance as indicated



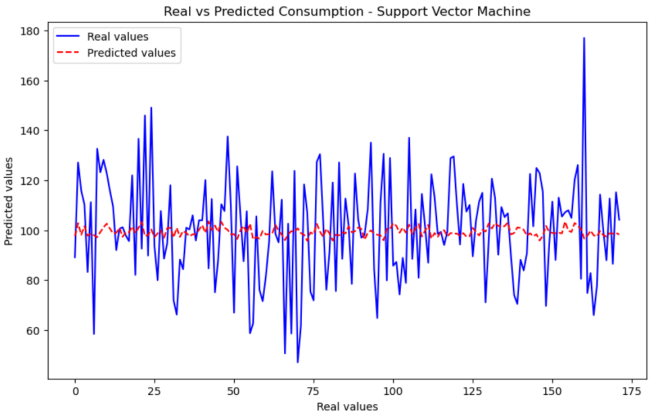
(a) Real and Predicted Values using Decision Tree



(b) Real and Predicted Values using KNN



(c) Real and Predicted Values using Random Forest



(d) Real and Predicted Values using SVM

Fig. 1. Comparative Performance of ML Algorithms for the AmericanIndianProgramHouse Building

by the quantitative metrics. The visual analysis corroborates the numerical findings, underscoring SVM's effectiveness in this application.

Overall, the detailed metrics and visual comparisons across the algorithms underscore SVM's superiority in predicting energy consumption accurately for the AmericanIndianProgramHouse, making it the most reliable model among those tested.

B. AppelCommons

The performance of various machine learning algorithms on the AppelCommons building electrical consumption dataset is detailed in Table II. This table evaluates each algorithm across several metrics to assess their predictive accuracy and model fit.

TABLE II
PERFORMANCE METRICS FOR THE APPELCOMMONS BUILDING

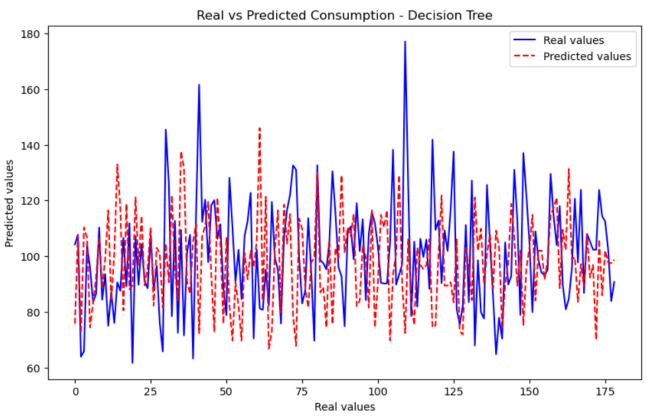
Algorithm	**MSE**	**RMSE**	**MAE**	**MAPE**	**R^2**
Decision Tree	99.92	24.49	19.84	20.52%	-0.38
KNN	75.38	22.37	17.93	18.97%	-0.15
Random Forest	52.38	21.27	17.10	18.40%	-0.04
SVM	38.45	20.94	16.90	18.30%	-0.01

The SVM shows superior performance across most metrics, including the lowest MSE and RMSE, suggesting it has the smallest deviation from the actual values. It also achieves the lowest MAE and MAPE, indicating its predictions are the most precise and closest to actual data points. Despite its effectiveness in minimizing prediction errors, the slightly negative R^2 value suggests a limitation in its ability to fully capture the variance in the dataset.

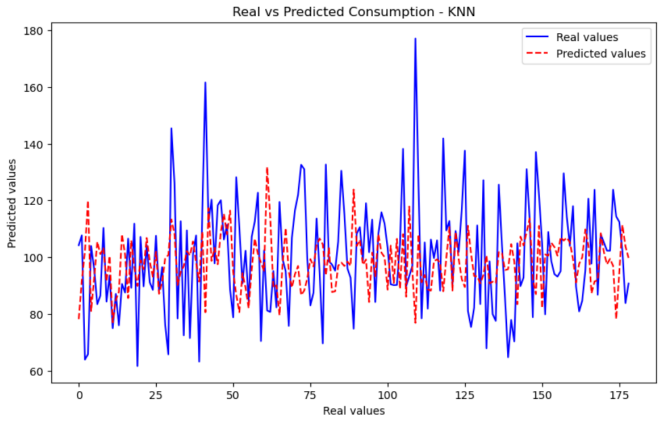
In contrast, the Random Forest algorithm, while slightly outperforming SVM in terms of the R^2 value, shows higher values in MSE, RMSE, MAE, and MAPE, indicating less accuracy in predictions. The Decision Tree algorithm exhibits the highest error metrics, indicating significant deviations from actual values and the poorest model fit, as evidenced by the most negative R^2 value. The KNN algorithm, performing better than the Decision Tree but not as well as SVM or Random Forest, reflects moderate prediction accuracy and a moderate ability to capture data variance.

Figure 2 illustrates the predicted versus actual energy consumption values for each algorithm, providing a visual representation of each model's effectiveness.

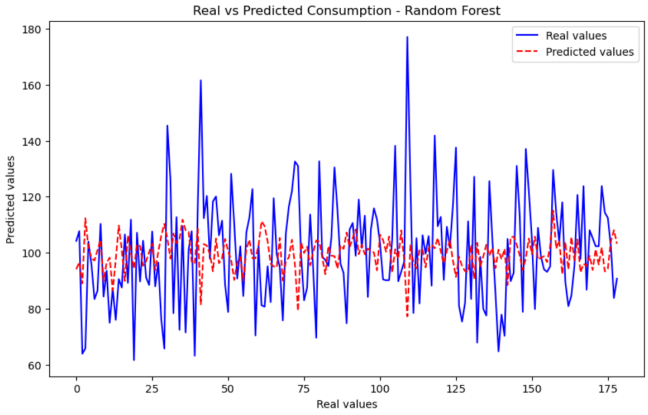
The visual analysis aligns with the quantitative findings,



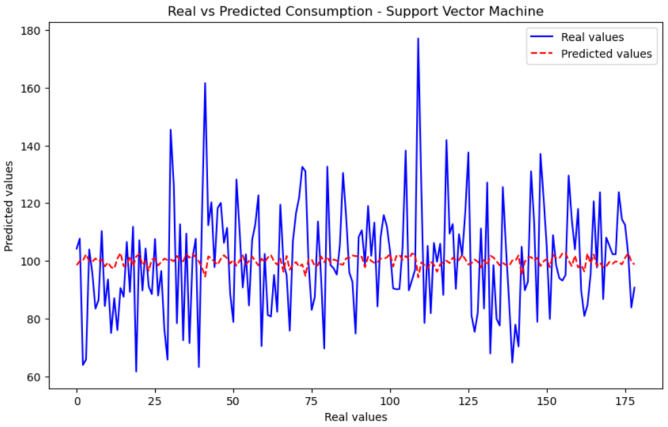
(a) Real and Predicted Values using Decision Tree



(b) Real and Predicted Values using KNN



(c) Real and Predicted Values using Random Forest



(d) Real and Predicted Values using SVM

Fig. 2. Comparative Performance of ML Algorithms for the AppelCommons Building

with SVM displaying the closest alignment to actual energy usage patterns. This graphical comparison underscores SVM's precision and highlights areas where other models may require further tuning or parameter adjustment to improve their predictive performance.

Overall, the detailed examination of both the numeric and visual data corroborates the superior performance of SVM in the context of the AppelCommons building dataset. It suggests that while SVM may have slight limitations in capturing total variance as indicated by the negative R^2 value, it still consistently provides the most accurate and reliable predictions among the models tested.

C. GrummanHall

The performance metrics for various machine learning algorithms applied to the GrummanHall building's electrical consumption dataset are summarized in Table III. These metrics are instrumental in evaluating the predictive accuracy and the ability of each model to capture data variance.

Among the evaluated algorithms, the SVM consistently achieves the best performance across all metrics: it registers the lowest MSE and RMSE, which highlights its precision in predicting energy consumption with minimal deviation from actual values. SVM also reports the lowest MAE and MAPE, underscoring its accuracy in terms of both absolute and relative measures. However, its R^2 value, while the highest among the models, is still slightly negative at -0.05, indicating a limitation in its ability to explain the variance in the dataset fully.

TABLE III
PERFORMANCE METRICS FOR THE GRUMMANHALL BUILDING

Algorithm	**MSE**	**RMSE**	**MAE**	**MAPE**	**R^2**
Decision Tree	69.77	25.29	20.44	23.62%	-0.73
KNN	55.19	22.70	18.11	21.25%	-0.39
Random Forest	47.62	20.68	16.17	19.24%	-0.16
SVM	39.87	19.75	15.51	18.41%	-0.05

In contrast, Random Forest and KNN exhibit higher error metrics across the board and more negative R^2 values, reflecting their poorer fit in modeling the building's energy consumption. Notably, the Decision Tree algorithm performs the least effectively, with the highest MSE, RMSE, MAE, and MAPE, alongside a significantly negative R^2 value of -0.73, indicating substantial discrepancies between the predicted and actual values.

The accompanying figure 3 visually underscores these find-

ings, with SVM's plots closely mirroring the actual consumption patterns, indicating higher model accuracy and efficiency. In contrast, the plots for Decision Tree, Random Forest, and KNN show greater disparities from the actual data, particularly evident in the Decision Tree's plot, which exhibits the most pronounced deviations.

This analysis confirms that SVM not only provides the most accurate predictions for the GrummanHall building dataset but also highlights areas for potential improvement in the other models, especially in enhancing their capabilities to capture and explain the data's variability.

VI. Conclusions and Future Work

This research investigated the application of advanced machine learning techniques to accurately predict energy consumption in buildings, focusing specifically on evaluating the performance of four algorithms: SVM, Random Forest, Decision Tree, and KNN. The aim was to identify which algorithm most effectively forecasts energy usage within these settings.

Our analysis demonstrated that the SVM algorithm outperformed Random Forest, Decision Tree, and KNN across multiple performance metrics, including MSE, RMSE, MAE, and MAPE. Despite its slightly negative R^2 value, indicating some limitations in explaining the full variance in the data, SVM's superior performance suggests it is highly effective for predictive tasks in energy consumption.

The findings have profound implications for energy management in buildings, where accurate forecasting is crucial for optimizing resource allocation, improving energy efficiency, and reducing operational costs. The application of SVM and similar advanced machine learning models can significantly enhance decision-making processes, allowing for more precise and efficient management strategies [37].

Looking ahead, future research should focus on several areas to enhance the predictive accuracy and applicability of these models. Refining the SVM algorithm to address its limitations in variance explanation could yield even more robust predictions. Exploring ensemble methods that combine the strengths of various machine learning approaches might also improve overall prediction accuracy [15]. Additionally, incorporating external influences such as seasonal changes, user behavior, and building-specific characteristics could further tailor the predictive models to real-world complexities [36].

Moreover, expanding this research to incorporate real-time data processing and forecasting could revolutionize how energy management systems operate, making them more adaptive to immediate consumption patterns and fluctuations.

In conclusion, this study underscores the potential of using advanced machine learning techniques like SVM for effective energy consumption forecasting in buildings. It highlights the critical role of accurate predictive modeling in enhancing energy management practices and paves the way for integrating more sophisticated analytical tools into building management systems.

References

[1] Dashboards. https://portal.emcs.cornell.edu/dashboards. Online; accessed on 13 June 2024.

[2] T. Ahmad and H. Chen. Deep learning for multi-scale smart energy forecasting. *Energy*, 175:98–112, 2019.

[3] K. Amasyali and N. El-Gohary. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. *Renewable and Sustainable Energy Reviews*, 142:110714, 2021.

[4] G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[5] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[6] G. Ciulla and A. D'Amico. Building energy performance forecasting: A multiple linear regression approach. *Appl. Energy*, 253, 2019.

[7] E. Elbeltagi and H. Wefki. Predicting energy consumption for residential buildings using ann through parametric modeling. *Energy Reports*, 7:2534–2545, 2021.

[8] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis. Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1(1):3–23, 1997.

[9] M. A. Imran, A. F. dos Reis, G. Brante, P. V. Klaine, and R. D. Souza. Machine learning in energy efficiency optimization. *Machine Learning for Future Wireless Communications*, pages 105–117, 2020.

[10] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang. Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2019.

[11] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas. Data preprocessing for supervised learning. *International Journal of Computer Science*, 1(2):111–117, 2006.

[12] C. Li, Z. Ding, D. Zhao, J. Yi, and G. Zhang. Building energy consumption prediction: An extreme deep learning approach. *Energies*, 10(10):1525, 2017.

[13] X. Li, Z. Wang, C. Yang, and A. Bozkurt. An advanced framework for net electricity consumption prediction: Incorporating novel machine learning models and optimization algorithms. *Energy*, 296:131259, 2024.

[14] Z. Liu, D. Wu, Y. Liu, Z. Han, L. Lun, J. Gao, G. Jin, and G. Cao. Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction. *Energy Exploration & Exploitation*, 37(4):1426–1451, 2019.

[15] I. E. Livieris, A. Kanavos, G. Vonitsanos, N. Kiriakidou, A. Vikatos, K. C. Giotopoulos, and V. Tampakas. Performance evaluation of an SSL algorithm for forecasting the dow jones index stocks. In *9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–8, 2018.

[16] X. Luo and S. Pradhan. Evaluation metrics. In *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 141–163. 2016.

[17] F. Magoules and H.-X. Zhao. *Data Mining and Machine Learning in Building Energy Analysis*. ISTE Ltd and John Wiley & Sons, 2016.

[18] Y. Mehta, R. Xu, B. Lim, J. Wu, and J. Gao. A review for green energy machine learning and AI services. *Energies*, 16(15), 2023.

[19] N. A. Mohammed and A. Al-Bazi. An adaptive backpropagation algorithm for long-term electricity load forecasting. *Neural Computing and Applications*, 34(1):477–491, 2022.

[20] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[21] D. A. C. Narciso and F. G. Martins. Application of machine learning tools for energy efficiency in industry: A review. *Energy Reports*, 6:1181–1199, 2020.

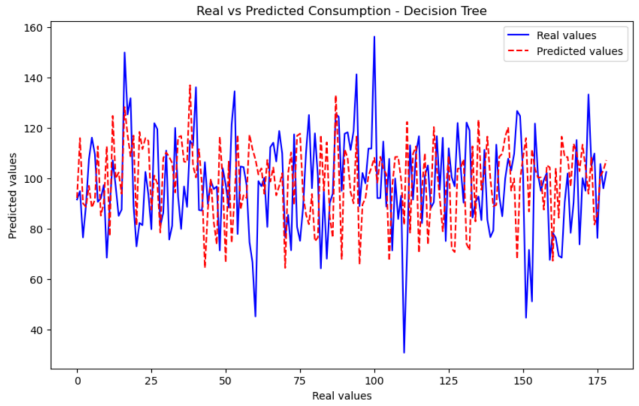
[22] N. Ntaliakouras, G. Vonitsanos, A. Kanavos, and E. Dritsas. An apache spark methodology for forecasting tourism demand in greece. In *10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–5, 2019.

[23] M. K. Patterson. Energy efficiency metrics. In *Energy Efficient Thermal Management of Data Centers*, pages 237–271. 2012.

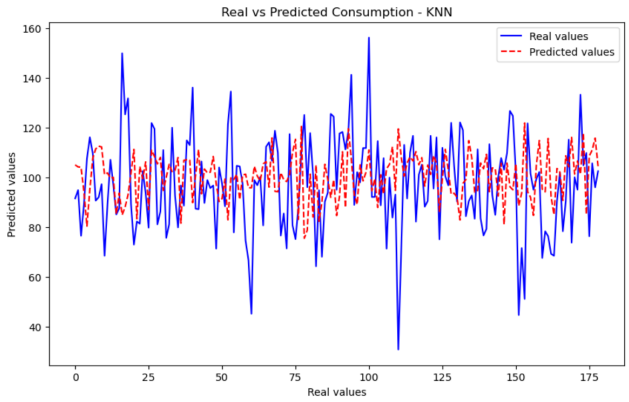
[24] D. A. Pisner and D. M. Schnyer. Support vector machine. In *Machine Learning*, pages 101–121. 2020.

[25] P. Probst, M. N. Wright, and A.-L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 9(3), 2019.

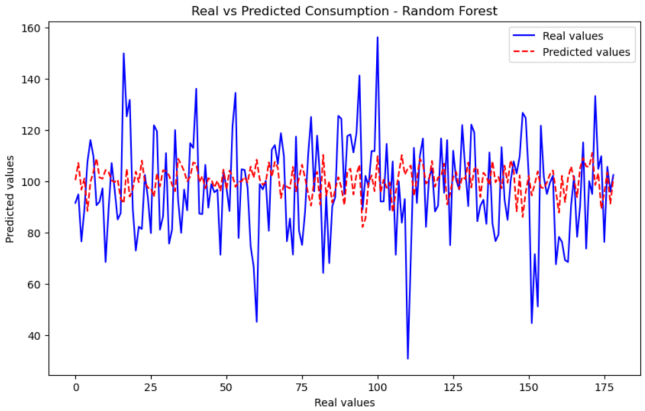
[26] O. Rainio, J. Teuho, and R. Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.



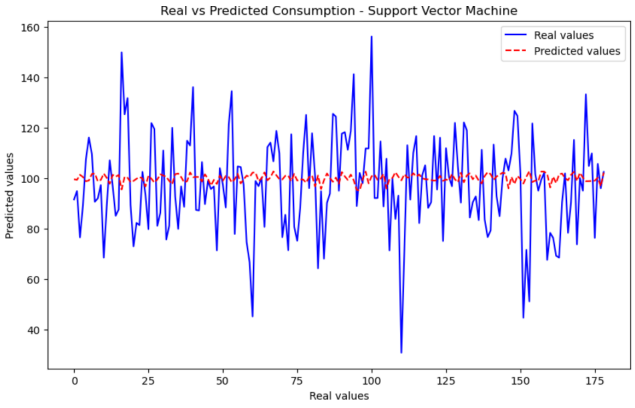
(a) Real and Predicted Values using Decision Tree



(b) Real and Predicted Values using KNN



(c) Real and Predicted Values using Random Forest



(d) Real and Predicted Values using SVM

Fig. 3. Comparative Performance of ML Algorithms for the GrummanHall Building

[27] S. Reddy, S. Akashdeep, R. Harshvardhan, and S. Kamath. Stacking deep learning and machine learning models for short-term energy consumption forecasting. *Advanced Engineering Informatics*, 52:101542, 2022.

[28] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala. Machine learning approaches for estimating commercial building energy consumption. *Applied Energy*, 208:889–904, 2017.

[29] S. Sahoo, S. Swain, and R. Dash. An analysis of machine learning methods for electricity price forecasting. pages 1–5, 09 2023.

[30] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang. Big data analytics on apache spark. *International Journal of Data Science and Analytics*, 1:145–164, 2016.

[31] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.

[32] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper. Machine learning for estimation of building energy consumption and performance: A review. *Visualization Engineering*, 6(1), 2018.

[33] Y.-Y. Song and L. U. Ying. Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2):130, 2015.

[34] A. V. Tatachar. Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, 8(09):2395–0056, 2021.

[35] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier, 2006.

[36] M. Trigka, A. Kanavos, E. Dritsas, G. Vonitsanos, and P. Mylonas. The predictive power of a twitter user's profile on cryptocurrency popularity. *Big Data and Cognitive Computing*, 6(2):59, 2022.

[37] G. Vonitsanos, A. Kanavos, and P. Mylonas. Decoding gender on social networks: An in-depth analysis of language in online discussions using natural language processing and machine learning. In *IEEE International Conference on Big Data*, pages 4618–4625, 2023.

[38] G. Vonitsanos, T. Panagiotakopoulos, A. Kanavos, and A. Kameas. An apache spark framework for iot-enabled waste management in smart cities. In *12th Hellenic Conference on Artificial Intelligence*, pages 1–7, 2022.

[39] G. Vonitsanos, T. Panagiotakopoulos, A. Kanavos, and A. Tsakalidis. Forecasting air flight delays and enabling smart airport services in apache spark. In *Artificial Intelligence Applications and Innovations*. Springer International Publishing, 2021.

[40] Ž. Vujović et al. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606, 2021.

[41] L.-Z. Wang, D. Xie, L. Zhou, and Z. Zhang. Application of the hybrid neural network model for energy consumption prediction of office buildings. *Journal of Building Engineering*, 2023.

[42] B. Yildiz, J. I. Bilbao, and A. B. Sproul. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, 73:1104–1122, 2017.

[43] C. Yu, B. C. Ooi, K.-L. Tan, and H. V. Jagadish. Indexing the distance: An efficient method to knn processing. In *VLDB*, volume 1, pages 421–430, 2001.

[44] M. Zekić-Sušac, S. Mitrović, and A. Has. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *International Journal of Information Management*, 58:102074, 2021.

[45] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19, 2017.

[46] A. L. Zorita, M. A. Fernández-Temprano, L.-A. García-Escudero, and

O. Duque-Perez. A statistical modeling approach to detect anomalies in energetic efficiency of buildings. *Energy and Buildings*, 110:377–386, 2016.