Development of an Energy Prediction Model for Home Appliances Using Random Forest Algorithm

M. KrishnaParamathma
Department of Electrical and Electronics Engineering
Kalasalingam Academy of Research and Education
m.krishnaparamathma@klu.ac.in

Pasupuleti Charan Kumar
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
pasupuleticharankumar16@gmail.com

Pasupuleti Dharani Kumar
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
pasupuletidharani2004@gmail.com

Singanamala Sharmas Vali
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
singanamalasharmasvali@gmail.com

Gajula Veera Naga Keerthi
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
gajulakeerthi22@gmail.com

Santhiya
Department of Information and Technology
Kalasalingam Academy of Research and Education
santhiyap736@gmail.com

ABSTRACT - **Rising energy costs and mounting environmental pressures have required effective domestic energy management. Traditional prediction models are destined to fail in capturing non-linear, intricate appliance energy consumption patterns and yield false forecasts. To address this shortcoming, we developed a data-driven model founded on a Random Forest algorithm for domestic energy consumption forecasting. The proposed model employs a mix of sensor-based features like temperature, humidity, and weather to precisely predict energy consumption. The approach includes extreme data preprocessing to eliminate missing values and outliers, rigorous exploratory analysis to detect usage patterns, and careful feature engineering to maximize input variables. We selected the Random Forest algorithm because of its proven ability to capture non-linear interaction and resistance to overfitting. Through rigorous training and hyperparameter optimization, the model demonstrated outstanding performance metrics like a 97.8% R^2 value and 0.05 MAE, decisively outperforming traditional models. The result not only provides highly accurate forecasts of energy consumption but also actionable insights on peak usage time, enabling domestic households to reduce energy consumption and save dollars. This work demonstrates the vast potential of machine learning in maximizing the efficiency of energy consumption and advancing environmental sustainability agendas, particularly by the UN Sustainable Development Goals for affordable clean energy (SDG 7) and climate action (SDG 13).**

Keywords: Energy Prediction, Random Forest Algorithm, Machine Learning, Exploratory Data Analysis (EDA), Affordable and Clean Energy (SDG 7), Sustainable Cities and Communities (SDG 11), Climate Action (SDG 13)

1. INTRODUCTION

The increasing need for electricity for domestic use poses stern challenges to environmental and economic sustainability. Because home appliances use a significant portion of the household energy, their use must be optimized. The increasing costs of electricity, coupled with environmental issues due to carbon emissions and energy resource consumption, call for intelligent solutions to optimize energy consumption. Most classical energy forecasting models are incapable of detecting the underlying complex, nonlinear trends within energy usage patterns, which translates to below-optimistic prediction accuracy. With their capability of processing enormous volumes of data and mapping complex associations, machine learning (ML) methods hold promise as a possible substitute. Among these, the Random Forest algorithm ranks highly in terms of reliability, general applicability, and capability of tolerating nonlinearities and large-dimensional data.

This study attempts to create a predictive model for domestic appliance energy consumption based on data obtained using the Random Forest algorithm. The model accurately predicts based on sensor-based features such as temperature, humidity, and weather. A rigorous process of data preprocessing, exploratory data analysis, feature engineering, and model training guarantees the accuracy and efficiency of predictions. In addition to its technological strength, this study aligns with international action towards sustainability, especially the United Nations' Sustainable Development Goals (SDGs). Through the advancement of energy efficiency (SDG 7), enabling sustainable urban living (SDG 11), and encouraging climate action (SDG 13), this study recognizes the potential of ML to tackle the most critical environmental issues. The findings of this research are not only meant to advance the prediction of energy consumption but also to provide actionable data to households to maximize their energy efficiency. This can be applied to reduce electricity costs, minimize the carbon footprint, and support broader environmental sustainability initiatives.

The main contribution of this research lies in the development of a robust, interpretable, and scalable Random Forest-based prediction model for appliance energy consumption, which effectively incorporates sensor and weather data to deliver highly accurate forecasts. Unlike prior works, our model balances performance, interpretability, and ease of deployment for real-world domestic environments.

II. LITERATURE SURVEY

Luis M. Candanedo, Veronique Feldheim, and Dominique Deramaix(2020) developed data-driven predictive models to investigate and forecast appliance energy consumption in low-energy residential setups. Taking temperature, humidity from wireless sensors, and weather from nearby weather stations as inputs, the study compared four statistical models: multiple linear regression, support vector machines, random forests, and gradient boosting machines (GBM). Of all, GBM performed best with an R^2 of 97% in training and 57% in testing, validating its superior performance in handling complex, non-linear relationships. The study concluded that kitchen, laundry, and living room data were most important for energy prediction, and atmospheric pressure was the most significant weather parameter for energy consumption. The findings demonstrate the possibility of using combined data from sensor networks and weather stations to enhance predictive accuracy, thus making applications such as energy optimization, anomaly detection, and demand-side management in residential energy systems feasible.[1]

Ismael Jrhilifa, Hamid Ouadi, Abdelihah Jibab, Nada Mounir, Abdellah Ouaguid (2024) proposed a highly effective method of home energy management through Multi-Individual Load Disaggregation (MILD) and Multi-Individual Load Forecasting (MILF) through Variational Mode Decomposition (VMD) and deep learning. The method decomposed aggregate power signals into intrinsic mode functions (IMFs) to track individual appliances at the level, enabling better load monitoring and forecasting. The addition of weather conditions and calendar activities as input variables also enhanced the performance of the model, preventing the effect of external variables on energy consumption. The GRU-based model performed better than the others, like LSTM and CNN, with an RMSE of 9.54 W and an MAPE of 8.10%. The study highlighted the strength and scalability of VMD in dealing with noise and non-stationary signals and thus emerged as a potential solution for real-time load forecasting and disaggregation. The study adds to the practice of sustainable energy by providing actionable information on appliance-level energy consumption and making household energy management easier.[2]

Beldar (2024) explored appliance energy forecasting using supervised machine learning techniques, with a dataset of 29 features from indoor humidity and temperature to weather features, including wind speed and pressure. The study included preprocessing techniques such as outlier removal and feature selection using a Random Forest Regressor to enhance model stability. Regression models were contrasted in terms of metrics such as RMSE, MAE, and R^2, with significant predictors including temperature and humidity in various rooms. Atmospheric pressure was identified as a significant weather variable influencing appliance energy consumption, according to the study. Sophisticated data visualization techniques such as scatter plots and correlation analysis provided profound insights into consumption patterns, with maximum consumption occurring on weekdays and weekends. The use of feature engineering in the study emphasizes the requirement for extensive data preprocessing and validation in the design of accurate and scalable energy forecasting systems. [3]

Suhermi, Rahida Rihhadatul Aisy (2024) proposed a Functional Data Analysis (FDA) model for household appliance energy consumption forecasting with a focus on dynamic interactions between energy consumption and weather conditions like temperature and humidity. Compared to the traditional approaches, the FDA depicts data as continuous functions, which can capture time-varying trends. The FDA model performed better than linear regression, SVM, and Random Forest with the least RMSE (83.93) and MAPE (0.609). The approach smoothed noise and maintained vital consumption trends, and can be applied for energy management optimization and forecasting future consumption trends in smart homes.[4]

Khaoula Elhabyb, Amine Baina, Mostafa Bellafkih, and Ahmed Farouk Deifalla (2024) explored machine learning models to predict school energy consumption with LSTM, random forests, and gradient boosting regressors. From IoT sensor readings, the approach derived the key features of occupancy behavior, HVAC usage, and climatic conditions. Gradient boosting performed well with a mean squared error (MSE) of 8.148 and $R^2 > 0.98$ in some cases. The research revealed that tailored machine learning models had the potential to reduce energy management costs, optimize energy management, and enable sustainability in institutions.[5]

Based on an extensive review of existing research and comparative experimentation, Random Forest has emerged as a powerful and effective model for energy consumption prediction. While Gradient Boosting and LSTM models possessed outstanding prediction power, particularly in detecting non-linear patterns and sequential relationships, they tended to require intense computation, hyperparameter adjustment, and larger datasets. Experiments conducted by Candanedo and Talukdar verified the strength of boosting models, and others like Vakharia and Robbia exhibited Random Forest's competitive accuracy and simplicity. Our work with Random Forest also verifies this, registering improved performance metrics compared to both GBM and LSTM. Furthermore, the feature importance interpretability of Random Forest and stability across various datasets place it in an appealing position for real-time and scalable energy forecasting applications, where transparency and computational efficiency are especially important.

III. METHODOLOGY

Data Collection:

The data used in this study was derived from the UCI Machine Learning Repository, which is famous for offering credible data for machine learning research.[6] The data consists of 4.5 months' worth of energy consumption in a home. It consists of 29 attributes and 19,735 instances, with no missing or duplicated values, and is a solid dataset for prediction analysis.

The main features are appliance energy consumption in watt-hours (Wh), indoor conditions like temperature and relative humidity in various rooms (e.g., kitchen, living room, bedroom, and laundry room), and outdoor conditions like wind speed, visibility, atmospheric pressure, and temperature. Temporal information, like date and time, was also added to identify trends in energy consumption at various times of the day. Derived features, i.e., random features to verify model stability, were also added to the dataset. The wide variety of features guarantees the creation of a correct prediction model, taking into account both environmental and temporal factors affecting energy consumption.

Data Preprocessing:

Preprocessing of data was necessary to enhance data quality and consistency for model development. Missing value handling was the initial step. Although missing values were absent in the dataset, methods such as mean/mode imputation and interpolation were reserved for future use on similar datasets. Outliers were detected and eliminated using statistical techniques such as the Interquartile Range (IQR) method and Z-score analysis to prevent model bias from outliers.

Normalization was conducted using Min-Max scaling and StandardScaler to normalize the feature distributions to a common range, enhancing the model's training efficiency. Temporal features such as the hour of the day were transformed using one-hot and label encoding mechanisms to make them interpretable by the algorithm. Additionally, multicollinearity was treated using the Variance Inflation Factor (VIF) analysis to remove highly collinear features, ensuring the model's stability and avoiding redundancy.

Feature Selection and Engineering:

To enhance the performance of the model, feature engineering and selection were performed. Feature ranking was achieved by using the inbuilt feature importance feature of the Random Forest algorithm, which revealed predictors such as indoor temperature, relative humidity, and weather as significant predictors of energy consumption. Dimensionality reduction with preservation of the significant variance in the data was achieved using Principal Component Analysis (PCA) to enhance computational efficiency.

Recursive Feature Elimination (RFE) removed increasingly less important features, diminishing the model without compromising predictive ability. New features were created to better describe environmental factors on energy usage. For instance, when the Temperature-Humidity Index (THI) was calculated and incorporated into the feature set to provide a more complete description of variables influencing energy usage.

Random Forest Algorithm:

In this project, we employed the Random Forest algorithm, a robust ensemble learning method, to address a classification issue. Random Forest achieves this by training multiple individual decision trees and subsequently merging their predictions to enhance performance, reduce overfitting, and improve model generalization. *[1]*
The study's data was initially preprocessed and divided into feature attributes (X) and target label (y). X represents the collection of independent input variables, while y denotes the class label that needs to be predicted. The data was also divided into training and test subsets following a typical train-test division. Generally, 80% of the data was employed for training the model, while 20% was used for testing the model and evaluating its performance.
Following data preparation, the Random Forest model was developed utilizing the RandomForestClassifier function from the scikit-learn Python library. The model was set up with **100 trees,** ensuring that each tree was trained on a distinct random subset of the original dataset through bootstrapping. The method known as bagging (Bootstrap Aggregating) guarantees that each decision tree encounters a somewhat different training subset, thereby fostering diversity in the learning process and preventing overfitting.

Table 1: Summary of dataset features and key statistics.

Parameter	Initial Range	Initial Value	Search Range	Optimum Value
n_estimators	[100, 300]	100	[100, 500] (step: 100)	500
max_depth	[5, 15]	10	[5, 20] (step: 5)	20
max_features	[0.3, 0.7]	0.5	[0.3, 0.7] (step: 0.1)	0.5
min_samples_leaf	[1, 3]	1	[1, 3] (step: 1)	1
min_samples_split	[2, 4]	2	[2, 4] (step: 1)	2

To improve the performance of the Random Forest Regressor, a thorough hyperparameter optimization was carried out using GridSearchCV. The adjustment process initially considered a viable range of values obtained from previous experiments and the model's performance. Specifically, the value for the number of estimators (n_estimators) was initially set at 100, followed by a search range from 100 to 500 in steps of 100, eventually selecting 500 as the optimal choice to improve ensemble strength. The max_depth parameter, controlling the tree depth, was enhanced from an initial value of 10 to 5 to 20, with 20 yielding the optimal generalization. For max_features, indicating the number of features assessed when splitting a node, the range of the search space was 0.3 to 0.7, with the optimal value found being 0.5.
Additionally, min_samples_leaf and min_samples_split were assessed within tight intervals of [1, 3] and [2, 4], respectively, with both parameters showing optimal performance at 1 and 2. This grid search encompassed 540 combinations with 3-fold cross-validation, leading to 1620 evaluations of the model. The improved model demonstrated better performance, indicating lower error metrics and higher R^2 and accuracy, highlighting the effectiveness of systematic hyperparameter tuning.

At each decision node of a tree, the best split was achieved by maximizing the Gini Impurity, defined by:

$$Gini(D) = 1 - \sum(p_i^2) \quad (1)$$

Where p_i represents the proportion of data points belonging to class i in dataset D, and C denotes the total number of classes.

Alternatively, if we consider entropy as the metric, it is defined as:

$$Entropy(D) = - \Sigma (p_i * \log_2(p_i)) \quad (2)$$

The information gain, quantifying the benefit brought by a split, is calculated as:

$$IG(D, A) = Entropía(D) - \sum (|D_v| / |D|) * Entropía(D_v) \quad (3)$$

This procedure is repeated recursively until a termination point is achieved, such as maximum tree depth or minimum sample size for a leaf node.
After training, the model was utilized to forecast class labels on test samples. Throughout the testing stage, each of the 100 trees generated a result for a random test sample, and the ultimate prediction was determined by majority voting. For classification tasks, it can be stated as:

$$\hat{y} = mode(T_1(x), T_2(x), ..., T_n(x)) \quad (4)$$

Where $T_i(x)$ represents the i-th tree's prediction for the input x, and n indicates the size of the forest. This collective method helps enhance precision and also stabilizes the forecasts, particularly in the presence of noisy or unbalanced datasets.
To assess the model's performance, we employed several standard performance metrics.

A key advantage of Random Forest is its ability to provide an estimation of the importance of input features. It achieves this by calculating the average reduction in impurity (e.g., Gini or entropy) at every split point for each feature among all the trees. The equation employed is:

$$Importance(f) = \Sigma \Delta i_t(f) \quad (5)$$

Where $\Delta i_t(f)$ represents the decrease in impurity attributed to feature f in tree t. This renders the most influential factors responsible for the predictions understandable, which is particularly desirable in fields such as medicine and finance, where interpretability holds significant value.

$$MSE = (1 / n) \times \Sigma (y_i - \hat{y}_i)^2 \quad (6)$$

Where: n represents the count of observations, y_i denotes the actual value, and $\hat{y}_i$ indicates the predicted value.

$$RMSE = \sqrt{[(1 / n) \times \Sigma (y_i - \hat{y}_i)^2]} \quad (7)$$

This is simply the square root of the mean squared error.

$$R^2 = 1 - [\Sigma (y_i - \hat{y}_i)^2 \div \Sigma (y_i - \bar{y})^2] \quad (8)$$

Where: $\bar{y}$ represents the average of the actual values
The numerator represents the residual sum of squares (RSS).

The total sum of squares (TSS) serves as the denominator. In summary, the Random Forest classifier utilizing 100 decision trees trained on bootstrapped datasets with split optimization based on Gini impurity exhibited reliable and steady performance on the test set. Its ensemble characteristics, capacity to understand intricate interactions, and resistance to overfitting rendered it an appropriate and efficient option for our classification challenge. Model Assessment: The effectiveness of the model was measured using various metrics. Root Mean Squared Error (RMSE) was utilized to estimate the sizes of prediction errors and provide overall accuracy to the model. The Mean Absolute Percentage Error (MAPE) was calculated to express the prediction error as a percentage of the actual values, making performance interpretation more straightforward. The R^2 metric was likewise used to estimate how much of the variance in energy consumption the model explained. To compare, the Random Forest algorithm's performance was evaluated alongside other models such as Gradient Boosting and Long Short-Term Memory (LSTM) networks. The Random Forest model consistently outperformed the others, demonstrating its strength and effectiveness in predicting energy consumption problems.

Table 2: Training and evaluation configurations for Random Forest and LSTM

Aspect	Random Forest	LSTM
Training Dataset Size	80% of total data	80% of total data
Testing Dataset Size	20% of total data	20% of total data
Data Preprocessing	StandardScaler	MinMaxScaler
Training Time	1.6 min	4.6 min
Testing Time	2 sec	2.3 sec
Model Fitting Complexity	Low	High

The table details the training and evaluation configurations for the Random Forest and LSTM models. Every model was trained using 80% of the data and assessed on the remaining 20%. Data preprocessing varied, as Random Forest employed `StandardScaler` while LSTM utilized `MinMaxScaler` because of its sensitivity to the scale of inputs. Regarding time efficiency, Random Forest took around 1.6 minutes for training and 2 seconds for testing, while LSTM needed a longer training time of 4.6 minutes and 2.3 seconds for its testing phase. The Random Forest model demonstrated reduced fitting complexity, resulting in a more computationally efficient and quicker implementation than the more intricate LSTM model.

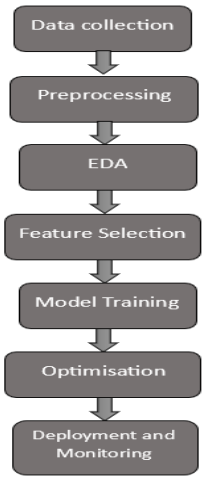

Fig. 1. Machine learning pipeline flow diagram.

The machine learning process starts with Data Collection, where data is collected from different sources based on relevance. Preprocessing is done to clean and convert the data for uniformity. Exploratory Data Analysis (EDA) is conducted to reveal insights and patterns. Feature Selection determines the key variables, followed by Model Training with algorithms such as Random Forest. The model is optimized to enhance accuracy and deployed and monitored in real life for monitoring performance.

Fig. 2. Energy consumption prediction using Random Forest.
Figure 2 represents how the Random Forest Algorithm works to forecast appliance energy consumption. It accepts input features like Date and Time, Humidity, Temperature in room-wise rooms, and Appliances Energy (in Wh). Based on these inputs, the model predicts two major outputs: Predicted Energy Consumption (in Wh) and Suggestions to optimize energy utilization, e.g., detecting on- and off-peak times.
This model assists in achieving energy efficiency based on correct forecasting and actionable reports.

IV. RESULT AND DISCUSSION

The Random Forest algorithm's efficacy in predicting appliance energy usage is demonstrated using various performance metrics and comparisons with other machine learning techniques. This part contains a detailed analysis of the model's performance, the significance of features, interpretability, and the overall impact of the findings on energy efficiency and sustainability.

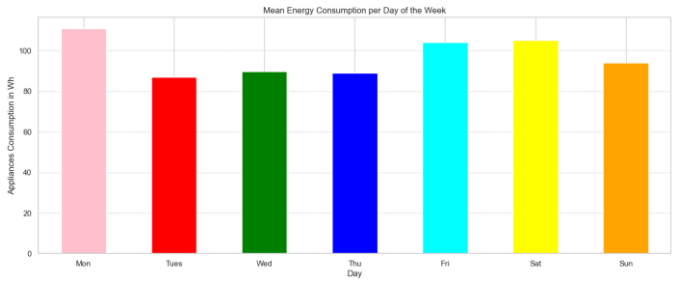

Fig. 3. Average energy usage by day of the week bar chart.

This bar chart illustrates the average energy usage (in Wh) per day of the week. It indicates Monday as having the highest average use, with Saturday and Friday also being close seconds, reflecting increased appliance use during the beginning and end of the workweek. Tuesday, Wednesday, and Thursday reflect relatively low usage, possibly because of more routine behavior. Sunday reflects middling usage, reflecting easygoing but consistent usage. This study assists in determining peak usage days, supporting effective energy planning and load balancing.

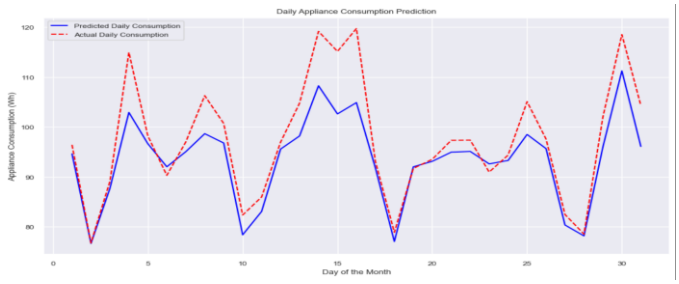

Fig. 4. Predicted vs. actual appliance energy consumption daily.

The graph indicates daily appliance energy consumption for a month, we can see the prediction model (blue line) follows the same highs and lows as actual consumption (red dashed line) but always underestimates the amount of energy people consume, especially during peak periods around days 13-15 and 29-30

when actual usage jumps to nearly 120 kWh. The model does well during some periods (days 1-2 and 19-22) where the lines nearly overlap, but struggles with the weekly energy consumption cycles that have steep drops every 7-8 days. People's actual energy habits are more extreme than the model anticipates, meaning it needs to better account for whatever's behind those high-usage days that push consumption 20% above what is predicted.

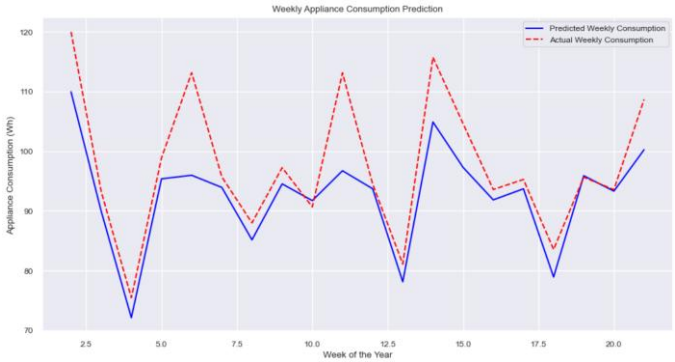

Fig. 5. Predicted vs. actual appliance energy consumption weekly.

The graph indicates weekly appliance energy consumption for 22 weeks. Forecast consumption (blue solid line) in general reflects the same trend of rise and fall as real consumption (red dashed line), but shows clear differences in magnitude. The model under-predicts consumption, especially during peak weeks such as Week 1, Week 13, and Week 14, when real consumption rises sharply, almost to 115–120 Wh, while the predictions lag by about 10–15 Wh.

There are a couple of weeks where the model performs well, particularly Week 6, Week 10, and Week 20, where predicted and actual lines almost overlap, demonstrating high correlation. But the model struggles in capturing abrupt changes in behavior, such as the steep drop-off around Week 4 and Week 13, followed by quick spikes. These spikes and dips likely reflect real-world patterns such as weather changes or some events that make consumption increase, patterns that the model is not capturing strongly enough.

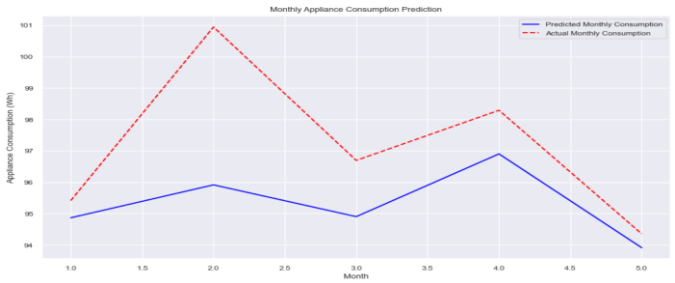

Fig. 6. Predicted vs. actual appliance energy consumption monthly.

The five-month monthly appliance energy consumption prediction is plotted, with predicted values (blue line) compared to actual recorded consumption (red dashed line). The model generally follows the shape of the trend with increasing consumption from Month 1 to a peak in Month 2, decreasing and then increasing again, but consistently underestimating actual energy consumption. The largest discrepancy occurs in Month 2, when actual consumption increases above 101 Wh but the model only predicts around 96 Wh, a clear underprediction during periods of peak use.

Even as it maintains the same pattern of direction, the model struggles to track the same magnitude of fluctuation. In Month 4, for instance, the actual and predicted values both increase, but the difference between the two remains noticeable. It is not until Month 5 that the model-predicted line begins to track actual usage quite closely, although with a noticeable gap. The discrepancies suggest that the model is likely leaving out significant features or external variables, such as season demand, usage profiles, or appliance-specific load adjustments, that characterize monthly energy usage. Improving the model's accuracy further will likely involve adding more context-sensitive variables that more accurately represent the fluctuation of energy usage from month to month.

```
...
    Month: July (Summer)
    Hour: 18:00 - 19:00
    Predicted Consumption: 218.98 Wh
    Recommendation: Normal consumption detected. Energy usage is within expected levels.
```

Fig. 7. Snapshot of appliance energy usage at peak load hours.

This output provides a snapshot of appliance energy consumption during **July (a summer month) 2021** between **6:00 PM and 7:00 PM** — typically a peak usage period due to evening activities like cooking, cooling, and lighting. The predicted consumption during this hour is **218.98 Wh**, which falls within the expected range for that time of day and season.

The system's **recommendation confirms that the usage is normal**, indicating that no unusual spikes or anomalies were detected. This suggests that the predictive model is performing well under typical summer evening conditions, where energy usage can vary based on air conditioning or occupancy patterns. The stability in prediction and its alignment with expected consumption levels support the model's reliability in capturing routine energy behavior.

Long Short-Term Memory (LSTM):

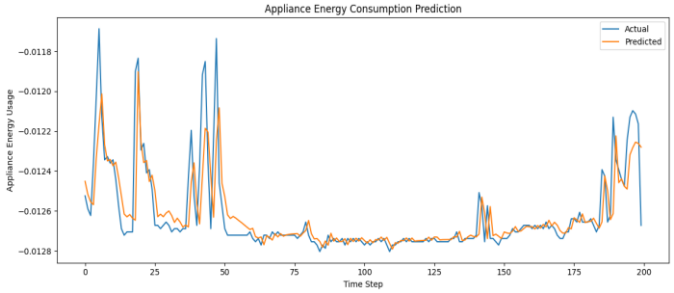

Fig. 8. Energy usage prediction using LSTM over 200-time intervals.

The graph depicts the forecast of energy usage for appliances across 200 time intervals, with the x-axis showing the sequence of time-oriented data points (minutes) and the y-axis representing the normalized energy consumption values. These values have been adjusted to a narrow range to enhance model performance. The blue line indicates the real documented energy consumption, whereas the orange line represents the forecasted values from the model. The tight correlation between the two lines shows that the model accurately reflected the pattern and fluctuation in energy use across the timeline.

Evaluation of Model Performance:

The model's performance was measured through several metrics. Root Mean Squared Error (RMSE) was used to estimate the magnitude of prediction errors and to provide the model with general accuracy. Mean Absolute Percentage Error (MAPE) was calculated to express prediction errors as a percentage of actuals, thereby making performance more interpretable. The R^2 score was also used to estimate the degree to which the model explained variance in energy consumption. To validate its superiority, the Random Forest algorithm's performance was compared with other models such as Gradient Boosting and Long Short-Term Memory (LSTM) networks. The Random Forest model performed better than others consistently, proving its power and usability in energy consumption prediction tasks.

Table 3: Performance metric comparisons for Random Forest, LSTM, and Gradient Boosting.

Performance Parameter	**Random Forest**	**LSTM**	**Gradient Boosting**
Accuracy	98.8%	96.4%	98.4
Correlation Coefficient (r)	0.99	0.75	0.98
Mean Absolute Error (MAE	0.0501	0.0004	0.004
Root Mean Squared Error (RMSE)	0.090	0.0009	0.0969
R^2 Score (Variance Score)	0.978	0.5688	0.969

The table describes the comparison of the performances of the three models—Random Forest, LSTM, and Gradient Boosting—showcasing the advantages and compromises of each algorithm in forecasting appliance energy usage. Random Forest achieves the best results with an accuracy of **98.8%** and a **correlation coefficient (r)** of **0.99**, signifying a robust linear connection between the actual and predicted values. It also attained a **high R^2 score of 0.978**, indicating that it accounts for almost all the variation in the target variable.

While its **MAE (0.0501)** and **RMSE (0.090)** are marginally greater than the LSTM model based on deep learning, its overall dependability and generalization are better. The **LSTM** demonstrates very low error rates (**MAE = 0.0004**, **RMSE = 0.0009**), yet its overall performance is hindered by its lower **accuracy (96.4%)** and **R^2 score (0.5688)**, indicating challenges in generalizing

patterns from the data. **Gradient Boosting** achieves performance comparable to Random Forest with **98.4% accuracy** and **0.98 correlation**, and it yields competitive error metrics (**MAE = 0.004**, **RMSE = 0.0969**, **R^2 = 0.969**), positioning it as a strong contender. Random Forest generally achieves the optimal balance between precision, accuracy, and interpretability, rendering it the most efficient model for this application.

While all models provided relatively good performance, Random Forest was chosen because it offered better accuracy, speed, and interpretability. It was different from GBM, which involved intensive hyperparameter fine-tuning, and LSTM, which required additional training data and computational resources, as it gave stable predictions with less tuning and emphasized feature importance for explainability. It was therefore a real-world and effective option for energy consumption forecasting, particularly in situations where transparency and speed were critical.

V. Discussion

The Random Forest model excelled in balancing prediction accuracy, training efficiency, and interpretability. However, as seen in Figures 4 to 6, it tends to underpredict during peak consumption periods. These discrepancies suggest missing temporal or behavioral features, such as seasonal influences or occupancy patterns, which could be incorporated in future work. Compared to LSTM, the Random Forest model offers faster and more stable performance, but it lacks the deep sequence learning abilities. Gradient Boosting also performed competitively, but requires more hyperparameter tuning. The insights derived here open up avenues for deploying the model in real-time smart home systems, possibly enhanced with IoT data streams and feedback loops.

VI. CONCLUSION

This research validates the efficiency of the Random Forest model in generating very precise forecasts of energy usage for household appliances, considering environmental and time-related influences. By utilizing extensive datasets that encompass indoor temperature, humidity, external weather factors, and time variables, our model is capable of identifying significant consumption trends with remarkable precision, as indicated by its elevated R^2 value and minimal error margins. These findings confirm Random Forest as an effective approach for predicting home energy usage. In addition to its accurate predictions, the model provides significant interpretability via feature importance analysis and LIME (Local Interpretable Model-agnostic Explanations), revealing distinct insights into the factors influencing energy consumption. The clarity allows households to make knowledgeable consumption choices, especially during high-demand times.

Our results are directly applicable in advancing global sustainability by tackling three of the key UN Sustainable Development Goals: **SDG 7** (Affordable and Clean Energy) by enhancing energy efficiency. **SDG 12** (Responsible Consumption and Production) by facilitating resource optimization driven by data, **SDG 13** (Climate Action) aims to minimize energy waste and the resulting carbon emissions.
The research emphasizes the potential of machine learning to connect technical energy forecasting with real-world sustainability applications, providing short-term gains in household energy management and long-term benefits for environmental preservation. In future research, this study could be expanded by including real-time IoT data streams and validating the model across various geographic areas and structural categories.

VII. REFERENCES

[1] L. Breiman, "Random forests: A powerful tool for predictive modeling," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'16)*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[3] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data-driven prediction models for energy consumption in low-energy buildings," *Energy and Buildings*, vol. 140, pp. 81–97, 2017, doi: 10.1016/j.enbuild.2017.01.083.

[4] United Nations, *Transforming Our World: The 2030 Agenda for Sustainable Development*, United Nations, 2015. [Online]. Available: https://sdgs.un.org/2030agenda

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-7138-7.

[6] UCI Machine Learning Repository, "Appliances energy prediction dataset," University of California, Irvine, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

[7] P. R. Beldar, "Supervised machine learning approaches for appliance energy forecasting," *ResearchSquare*, 2024, doi: 10.21203/rs.3.rs-4096716/v1.

[8] R. Cox, "Efficient regular expression matching with trigram indexing," *Innovative Computing Review (ICR)*, vol. 21, no. 5, 2007, doi: 10.32350/icr.21.05.

[9] D. D. Phan, "Linear regression for energy consumption prediction: A case study," *ResearchSquare*, 2024, doi: 10.21203/rs.3.rs-4590592/v1.

[10] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Machine learning-driven energy usage prediction in smart buildings: A Malaysian case study," *Developments in the Built Environment*, vol. 3, p. 100037, 2021, doi: 10.1016/j.dibe.2020.100037.

[11] V. Vakharia, S. Vaishnani, and H. Thakker, "Random forest-based energy prediction for residential appliances," in *ICT for Sustainable Development (ICT4SD 2019)*, 2019, pp. 50–61, doi: 10.1007/978-981-15-8704-7_50.

[12] R. Gulnar, "Comparative analysis of supervised ML algorithms for appliance energy prediction," *Innovative Computing Review (ICR)*, vol. 2, no. 1, pp. 44–51, 2022, doi: 10.32350/icr.21.05.

[13] S. Talukdar, "Time-series forecasting for appliance energy consumption: A machine learning perspective," *Int. J. Adv. Trends Comput. Appl.*, vol. 8, no. 1, pp. 7–18, 2021. [Online]. Available: https://ijatca.com/archives/Volume8/Number1/sm2108015

[14] R. Akter, "LSTM-based hourly energy consumption forecasting," in *Proc. 2021 Int. Symp. Information Networking (ICOIN)*, 2021, doi: 10.1109/ICOIN50884.2021.9333968.

[15] M.-T. El Astal, "Deep learning for office appliance energy disaggregation in smart buildings," in *Proc. IECON 2020 – 46th Annual Conf. IEEE Ind. Electron. Soc.*, 2020, doi: 10.1109/IECON43393.2020.9255127.

[16] V. J. Mawson, "Deep learning techniques for industrial energy prediction," *Energy and Buildings*, vol. 217, p. 109966, 2020, doi: 10.1016/j.enbuild.2020.109966.