Time series forecasting on hourly energy consumption

1st Hogan Lin
M.S. Data Science
Stevens Institute of Technology
Hoboken, NJ, US
hlin31@stevens.edu

2nd Yian Chen
M.S. Applied Artificial Intelligence
Stevens Institute of Technology
Hoboken, NJ, US
ychen17@stevens.edu

3rd Pranay Reddy Baireddy
M.S. Applied Artificial Intelligence
Stevens Institute of Technology
Hoboken, NJ, US
pbairedd@stevens.edu

***Abstract*—Due to present day technological advances, energy consumption is one of the most critical issues on our environmental and economic impact. Optimizing energy usage becomes an important task to effectively allocate these resources and save cost. It is therefore crucial to have accurate energy consumption forecasting for such effective resource management and demand planning. This study leverages machine learning methods, including Random Forest, XGBoost, and LSTM, to predict hourly energy consumption patterns using data from the PJM regional transmission organization. Advanced feature engineering techniques, such as sine-cosine transformations for cyclical data and feature selection, were employed to enhance model performance. The findings highlight the importance of tailored feature engineering and model optimization for robust energy forecasting. Future work includes integrating hybrid models and contextual variables to refine predictions further.**

I. Introduction

Accurate energy consumption prediction is vital for effective power system planning, demand management, and sustainable resource allocation. Electricity is expected to replace other energy sources as one of the primary energy source for home, businesses, and transportation in the near future [9]. The demand for electricity in the United States in 2022 was about 4.07 trillion kWh, which is 14 times greater than electricity use in 1950 [11]. Energy consumption is expected to double by reach 360 TWh, and the demand for electricity will increase to 120 GW per year [2]. Efficient electricity usage is therefore critical for conserving resources, and this requires effective resource planning. Optimal energy consumption forecasting ensures grid stability, efficient resource allocation, and cost reduction. As demand for energy fluctuates due to various factors, such as time of day, season, and operational schedules, advanced prediction methods are necessary to ensure accuracy of the predictions. This study investigates various machine learning methods to analyze historical hourly energy consumption data from PJM regional transmission organization (RTO) in the United States that coordinates movement of wholesale electricity across states. Our study focuses on both overall trends and specific influences like holiday and seasonal effects on usage patterns. By utilizing a comprehensive dataset, we gain insights into how energy consumption varies over time and under different conditions. Our models allow us to make future forecast based on historical data for any future business or resource-allocation planning needs. Through this approach, our goal is to develop models that not only yield high accuracy but also offer valuable insights to inform future energy demand forecasts and promote efficient energy resource management.

II. Related Work

Previous research on energy consumption prediction has primarily focused on time series analysis and machine learning techniques, each providing unique benefits depending on the data's characteristics and application context [5], [7]. Traditional time series models like Autoregressive Integrated Moving Average (ARIMA) are widely used for forecasting because of their capacity to capture linear dependencies and temporal patterns [10]. However, these models often struggle to model the complex, non-linear trends commonly observed in real-world energy datasets, where machine learning methods such as neural networks and support vector machines have shown superior performance in capturing these dynamics [1], [4].

The ARIMA and Holt-Winters model on time series analysis are also compared in Aurna et al. [3] to predict the energy consumption data in Ohio and Kentucky. Their studies reveal that the Holt-Winters model is more effective for long-term forecasting.

Machine learning methods, on the other hand, introduce greater flexibility by enabling models to learn intricate relationships and non-linear patterns within the data. Techniques such as neural networks, including Long Short-Term Memory (LSTM) networks, and decision trees have shown promise in energy forecasting tasks, particularly for datasets with substantial temporal and non-linear dependencies. LSTMs are well-suited for sequence data, as they excel at capturing temporal dependencies over extended periods. Decision tree-based methods, such as Random Forests and Gradient Boosting Machines (GBMs), offer robust performance across various data types, adapting well to skewed and imbalanced datasets often encountered in energy usage scenarios.

In recent studies, researchers have explored hybrid models that integrate time series analysis with machine learning to leverage the strengths of both approaches for improved predictive accuracy [7], [4]. For instance, certain studies preprocess data using time series decomposition methods before

applying machine learning algorithms, enabling models to more effectively capture seasonal and trend components [10]. Feature engineering and normalization techniques are also widely applied to enhance model performance, particularly in the context of energy data where skewed distributions are common [6].

Elhabyb et al. (2024)[8] discusses the use of supervised machine learning methods, such as Random Forest and Gradient Boosting Regressor (GBR), to model energy usage in educational buildings. The studies from [8] shows that the GBR method performed the best overall on most buildings, and that it is worth exploring hybrid or ensemble methods. Similar to their approach, our project also train the Random Forest, Gradient Boosting Regressor, and RNN model on LSTM algorithm for energy consumption predictions, where the dataset from the latest year as our ground-truth consumption levels.

Our work builds upon existing research by training and analysing the performance of similar machine learning methods, while also incorporating feature engineering and external features such as holiday, seasonal effects, and multivariate analysis into the prediction process. Through these approaches, we aim to contribute a more comprehensive understanding of energy usage patterns and enhance predictive accuracy in diverse consumption contexts.

III. Our Solution

A. Description of Dataset

The dataset, sourced from Kaggle[1], provides comprehensive hourly energy consumption data from PJM Interconnection LLC (PJM), a regional transmission organization (RTO) that coordinates electricity transmission across multiple U.S. states, including Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia. This organization operates as part of the Eastern Interconnection grid, managing power distribution across a significant portion of the Eastern United States.

The dataset includes consumption data in megawatts (MW) for several prominent energy regions, such as AEP, COMED, and DAYTON, each representing different service areas. Each file in the dataset represents hourly records, spanning various time periods depending on the specific region. This variance occurs due to the historical changes in regions managed by PJM, meaning that certain areas may have data coverage only for specific dates. To provide a clear understanding of the dataset's characteristics, we generate statistical summaries, including figures and tables displaying the mean, median, variance, and skewness of energy consumption across different regions and time periods. Skewness analysis reveals a right-skewed distribution (Figure 1), indicating the presence of extreme outliers where energy consumption spikes significantly on certain dates. Additionally, Figure 2 illustrates the wide range of outliers across different energy companies, reflecting the expected high fluctuations in collective energy consumption due to various factors. These insights emphasize the need for robust models capable of handling variability and evaluation metrics sensitive to large deviations caused by such outliers.

[1]https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption

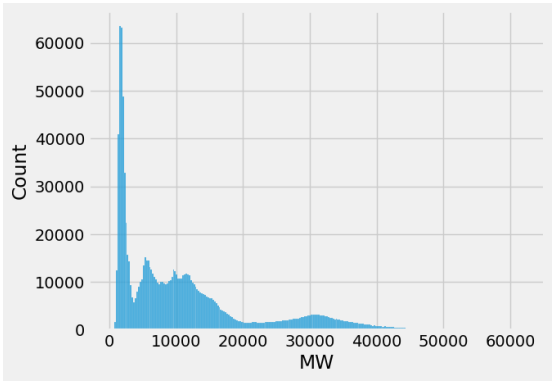

Fig. 1. Distribution of aggregate energy consumption data. A right skewness can be observed.

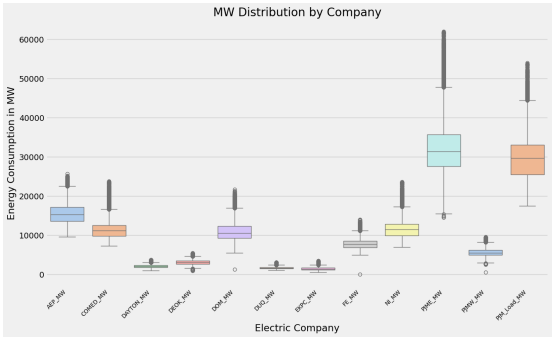

Fig. 2. Significant outliers can be observed across the energy consumption data across varying companies.

Data Preprocessing To prepare the dataset for analysis, we performed several preprocessing steps. First, we merged all the datasets to create a unified dataset encompassing all regions. While there are overlapping time periods of the data across the datasets of different companies, the start and end dates of the data are different for each company. To address this, we synchronized the dataset by including data from the earliest start date up to one year before the latest end date. This approach ensures the training data captures a long historical record while excluding the most recent year. The data points from the cut-off date (one year before the latest end date) to the most recent date were reserved for testing, allowing our model to be evaluated on the most up-to-date data.

We then standardized column names for consistency and identified missing data points, which we addressed through appropriate imputation methods where feasible. This cleaned dataset allowed for more robust analysis, helping us to accurately evaluate consumption patterns, detect anomalies, and assess the effect of external factors, such as holidays, on energy usage across these regions.

Feature Engineering For feature engineering we came up with some hypotheses that are common to time series trends. The first hypothesis was that there was more energy consumption on holiday days. We utilized the holidays library to examine the correlation between US holidays and energy consumption usage, but the correlation was -0.01, indicating virtually no linear relationship. We then checked the average daily use of energy on major US winter holidays like Christmas and Thanksgiving, and also found that there is little to no correlation as shown in Figure 3 . Based on these findings, we conclude that hypothesis 1 is False.

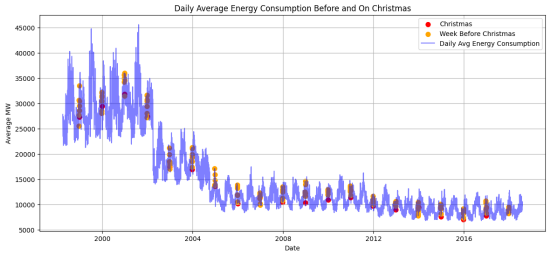

Fig. 3. Daily average energy consumption of all companies a week leading up to Christmas and on Christmas; we observe there is very low correlation.

Next we took a look at correlation between seasons and energy consumption. We propose hypothesis 2, in which seasons affect energy consumption and seasons with lower or higher temperatures would have higher energy needs. We observe a general pattern where Winter and Summer seasons have higher energy expenditure (Figure 4); therefore hypothesis 2 is true. Even though holidays have low correlation, we added it to our features since we are uncertain of the relevance from correlation alone. We also added the four seasons as a feature to our dataset.

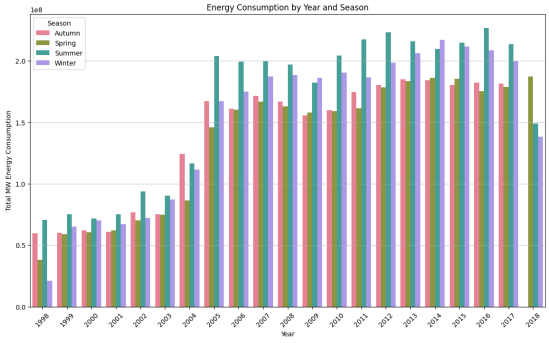

Fig. 4. Average Daily Energy Consumption based on Seasons. We observe that Winter and Summer seasons have higher usages.

Since time-series data is usually cyclical (as observed by the sinusoidal patterns of the data), we transformed the features to cyclical features using sine and cosine encoding for our models to better learn the cyclical patterns. We apply the sine and cosine transformations on our features using the formula below:

$$x_{\sin} = \sin\left(\frac{2 \cdot \pi \cdot x}{\max(x)}\right)$$
$$x_{\cos} = \cos\left(\frac{2 \cdot \pi \cdot x}{\max(x)}\right)$$

This essentially projects our data points onto a circle, so the models can learn the cyclical natures of the data.

This dataset serves as a foundation for our study, providing detailed temporal granularity and allowing us to apply machine learning models to understand and predict hourly energy consumption patterns across a diverse set of U.S. regions.

B. Machine Learning Algorithms

For the first part of our machine learning technique we use Random Forest to predict hourly energy consumption. Random Forest is well-suited for time series forecasting as it captures complex, non-linear relationships between input features and the target variable. Unlike traditional time series models, Random Forest assumes no specific data distribution, making it highly flexible. It is also a model that mitigates overfitting due to its random feature selection at the decision tree splits. This randomness ensures that the model does not rely too heavily on specific features. In addition, is uses bootstrapping, where each tree is trained on randomly sampled subset of the training data. Such variation ensures trees are not identical and helps them generalize better to new data. For this task, we use a set of engineered features such as the hour of the day, day of the week, and month, along with the historical energy consumption data and the holiday data to train the model. The model is trained on a training set and validated on a separate test set to assess performance. Random Forest helps us capture both seasonal patterns and abrupt changes in the data, providing accurate and robust predictions for hourly energy consumption across multiple regions. Some of challenges we faced with the use of Random Forest include computational efficiency and intensity, especially when we try to do a grid search or random search on the model to fine-tune it with different hyperparameters.

Our second model employs the Extreme Gradient Boosting (XGBoost) model for time prediction. XGBoost is a commonly-chosen model for time-series prediction, particularly in energy usage. The model can handle complex feature interactions and has built-in regularization to mitigate model overfitting, which works differently from the Random Forest. We decided on XGBoost as the model also supports sophisticated feature engineering, which helps improve the accuracy of our forecasts. Finally, it can complement other machine learning models, making it well-suited for future ensemble learning approaches.

For our third model we apply neural networks (NN) and traditional time series models to predict hourly energy consumption, leveraging both the flexibility of neural networks in capturing complex patterns and the structured approach of time series analysis. We chose Recurrent Neural Networks (RNN) because they are designed specifically for sequential data,

making them well-suited for time series forecasting where past data influences future predictions. RNNs can learn patterns in the sequential order of the data (e.g. trends, cycles, seasonality) by processing input sequences step by step. However, standard RNNs suffer from limitations such as the vanishing gradient problem, which prevents them from capturing long-term dependencies effectively. To address vanishing gradient problem, we employed the Long Short-Term Memory (LSTM) model. The LSTM is a type of RNN that can capture long-term dependencies in time series data, which is important for patterns spread over long time intervals.

Our neural network architecture includes an input layer, three hidden layers, and an output layer designed for regression. The hidden layers employ ReLU (Rectified Linear Unit) activation functions, which improve model convergence by addressing vanishing gradient issues often seen in deep networks. To further optimize the model, we implement dropout regularization between hidden layers, reducing the likelihood of overfitting by randomly deactivating a subset of neurons during training. Hyperparameters, such as learning rate and batch size, are critical to model performance. We begin with initial values informed by previous studies on similar data but fine-tune these parameters during training using grid search and cross-validation techniques. The learning rate dictates how quickly the model updates weights in response to errors, while batch size affects memory usage and training speed. We conduct training in mini-batches, balancing computational efficiency and gradient stability, and adaptively adjust the learning rate to enhance convergence.

For our Random Forest Model, we fine-tuned the hyperparameters using random search and tested out different combinations of hyperparameters, including maximum features, maximum depth of tree, number of trees, and minimum sample leaf using three fold cross-validation. We also tested different parameters for the XGBoost model, which shows improvement in model performance. For our LSTM model, we fine-tuned the hyperparmeters by testing different number of layers, changing dropout rates, number of neuron units in the layers, as well as the learning rate.

By combining these machine learning approaches, we aim to achieve robust, accurate predictions that account for both complex, non-linear interactions in the data and time-dependent patterns that traditional statistical models excel in capturing. This dual approach allows us to evaluate the effectiveness of different predictive methodologies in forecasting energy consumption at an hourly level across multiple regions.

C. Implementation Details

Previously we established that we transformed our features using sine and cosine encoding, which essentially normalizes the data values between -1 to 1. This cyclical encoding not only helps the model better learn time-series patterns, but the normalization also helps improve model accuracy and convergence rates by reducing bias toward larger numerical values. This normalization ensures that all variables contribute equally to the training process, particularly important when analyzing time-series data with large fluctuations in power consumption.

Our combined datasets include energy consumption data of different regions of PJM over the years. Since the date ranges of different regions are inconsistent, we find individual cut-off points for each company. For our training data, we used data points from the earliest data up to one year before the last date. For the test set we use the most recent data points from the cut-off date, or one year before the latest date. This way we ensure we are using historical data for the training sets and the data points of the most recent dates for our test set to evaluate model predictions. We built a custom pre-processor and applied our cylical encoding to our data, as well as one-hot encoded our holiday features (holiday v.s. not holiday), and numerically encoded our season features to ordinal values.

Throughout the project, we document our findings in both visual and tabular formats, including model evaluation metrics and comparisons across different algorithms. This structured approach provides a comprehensive view of the dataset, model implementations, and insights derived from holiday impact analysis. The detailed tracking of data processing steps, model adjustments, and final performance results ensures reproducibility and clarity in our workflow.

D. Evaluation

For our evaluation we use the R-squared (R^2), which is a statistical measure of how well the regression model fits the data. It explains the proportion of variance in the target variable that is predicted from the features. The formula for (R^2) is:

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

In addition, we also used the Root Mean Squared Error (RMSE), which measures the average magnitude of the residuals, providing insight into how well the model predicts the actual values. Since RMSE is expressed in the same units as the target variable (the megawatts MW energy consumption), it makes it easier to interpret the error magnitude in the context of energy consumption. RMSE also penalizes large errors because of its squaring term, which could be useful for energy consumption forecasting as large errors could indicate significant deviations from the true demand, which could affect our grid search optimization. In addition, we saw from our data exploration that energy consumption data often have fluctuations or outliers, and RMSE's sensitivity to large deviations helps in identifying and addressing these outliers. The RMSE is defined as square root of the average of the squared differences between predicted values and observed values. The formula is:

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

For energy consumption problems measured in megawatts (MW), the Mean Absolute Error (MAE) is also a useful metric

to measure the average magnitude of errors between actual and predicted values, it is expressed as:

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

Finally, we also use Mean Absolute Percentage Error (MAPE) as our evaluation metric. MAPE measures the average percentage error between actual and predicted values, making it easier to interpret as a relative error. MAPE reflects how much the predicted usage deviates from actual percentages, so the lower the MAPE, the better the model is. It is worth noting that MAPE is sensitive to small values of y, which can inflate the error. MAPE is calculated as:

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

The addition of MAE and MAPE allows us to evaluate both the absolute and relative performance of our model, providing a more comprehensive understanding of its accuracy. One limitation of using MAPE is that when the actual values have instances of zero, it would lead to division by zero. As the actual values approach zero, the percentage errors can become extremely large even if the forecasted values are close to the actual values. This can distort the MAPE, and when there are actual values that are 0, it would lead to division by zero. We noted that when using MAPE as a evaluation metric on our dataset, we would get infinity in our evaluation result. This is because there are instances of zeros in our actual values. To address this, several options for similar metric were weighed. Some common metrics include WMAPE (Weighted Mean Absolute Percentage Error), SMAPE(Symmetric Mean Absolute Percentage Error), MASE (Mean Absolute Scaled Error), or WAPE(Weighted Absolute Percentage Error). Each have different advantages and drawbacks, which will not be discussed in this paper as it is out of scope of this project. Our evaluation uses the Weighted Absolute Percentage Error (WAPE) because we want a general, unbiased view of forecast accuracy across the entire dataset without giving specific priority or weight assigned to certain time periods, companies, or seasons. WAPE is robust to low actual values (y) as compared to MAPE. WAPE weights the error by adding the total energy consumption, and is calculated as:

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

1) Random Forest Regressor: The evaluation results, summarized in Table I, indicate strong performance from the initial Random Forest model. The R-squared scores are 0.9788 for the training set and 0.9621 for the test set, suggesting a high level of predictive accuracy. The RMSE values of 1537.14 (training set) and 1987.45 (test set) further confirm the model's effectiveness in capturing the underlying patterns of the data. The MAE on the test set was 1130.46. The MAPE metric was 10.66%, and WAPE on test set was 10.23%.

For the next part of our Random Forest Model, we extracted the top five important features and fit our data in the data pipeline again to examine if model performance improves. We see in Table II that the R-squared scores improved and the RMSE on the test set decreases. The R-squared score remains high, with 0.9736 for the training set and 0.9705 for the test set, suggesting the model's strong predictive capability. Additionally, the RMSE values show a slight improvement, with 1717.78 MW for the training data and 1752.68 MW for the test data. The MAPE score was 9.39%, showing a 1% improvement from the previous model. The WAPE score also shows a 1% increase to 9.03%, indicating that the feature selection process helped reduce prediction errors and is good at generalizing to unseen data.

TABLE I
RANDOM FOREST MODEL EVALUATION METRICS

Metric	**Training Set**	**Test Set**
R^2 Score	0.9788	0.9621
RMSE (MW)	1537.14	1987.45
MAE (MW)	852.83	1130.46
MAPE (%)	inf	10.66
WAPE (%)	7.60	10.23

TABLE II
RANDOM FOREST MODEL EVALUATION METRICS (WITH FEATURE EXTRACTION)

Metric	**Training Set**	**Test Set**
R^2 Score	0.9736	0.9705
RMSE (MW)	1717.78	1752.68
MAE (MW)	964.28	997.16
MAPE (%)	inf	9.39
WAPE (%)	8.59	9.03

The results indicate that the selecting only the top k important features further improves our model predictions. The performance improvement after feature selection suggests that dimensionality reduction eliminated less relevant or noisy features, which can otherwise obscure the learning process of the Random Forest model. By focusing on the most important features, model complexity is also reduced and the model is able to generalize better and avoid overfitting, leading to improved accuracy and efficiency

In addition to feature selection to improve the model, we also used randomized search to fine-tune the hyperparameters of the Random Forest Model. The results of the model using the test dataset are shown in table III. The RMSE score for the Random Forest Model on the test set after fine-tuning was 1966.08, which is slightly higher than the previous Random Forest Model without fine-tuning. The MAE also increased to 1120.29, indicating higher error between actual and predicted values. Both MAPE and WAPE scores also increased by a little over 1% compared to the previous model. This indicates that the fine-tuned model perhaps was not using the best parameters, and more testing on different hyperparameter

combination may be needed for further optimization.

TABLE III
RANDOM FOREST MODEL EVALUATION METRICS (FINE-TUNED MODEL)

Metric	Value
Test R^2 Score	0.9629
Test RMSE (MW)	1966.08
Test MAE (MW)	1120.29
Test MAPE (%)	10.58
Test WAPE (%)	10.14

TABLE IV
XGBOOST MODEL1 EVALUATION METRICS

Metric	Training Set	Test Set
R^2 Score	0.9631	0.9616
RMSE (MW)	2031.14	1999.72
MAE (MW)	1275.01	1263.44
MAPE (%)	Inf	17.63
WAPE (%)	11.36	11.44

IV. INSIGHTS FROM XGBOOST MODEL 1 EVALUATION

- **High Predictive Performance:** The model achieves an R^2 score of 0.9616 on the test set, explaining over 96% of the variance in energy consumption data.
- **Low Prediction Errors:** Metrics such as RMSE (1999.72), MAE (1263.44), and WAPE (11.44%) reflect minimal error and strong generalization to unseen data.
- **Handling Zero Targets:** The inf MAPE observed in training is due to zero or near-zero target values. Preprocessing adjustments are recommended to avoid this anomaly. The WAPE metric handles this issue.

TABLE V
XGBOOST MODEL2 EVALUATION METRICS

Metric	Training Set	Test Set
R^2 Score	0.9755	0.9698
RMSE (MW)	1654.82	1775.40
MAE (MW)	944.01	1033.69
MAPE (%)	Inf	10.47
WAPE (%)	8.41	9.36

V. INSIGHTS FROM XGBOOST MODEL 2 EVALUATION

- **High Predictive Performance:** The model achieves an R^2 score of 0.9698 on the test set, explaining approximately 97% of the variance in energy consumption data.
- **Low Prediction Errors:** Metrics such as RMSE (1775.40), MAE (1033.69), and WAPE (9.36%) demonstrate minimal error and strong generalization to unseen data.
- **Handling Zero Targets:** The inf MAPE observed in training is due to zero or near-zero target values. WAPE is used to handle this issue. The MAPE value decreased by 7% from the first XGBoost model, and WAPE decreased by 2%, indicating better model performance with improved predictions after hyperparameter-tuning.

TABLE VI
EVALUATION METRICS FOR RNN MODEL

Metric	Training	Test
R^2 Score	0.7331	0.6104
RMSE (MW)	3333.62	4025.63
MAE (MW)	2468.98	3050.21
MAPE (%)	7.45	9.60
WAPE (%)	7.65	9.74

TABLE VII
EVALUATION METRICS FOR LSTM MODEL

Metric	Training	Test
R^2 Score	0.7455	0.6194
RMSE (MW)	3254.84	3979.01
MAE (MW)	2439.24	3041.33
MAPE (%)	7.46	9.61
WAPE (%)	7.55	9.71

Insights and Observations

The evaluation metrics for both the RNN and LSTM models provide valuable insights into their performance in predicting energy consumption. Below are the key observations:

- **RNN Model:**
  - The R^2 score of 0.7331 (training) and 0.6104 (test) indicates that the model captures a substantial portion of the variance in the data but performs slightly less well on the test set.
  - The RMSE and MAE values on the test set (4025.63 and 3050.21, respectively) are higher than the training set values, suggesting that the model might slightly underfit the training data or struggle with generalization.
  - The MAPE and WAPE percentages are relatively low (9.60% and 9.74%, respectively), indicating that the model performs reasonably well in terms of relative error but may require fine-tuning to handle outliers or extreme fluctuations.
- **LSTM Model:**
  - The R^2 score of 0.7455 (training) and 0.6194 (test) suggests that the LSTM model performs marginally better than the RNN in capturing sequential patterns, but there is still room for improvement.
  - The RMSE and MAE values on the test set (3979.01 and 3041.33, respectively) are slightly lower than those of the RNN, indicating better predictive accuracy.
  - The MAPE and WAPE percentages (9.61% and 9.71%, respectively) are comparable to the RNN,

showing that the LSTM model has a similar level of relative prediction accuracy.

- **Comparison of RNN and LSTM Models:**
  - Both models show a slight drop in performance on the test set compared to the training set, highlighting the need for further regularization or hyperparameter tuning to improve generalization.
  - The LSTM model marginally outperforms the RNN in most metrics, particularly in RMSE and MAE, suggesting its better ability to capture temporal dependencies.
  - Despite the better performance of the LSTM, both models have comparable MAPE and WAPE values, indicating similar relative prediction accuracy in percentage terms.

VI. Conclusion

This project developed an energy consumption prediction model using historical data from multiple U.S. companies to improve power system planning and demand forecasting. The results highlight the importance of feature engineering and data normalization in addressing the skewed distribution of energy consumption data, which significantly improves model convergence and performance. Our experiments reveal that while holidays exhibit minimal correlation with energy consumption trends, other temporal patterns, such as hourly, daily, and seasonal fluctuations, play a more significant role. Key data preprocessing steps, including data synchronization, aggregation, and feature encoding, ensure the model has sufficient and well-structured data to learn effectively. Additionally, feature engineering on our time-series data proved essential for capturing nuanced relationships and fluctuations within energy usage patterns. The low error values achieved by our models demonstrate the effectiveness of the preprocessing pipeline, as well as the careful model selection and optimization strategies employed in this study.

The application of the Random Forest Regressor to the energy consumption forecasting task proved to be highly effective. Initially, the model displayed strong performance, with relatively low RMSE and MAE values and low MAPE, WAPE percentages. After performing feature selection and retaining only the top five most important features, the model demonstrated even better generalization capabilities with lowered error rates. This improvement suggests that the selected features were indeed the most relevant predictors, reducing noise and enhancing the model's ability to generalize to unseen data.

The XGBoost model also demonstrates great predictive performance with great consistency across training and test datasets, with relatively low RMSE and MAE values on test data. The minimal difference between training and test metrics suggests that the XGBoost model is robust to overfitting, achieving good generalization across different datasets. After hyperparameter tuning, the XGBoost model demonstrated improved prediction accuracy, with the MAPE value decreasing over 7% and the WAPE value decreasing 2%.

Although the LSTM model effectively captures non-linear patterns and demonstrates high predictive accuracy for this dataset, traditional machine learning models outperformed both the LSTM and RNN on the test set. Furthermore, the LSTM and RNN models appear to generalize less effectively to unseen data, as evidenced by their higher error rates on the test set.

The feature selection process helped streamline our models by focusing on the most predictive attributes, ultimately leading to more efficient and accurate models. These results indicate that the Random Forest model, particularly when optimized through feature selection, is well-suited for time series forecasting of energy consumption, demonstrating both robustness and predictive power.

Overall, the XGBoost model demonstrated strong robustness against overfitting and achieved the best generalization, as it exhibited the smallest variance between training and testing performance among all the models. The Random Forest model is observed to have the best prediction capabilities as it has the lowest error values. The application of feature engineering, model selection, and systematic tuning underscores the importance of a comprehensive approach to machine learning in the energy domain.

VII. Future Work

Future work should explore optimization methods on current model, as well as integration of advanced machine learning architectures, such as hybrid models combining LSTM and CNN, and attention mechanisms to further refine model accuracy. Additionally, incorporating other contextual variables—such as extreme weather conditions or economic indicators—may help capture complex influences on energy usage, paving the way for even more reliable predictions.

Ultimately, the insights and methodologies developed in this project contribute to the growing field of energy forecasting, highlighting opportunities for machine learning applications to improve energy management and support decision-making in power systems.

References

[1] Ahmed, N.K., Atiya, A.F., Gayar, N.E., El-Shishiny, H.: An empirical comparison of machine learning models for time series forecasting. Econometric Reviews **29**(5-6), 594–621 (2010). https://doi.org/10.1080/07474938.2010.481556

[2] Alanbar, M., Alfarraj, A., Alghieth, M.: Energy consumption prediction using deep learning technique: Case study of computer college. International Journal of Interactive Mobile Technologies (iJIM) **14**(10), 143–153 (2020). https://doi.org/10.3991/ijim.v14i10.14383

[3] Aurna, N.F., Rubel, M.T.M., Siddiqui, T.A., Karim, T., Saika, S., Arifeen, M.M., Mahbub, T.N., Reza, S.M.S., Kabir, H.: Time series analysis of electric energy consumption using autoregressive integrated moving average model and holt winters model. TELKOMNIKA Telecommunication, Computing, Electronics and Control **19**(3), 991–1000 (2021). https://doi.org/10.12928/TELKOMNIKA.v19i3.15303

[4] Bandara, K., Bergmeir, C., Smyl, S.: Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. International Journal of Forecasting **36**(1), 1–19 (2020). https://doi.org/10.1016/j.ijforecast.2019.03.010

[5] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley (2015)

[6] Chou, J.S., Lin, C.H.: Feature engineering in machine learning for energy prediction. Energy Reports **7**, 745–754 (2021). https://doi.org/10.1016/j.egyr.2021.06.078
[7] Deb, C., Lee, S.E., Santamouris, M., Cheng, C.K.: Forecasting energy consumption for commercial buildings using hybrid artificial intelligence modeling schemes. Energy and Buildings **146**, 141–151 (2017). https://doi.org/10.1016/j.enbuild.2017.04.003
[8] Elhabyb, K., Baina, A., Bellafkih, M., Deifalla, A.F.: Machine learning algorithms for predicting energy consumption in educational buildings. International Journal of Energy Research (2024). https://doi.org/10.1155/2024/6812425
[9] Hu, M., Huang, Chen, Y.: A hybrid strategy combining machine learning and wavelet transform for estimating power usage. IEEE Access **8**, 20632–20644 (2020)
[10] Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts (2018)
[11] U.S. Energy Information Administration: Electricity explained: Use of electricity (2022), https://www.eia.gov/energyexplained/electricity/use-of-electricity.php, accessed: 2024-12-15