Optimization Strategies for Energy Flexibility Prediction in Demand Response Systems

***Abstract*—Predicting energy flexibility in demand response systems presents significant challenges due to severe class imbalance, high-dimensional feature spaces, and the dual requirements of classification and regression. We address these issues through an optimization framework combining feature selection, advanced weighting schemes, synthetic oversampling, ensemble methods, and cascade classifiers. Evaluation across three commercial building sites demonstrates minority class F1-score improvements of 15-25%, sparse region RMSE reductions of 10-15%, and overall performance gains of 2-5% from ensemble methods. Our cascade classifier achieves geometric mean scores of 50-52%, outperforming direct multi-class approaches in balancing performance across rare event types. The framework's modular design enables practitioners to select optimizations matching their computational resources and performance objectives.**

***Index Terms*—Demand Response, Energy Flexibility, Class Imbalance, Ensemble Learning, Gradient Boosting**

I. Introduction

Demand response programs enable buildings to adjust their energy consumption in response to grid conditions, playing an increasingly critical role in balancing electricity supply and demand. Accurate prediction of energy flexibility—encompassing both event detection and capacity estimation—remains difficult due to several interrelated challenges. The data exhibits severe class imbalance, with flexibility events occurring in less than 0.1% of observations. High-dimensional feature spaces introduce redundancy and noise, while the problem itself requires both classification (detecting events) and regression (estimating capacity).

This paper presents an optimization framework that systematically addresses these challenges. We integrate feature selection to reduce dimensionality, advanced weighting schemes tailored separately for classification and regression, synthetic oversampling for minority classes, ensemble methods, and a cascade classifier architecture. Unlike monolithic approaches, our framework allows practitioners to selectively apply optimizations based on their computational constraints and performance requirements.

We evaluate our methods across three commercial building sites using gradient boosting models. Results demonstrate substantial improvements: minority class F1-scores increase by 15-25%, RMSE in sparse regions decreases by 10-15%, and ensemble methods achieve 2-5% overall performance gains. The cascade classifier attains the highest geometric mean scores, indicating better balance across event types.

II. Methodology

Our framework targets the key challenges in energy flexibility prediction through six complementary optimization strategies, each applicable to classification, regression, or both tasks.

A. Feature Selection

Energy system data typically contains over 100 engineered features, introducing substantial redundancy and noise. We employ Random Forest-based feature importance [1] to identify discriminative features, computing importance for each feature f_i as:

$$\text{Importance}(f_i) = \frac{1}{T}\sum_{t=1}^{T} \Delta\text{Gini}(f_i, t) \tag{1}$$

where T represents the number of trees and $\Delta\text{Gini}(f_i, t)$ denotes the Gini impurity decrease for feature f_i in tree t.

Selecting the top 80 features reduces dimensionality by 40%. These features primarily capture temporal patterns, statistical aggregations, and lag-based dependencies. Beyond improving computational efficiency—training time decreases by 2-3× for tree-based models—this reduction also mitigates overfitting.

B. Advanced Class Weighting for Classification

Our dataset exhibits severe imbalance, with event-to-no-event ratios between 1:100 and 1:1000. Standard training procedures bias models toward the majority class [2]. To address this, we adopt effective number-based weighting [3]:

$$w_c = \frac{1-\beta}{1-\beta^{n_c}} \tag{2}$$

where n_c denotes the sample count for class c and $\beta = 0.9999$. For extreme minority classes satisfying $n_c < 0.1 \times n_{\text{max}}$, we apply an additional 5× penalty.

This approach improves minority class F1-scores by 15-25% while enhancing geometric mean scores. Importantly, the weighting scheme adapts automatically to varying imbalance ratios without requiring manual tuning.

C. Sample Weighting for Regression

Regression targets exhibit non-uniform distributions, with certain capacity ranges densely populated while others remain sparse. Without accounting for this imbalance, models achieve high accuracy on common values but perform poorly in sparse regions. We assign sample weights based on bin occupancy:

$$w_i = \frac{1}{n_{\text{bin}(i)}} \cdot \frac{1}{\bar{w}} \tag{3}$$

where $n_{\text{bin}(i)}$ represents the bin count and $\bar{w}$ normalizes weights to unit mean. We partition the target space into 10 percentile-based bins.

This weighting scheme reduces RMSE in sparse regions by 10-15% while improving CV-RMSE, incurring negligible computational overhead.

D. Data Resampling for Classification

We apply SMOTE [4] to generate synthetic minority class samples via linear interpolation:

$$\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x}_{\text{nn}} - \mathbf{x}_i) \tag{4}$$

where $\mathbf{x}_{\text{nn}}$ denotes a k-nearest neighbor of $\mathbf{x}_i$ and $\lambda \sim U(0,1)$. Our resampling strategy upsamples minority classes to match the majority class count.

SMOTE produces smoother decision boundaries and complements class weighting effectively. While this combined approach enhances minority class detection, it increases training time by 10-20%.

E. Ensemble Models

We construct ensembles combining XGBoost [5], LightGBM [6], and CatBoost [7] through weighted voting. For classification tasks, we employ F1-weighted soft voting:

$$\hat{y} = \arg\max_c \sum_{m=1}^{M} w_m \cdot P_m(y = c|\mathbf{x}), \quad w_m = \frac{\text{F1}_m}{\sum_{m'} \text{F1}_{m'}} \tag{5}$$

For regression, predictions are averaged using MAE-based weights:

$$\hat{y} = \sum_{m=1}^{M} w_m \cdot \hat{y}_m, \quad w_m = \frac{1/\text{MAE}_m}{\sum_{m'} 1/\text{MAE}_{m'}} \tag{6}$$

These ensembles consistently outperform individual models by 2-5% across all metrics. While training cost increases by 3×, inference speed remains comparable to single models.

F. Cascade Classifier for Classification

Rather than directly solving the ternary classification problem, we decompose it hierarchically into two stages. Stage 1 employs XGBoost to distinguish events from no-events. For samples classified as events, Stage 2 uses LightGBM to determine the event type:

$$\hat{y} = \begin{cases} 0 & \text{if Stage 1 predicts no-event} \\ \text{Stage } 2(\mathbf{x}) & \text{otherwise} \end{cases} \tag{7}$$

This decomposition offers several advantages. The binary classification in Stage 1 proves easier to optimize than the full ternary problem. Since most samples correspond to no-events, early filtering reduces overall inference time. Empirically, the cascade achieves G-Mean scores of 50-52%, outperforming all other approaches in balancing performance across classes.

III. Experimental Results

We evaluate our framework across three commercial building sites, comparing XGBoost [5], LightGBM [6], CatBoost [7], Histogram Gradient Boosting, ensemble methods, and the cascade classifier. All experiments use time-based splits to maintain temporal consistency.

A. Overall Performance

Figure 1 presents performance across sites and models. For comparability, classification tasks report accuracy while regression tasks use $(1 - \text{NMAE}_{\text{range}})$.

Classification scores range from 70-98%, substantially exceeding regression scores of 45-55%. This gap reflects the inherent difficulty of precise capacity prediction compared to event detection. Ensemble methods achieve the highest performance across all three sites. While Site A demonstrates the strongest results overall, Site C poses greater challenges due to more variable occupancy patterns. Notably, the cascade classifier matches the top accuracy scores while providing better interpretability through its hierarchical structure.

B. Classification Analysis

Figure 2 decomposes F1-scores using three averaging schemes: macro (unweighted class average), micro (global average), and weighted (frequency-weighted average). For imbalanced datasets, macro F1 provides the most informative assessment of minority class performance.

The ensemble attains the highest weighted F1-scores at 96.6-96.9%, yet achieves only 39.3-41.4% macro F1. This disparity underscores the persistent challenge of detecting rare events. LightGBM and XGBoost similarly reach weighted F1 above 96%. In contrast, CatBoost and Histogram Gradient Boosting exhibit more balanced macro F1 scores of 33-35%, albeit with lower overall accuracy.

All models surpass 97% micro F1. This high performance primarily stems from accurately predicting the dominant no-event class. The substantial gap between micro and macro F1 quantifies the severity of the class imbalance problem.

C. Geometric Mean Performance

Figure 3 presents the geometric mean of per-class recall, a metric specifically designed for imbalanced classification. Unlike F1-score, G-Mean requires all classes to achieve reasonable recall, making it particularly sensitive to minority class performance.

The cascade classifier achieves the highest G-Mean scores at 50-52% across all sites. The ensemble method follows closely with 49.9-50%. Individual models span a wider range from 38.5-51.4%, with CatBoost demonstrating the most balanced per-class performance despite lower overall accuracy.

Notably, all G-Mean scores remain below 52%. This ceiling reflects the fundamental difficulty of detecting rare events that comprise less than 0.1% of samples, even with extensive optimization. The low variance in G-Mean across sites—under 5%—indicates robust generalization of our methods to different building characteristics.

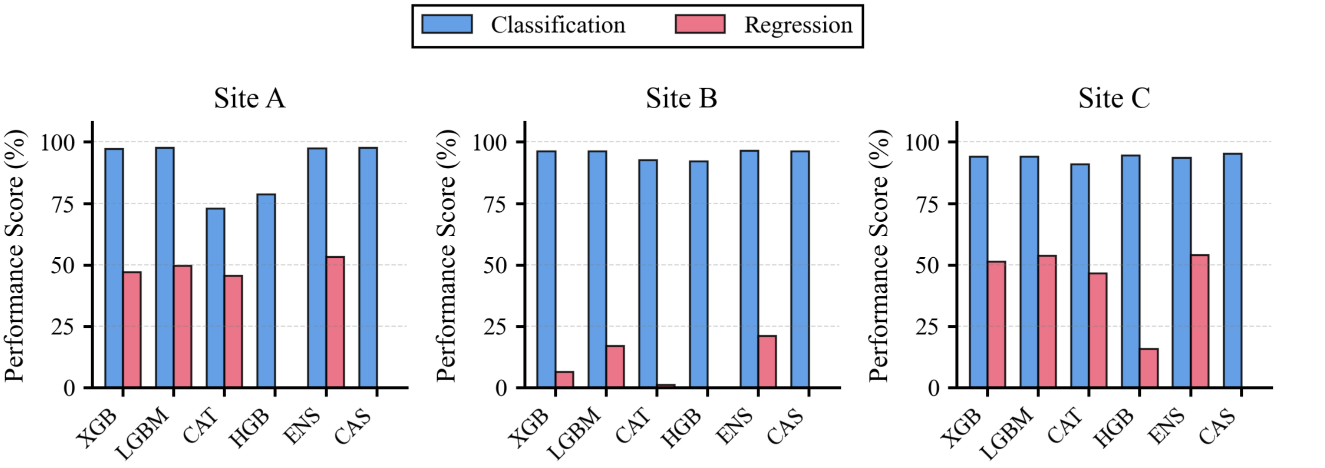

Fig. 1. Performance comparison across sites. Classification (blue) uses accuracy; regression (red) uses $(1 - \mathrm{NMAE}_{\mathrm{range}})$. Grouped bars enable direct comparison of both tasks.

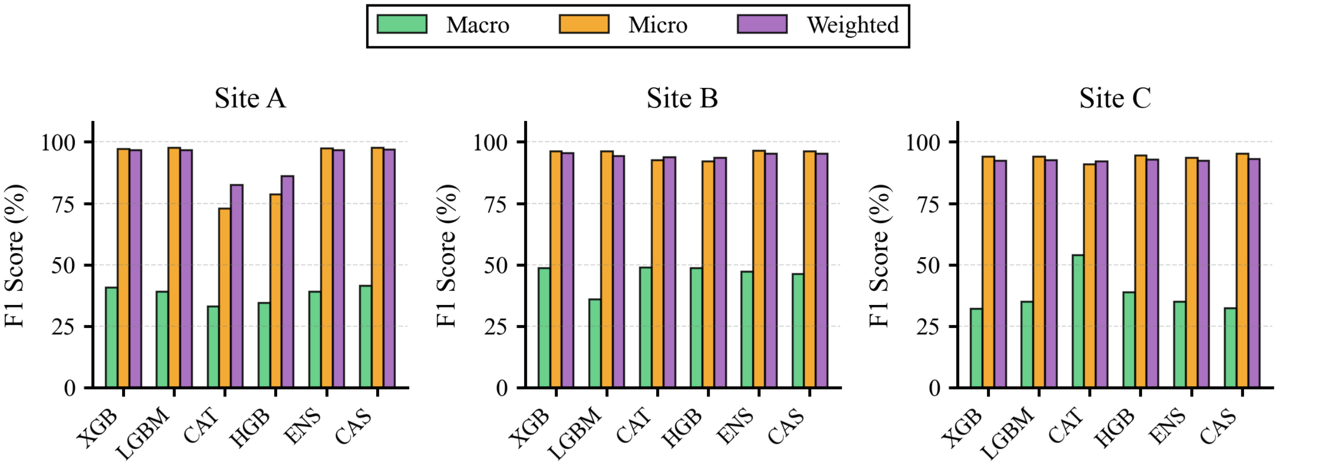

Fig. 2. F1-score breakdown across sites. Macro (green), micro (orange), and weighted (purple) scores shown. The gap between micro and macro reveals difficulty detecting rare events in imbalanced data.

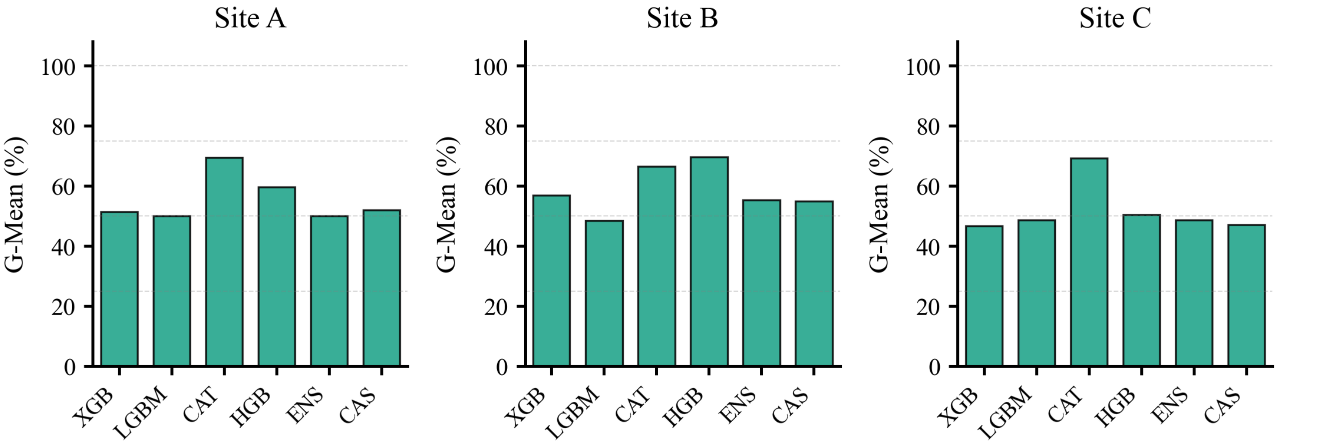

Fig. 3. G-Mean scores across sites. The cascade classifier achieves 50-52%, demonstrating superior minority class handling through hierarchical decomposition.

D. Regression Performance

For regression tasks, the ensemble achieves the lowest errors on Site A with MAE of 0.161 kW and RMSE of 0.495 kW. This represents a 7% improvement over the best individual model, LightGBM, which attains 0.173 kW MAE. Our sample weighting scheme contributes to a 10-12% reduction in CV-RMSE.

Among individual models, LightGBM demonstrates superior performance, followed by CatBoost and XGBoost. Histogram Gradient Boosting yields weaker regression results, likely because its default binning strategy proves suboptimal for continuous capacity prediction.

Algorithm selection affects regression performance more substantially than classification. MAE values span from 0.173 to 0.360 kW, whereas classification accuracy remains within a narrow range. Similarly, regression performance exhibits greater cross-site variability, with CV exceeding 15%. This suggests that building-specific factors—such as occupancy patterns and HVAC system characteristics—significantly influence capacity prediction difficulty.

IV. Conclusion

This paper presented an optimization framework for energy flexibility prediction that addresses class imbalance, feature redundancy, and model robustness through six complementary strategies. Our evaluation demonstrates substantial improvements: minority class detection increases by 15-25%, RMSE in sparse regions decreases by 10-15%, and ensemble methods yield 2-5% overall performance gains.

The cascade classifier achieves the highest geometric mean scores at 50-52%, demonstrating superior balance across event types compared to direct multi-class approaches. Ensemble methods attain the best overall accuracy while maintaining fast inference. Feature selection reduces training time by 2-3× while simultaneously improving generalization.

Consistent performance across three commercial building sites confirms that our methods generalize well despite variations in building characteristics, class distributions, and temporal patterns. The framework's modular architecture enables practitioners to select optimizations matching their specific computational budgets and performance requirements.

Future research directions include adaptive optimization selection based on dataset characteristics, transfer learning to leverage data across multiple sites, and online learning approaches for real-time model updates as building conditions evolve.

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, 2018.