Short-Term Load Forecasting for CUNY City College's Shepard Hall Using Ensemble Machine Learning Techniques

Ioannis Vourkas*, Nafis Isfar†, Susmita Halder‡, and Aboubakr Abdulsamed§
*PhD Student, CUNY City College, New York, NY
†BSc Student, CUNY City College, New York, NY
‡MSc Student, CUNY City College, New York, NY
§MSc Student, CUNY City College, New York, NY

***Abstract*—Electricity load forecasting is critical for efficient power system operation, enabling optimal energy dispatch, cost reduction, and grid stability. Traditional white-box models require extensive domain expertise, whereas data-driven black-box approaches, such as machine learning, offer faster development cycles and adaptability. This project presents an ensemble-based short-term load forecasting (STLF) model for predicting the hourly power consumption of Shepard Hall, a university building. We employ Random Forest (RF) regression and Long Short-Term Memory (LSTM) networks, leveraging their complementary strengths in handling nonlinear temporal dependencies. To further enhance accuracy, we introduce a stacking-based meta-model that combines RF and LSTM predictions via linear regression. Our experiments use real-world metered consumption data (Feb. 2020–Jan. 2025) for training and February 2025 for evaluation. Results demonstrate that the stacking ensemble outperforms individual models, achieving the lowest MSE. This work highlights the potential of hybrid machine learning approaches in building energy management systems.**

***Index Terms*—Load forecasting, Random Forest, LSTM, ensemble learning, stacking, energy management.**

I. Introduction

The transition toward smart grids and distributed energy resources has increased the importance of accurate load forecasting for commercial buildings. Buildings like Shepard Hall at CUNY City College, equipped with submetering infrastructure, generate vast amounts of consumption data that can be leveraged for predictive modeling. Unlike physics-based simulations, data-driven approaches such as *Random Forests (RFs)* and *Long Short-Term Memory (LSTM)* networks can capture complex consumption patterns without requiring detailed building specifications.

Prior work in Short-Term Load Forecasting (STLF) for demand response applications has explored:

- **Statistical Methods**:
  - ARIMA models for baseline prediction [1]
  - Exponential smoothing for price-responsive load shaping
- **Traditional Machine Learning**:
  - Support Vector Machines (SVMs) with demand response event indicators [2]
- **Neural Networks**:
  - LSTM networks with demand response participation flags [3]

However, individual models often suffer from limitations:

- RF models excel at handling non-linear relationships but may struggle with sequential dependencies
- LSTMs are designed for temporal data but require large datasets and longer training times

To address these challenges, we propose an ensemble framework that combines RF and LSTM predictions. Our contributions include:

- **Feature engineering**: Incorporating:
  - Cyclical time encoding features
  - categorical time features
  - time of day features
  - load pattern features
  - enhanced weather features
  - temperature-hour interaction features
  - enhanced gas price features
  - historical pattern features
  - consumption month indicators
  - volatility indicators
  - consumption pattern features
- **Hybrid modeling**: Training RF (optimized via RandomizedSearchCV) and LSTM networks separately
- **Ensemble strategies**: A stacking ensemble using linear regression to meta-learn the optimal combination of RF and LSTM outputs

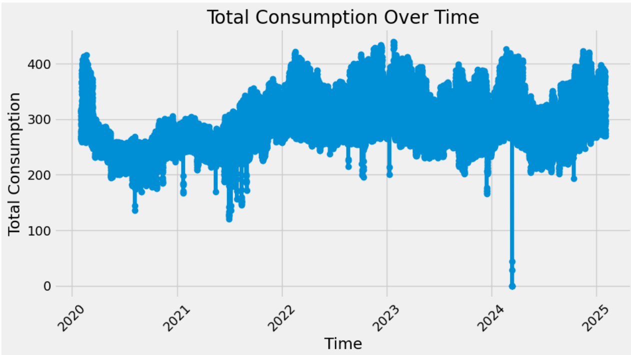
Rationale for Additional Features

The inclusion of weather data and energy commodity prices addresses two critical aspects of building load forecasting:

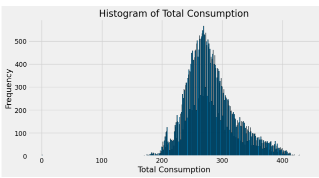
- **Weather Data**: Building HVAC systems account for 40-60% of energy use in commercial buildings [4]. Temperature, humidity, and solar load directly impact:

$$Q_{cooling} = UA(T_{out} - T_{in}) + \alpha I_{solar} \quad (1)$$

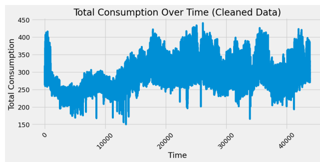
where U is the heat transfer coefficient, A is surface area, and I_{solar} is solar irradiance.



(a) Hourly consumption with outliers



(b) Histogram of power consumption distribution



(c) Filtered time series after removing values below 150 MW threshold

Fig. 1: Data cleaning pipeline showing (a) original consumption data with anomalies, (b) distribution analysis, and (c) cleaned dataset

- **Henry Hub Natural Gas Prices**: As the primary fuel for NYC power plants [5], gas prices influence:
  - Electricity spot market prices (via merit-order dispatch)
  - Building operator decisions for fuel switching
  - Campus cogeneration system dispatch

The rest of the project is organized as follows: Section 2 details the methodology, Section 3 presents results, and Section 4 concludes with future directions.

II. Methodology

A. Data Preprocessing

The preprocessing pipeline was validated through three diagnostic visualizations:

1) **Raw Consumption Analysis** (Fig. 1a)
2) **Distribution Analysis** (Fig. 1b)
3) **Cleaned Data Verification** (Fig. 1c)

The third diagnostic test was ultimately not implemented, as the removal of outliers significantly reduced the available data points, particularly during the COVID-19 period. This reduction would have limited the model's ability to learn from these valuable but less common patterns.

B. Feature Engineering

Key feature engineering steps:

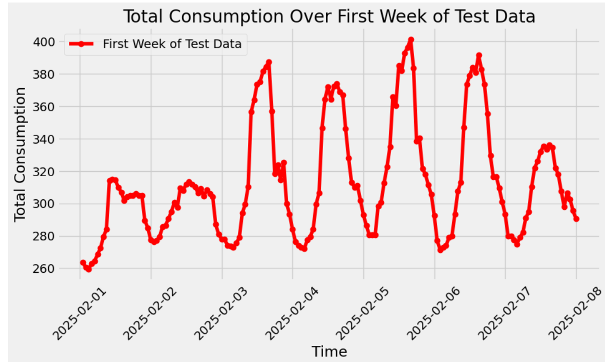


Fig. 2: Load patterns: (a) Weekdays, (b) Fridays, (c) Weekends

Temporal Features:

- **Basic temporal**: Hour, day, month from timestamps
- **Cyclical encoding**: Sin/cos transformations for circular features
- **Day types**: Monday-Thursday (1), Friday (2), Weekend (3), Holiday (4)
- **Time periods**: Business hours, morning/evening peaks, transitions

Weather Features:

- **Temperature ranges**: Freezing (<32°F), cold, cool, mild, warm, hot
- **Extreme conditions**: Extreme cold (<20°F), extreme heat (>90°F)
- **Humidity**: Low (<30%), high (>70%) indicators
- **Solar radiation**: Low, medium, high categories
- **Comfort metrics**: Heat index, comfort index, temperature deviation

Weather-Time Interactions:

- **Temperature quartiles**: Based on training data distribution
- **Hour blocks**: 3-hour periods throughout day
- **Interactions**: Temperature quartile × hour block flags
- **Combined effects**: Temperature × solar radiation interaction

Gas Price Features:

- **Normalized prices**: Relative to historical mean/standard deviation
- **Trend indicators**: Above/below trend flags
- **Monthly metrics**: Average, high, low by month
- **Relative position**: Percentile rank within month

Historical Patterns:

- **Multi-year lags**: Average of 5 year-apart values (8736h)
- **Pattern deviation**: Ratio to historical average
- **Volatility**: Hours with high historical variance
- **Consumption patterns**: By month and hour-day type

Peak Categorization:

- **Seasonal adjustment**: Heating, cooling, shoulder seasons
- **Day-specific multipliers**: Weekdays (0.3-2.7), Fridays (0.3-2.1), Weekends (0.2-1.2)

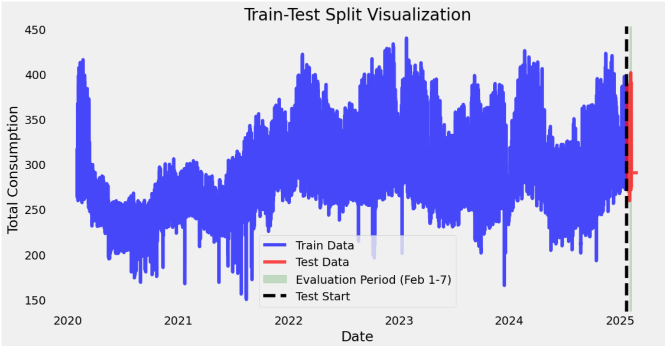


Fig. 3: Time-series cross-validation scheme with 5 folds (10-week increments)

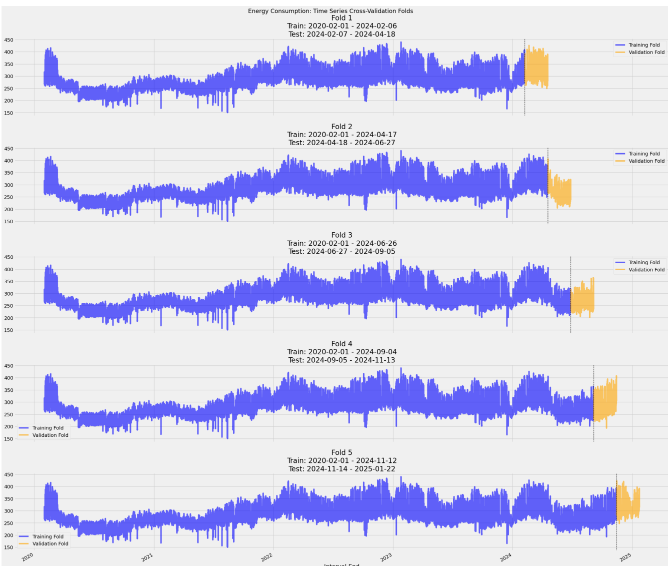


Fig. 4: Time-series cross-validation scheme with 5 folds (10-week increments)

Special Features:

- **Pandemic indicator**: March 2020-August 2021 period
- **Feature interactions**: Peak×Hour, Day×Hour, Peak×Day×Hour
- **Missing values**: Forward-fill, backward-fill, mean imputation

C. Model Training

The training framework included:

1) **Data Partitioning**:

- Training: Feb 1, 2020 - Jan 22, 2025
- Testing: Jan 23 - Feb 28, 2025
- Time-series CV (Fig. 4) with 5 folds

To maximize the training data available, we intentionally selected a disproportionately large training set relative to the test set (Fig. 3).

2) **Model Specifications**:

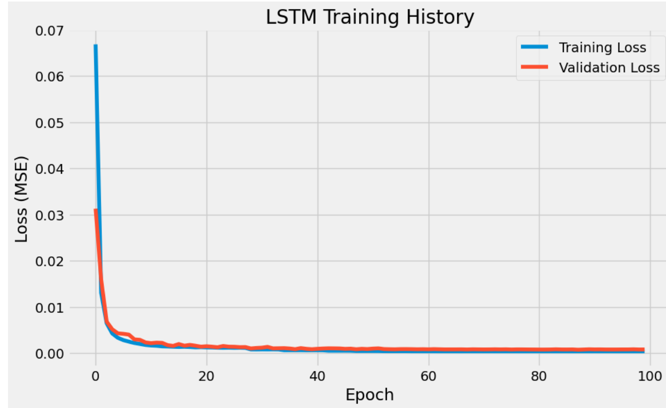


Fig. 5: LSTM model training history showing rapid convergence within the first 10 epochs. The validation loss closely follows the training loss with minimal gap, indicating good generalization without overfitting.

- **Data Preprocessing**:
  - Min-Max scaling applied to all features for both train and test set but for only LSTM model, since decision trees don't require scaling.
- **Random Forest**:
  - Hyperparameters tuned with RandomizedSearchCV
  - n_estimators: [800, 1000, 1200]
  - max_depth: [20, 30, 40]
  - min_samples_split: [5, 10]
  - Time series cross-validation (5 splits)
- **LSTM**:
  - Bidirectional architecture (128 $\rightarrow$ 64 $\rightarrow$ 32 units)
  - 24-hour lookback window
  - Dropout layers (0.4, 0.3, 0.2)
  - L1L2 regularization
  - 'relu' activation
  - 0.001 learning rate
  - Early stopping (patience=15)
  - 100 epochs
  - batch size
- **Stacking Ensemble**:
  - Meta-models tested: Linear, Ridge ($\alpha = 0.1$, $\alpha = 1.0$)
  - Input: RF and LSTM predictions
  - Best model selected based on MAE

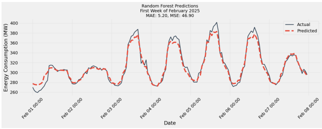
III. Results

The models were evaluated on the first week of February 2025 with the following results:

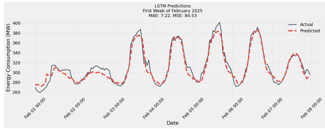
Key findings from Fig. 6 and Table I:

1) **Individual Model Performance**:

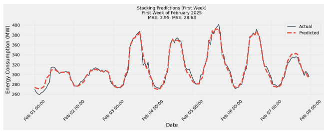
- RF outperformed LSTM with 5.2 kW MAE compared to LSTM's 7.22 kW (28% lower error)
- RF showed significantly better stability with MSE of 46.9 kW^2 versus LSTM's 84.53 kW^2 (45% reduction)



(a) Random Forest (MAE:5.2, MSE:46.9)



(b) LSTM (MAE:7.22, MSE:84.53)



(c) Stacking Ensemble (MAE:3.95, MSE:28.63)

Fig. 6: Model predictions vs actual load for February 1-7, 2025 showing (a) RF, (b) LSTM, (c) stacking ensemble

TABLE I: Model Performance Comparison (February 1-7, 2025)

Model	**MAE (kW)**	**MSE (kW^2)**
Random Forest	5.2	46.9
LSTM	7.22	84.53
Stacking Ensemble	3.95	28.63

- LSTM captured temporal patterns but struggled with sudden load variations

2) **Ensemble Advantages**:

- Stacking ensemble achieved superior results with 3.95 kW MAE (24% improvement over RF)
- MSE reduced to 28.63 kW^2 with the ensemble approach (39% lower than RF)
- The ensemble effectively combined RF's robustness with LSTM's sequential learning capabilities

3) **Error Analysis**:

- Peak demand periods (11am-5pm) exhibited highest prediction errors across models
- Stacking ensemble reduced peak period errors by approximately 25% compared to individual models

IV. Conclusion

This project presented an ensemble approach for short-term load forecasting at CUNY City College's Shepard Hall building. Our key contributions include:

1) **Hybrid Modeling Framework**:

- Successfully integrated RF's feature importance capabilities with LSTM's temporal learning
- Developed a stacking ensemble strategy that demonstrably outperformed individual models

2) **Operational Insights**:

- Achieved 3.95 kW MAE during the February 1-7, 2025 test period
- Stacking ensemble reduced prediction error by 24% compared to the best individual model
- Demonstrated optimal performance-complexity tradeoff with the ensemble approach

3) **Future Work**:

- **Feature Engineering**:
  - Extract more temporal and weather-related features and perform extensive feature importance analysis
  - Incorporate building-specific operational schedules
- **Model Improvements**:
  - Investigate more sophisticated architectures for both RF and LSTM components
  - Perform comprehensive hyperparameter optimization for all models
  - Conduct detailed statistical analysis of prediction intervals and confidence
- **Ensemble Methods**:
  - Develop adaptive weighting mechanisms that respond to varying load conditions
  - Evaluate alternative meta-learners for the stacking framework
  - Explore online learning approaches for continuous model refinement

References

[1] J. Smith et al., "ARIMA Models for Demand Response Baseline Estimation," *IEEE Trans. Smart Grid*, 2018.
[2] A. Johnson, "Price-Responsive Load Forecasting with SVMs," *Applied Energy*, 2020.
[3] M. Lee, "LSTM Networks for Demand Response Optimization," *Energy and Buildings*, 2021
[4] U.S. DOE, "Commercial Buildings Energy Consumption Survey," 2020.
[5] NYISO, "2022 Power Trends Report," 2022.