EnsembleCI: Ensemble Learning for Carbon Intensity Forecasting

Leyi Yan
University of Waterloo
Waterloo, Canada
l52yan@uwaterloo.ca

Linda Wang
University of Waterloo
Waterloo, Canada
lindaw@uwaterloo.ca

Sihang Liu
University of Waterloo
Waterloo, Canada
sihangliu@uwaterloo.ca

Yi Ding*
Purdue University
West Lafayette, IN, USA
yiding@purdue.edu

Abstract

Carbon intensity (CI) measures the average carbon emissions generated per unit of electricity, making it a crucial metric for quantifying and managing the environmental impact. Accurate CI predictions are vital for minimizing carbon emissions, yet the state-of-the-art method (CarbonCast) falls short due to its inability to address regional variability and lack of adaptability.

To address these limitations, we introduce EnsembleCI, an adaptive, end-to-end ensemble learning-based approach for CI forecasting. EnsembleCI combines weighted predictions from multiple sublearners, offering enhanced flexibility and regional adaptability. In evaluations across 11 regional grids, EnsembleCI consistently surpasses CarbonCast, achieving the lowest mean absolute percentage error (MAPE) in almost all grids and improving prediction accuracy by an average of 19.58%. While performance still varies across grids due to inherent regional diversity, EnsembleCI reduces variability and exhibits greater robustness in long-term forecasting compared to CarbonCast and identifies region-specific key features, underscoring its interpretability and practical relevance. These findings position EnsembleCI as a more accurate and reliable solution for CI forecasting. EnsembleCI source code and data used in this paper are available at https://github.com/emmayly/EnsembleCI.

CCS Concepts

• Social and professional topics → Sustainability; **• Computing methodologies → Ensemble methods**.

Keywords

Grid Carbon Intensity, Multi-day Forecasting, Ensemble Learning

ACM Reference Format:
Leyi Yan, Linda Wang, Sihang Liu, and Yi Ding. 2025. EnsembleCI: Ensemble Learning for Carbon Intensity Forecasting. In *The 16th ACM International Conference on Future and Sustainable Energy Systems (E-ENERGY '25), June 17–20, 2025, Rotterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3679240.3734630

1 Introduction

The rapid growth of computing infrastructure, particularly in datacenters and AI systems, has significantly increased carbon emissions, with datacenters alone accounting for 4% of global electricity consumption [3, 20, 36]. AI workloads, like training and serving large language models, further amplify this impact [13, 18, 28, 31, 37]. The electricity powering these systems comes from both non-renewable and renewable sources, with non-renewables having a much higher carbon emissions [34, 35].

Carbon intensity (CI) is a critical metric used to quantify the average carbon emissions associated with electricity generation. Expressed in gCO_2e/kWh, it represents the equivalent carbon emissions per kilowatt-hour of electricity. In this paper, we adopt the generation-based definition of CI, which accounts for the average carbon emissions from power generation at each source. Accurate CI predictions are essential for reducing the operational carbon emissions of computing and energy systems through load shifting and workload scheduling [3, 21, 32].

However, predicting CI is difficult due to regional energy variability and the complexity of long-term forecasting [25], which make universal modeling and accurate temporal prediction challenging. The state-of-the-art CarbonCast [26, 27] approach improves CI forecasting with a two-tier neural network using historical energy, carbon, and weather data. However, it has two limitations: a fixed, manually designed architecture that does not adapt to regional or energy source differences, and an inefficient two-tier structure that, as our analysis shows, offers no clear benefit over simpler alternatives (§2.3).

To address these limitations, we present EnsembleCI, an adaptive, end-to-end ensemble learning-based approach that integrates weighted predictions from a pool of predictive models (i.e., sublearners) for improved flexibility and accuracy, instead of relying on a single, fixed model architecture. This ensemble approach *accommodates variations and diversity* across regional girds to achieve more accurate and reliable CI predictions, and enables an *end-to-end architecture* that streamlines training and prediction by reducing redundant computations.

We evaluate EnsembleCI against CarbonCast across 11 grids (6 in the US and 5 in the EU) by predicting CI based on direct emission factors up to 4 days into the future. We use *mean absolute percentage error* (MAPE) as the metric. Overall, EnsembleCI achieves the lowest MAPE in 10–11 grids, while CarbonCast only achieves the lowest MAPE in at most 1 grid. Furthermore, for day-1 through day-4 forecasts, EnsembleCI outperforms CarbonCast by 18.1%, 17.13%, 19.69%, and 23.4%, respectively. Additionally, EnsembleCI demonstrates greater robustness in long-term forecasting, with smaller MAPE increases by day 4 (EnsembleCI: 3.63%, CarbonCast: 5.43%). Using permutation feature importance, we identify that important features vary by grids, while sublearners strongly align in selecting the most impactful features within each grid, emphasizing their robustness. We summarize our contributions as follows.

- Identifying and analyzing the limitations of the state-of-the-art CI forecasting method CarbonCast.

*Corresponding faculty author: yiding@purdue.edu.

E-ENERGY '25, Rotterdam, Netherlands

ACM ISBN 979-8-4007-1125-1/25/06
https://doi.org/10.1145/3679240.3734630

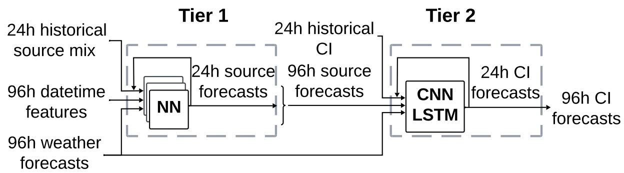

Figure 1: The architecture design of CarbonCast [26, 27].

- Developing EnsembleCI, an adaptive, end-to-end ensemble learning-based approach for accurate and flexible CI forecasting.
- Conducting extensive evaluation to demonstrate the effectiveness and robustness of EnsembleCI.

2 Background and Motivation

This section introduces CarbonCast, evaluates its forecasting performance across grids, analyzes the energy mixes behind its inconsistencies, and explores alternative predictors for improvement.

2.1 The State-of-the-Art: CarbonCast

To forecast carbon intensity using publicly available data, such as the U.S. Energy Information Administration (EIA) [4], without relying on third-party services like ElectricityMap [1] and WattTime [2], CarbonCast has been introduced. This state-of-the-art open-source tool enables hourly carbon intensity predictions for up to 96 hours into the future [26, 27]

Figure 1 illustrates CarbonCast's model architecture design, which employs a two-tier architecture. Tier 1 trains individual neural network models for each energy source to forecast energy production up to 4 days ahead, based on historical energy production data, weather forecasts (for renewables only), and temporal features. Tier 2 combines these energy forecasts with historical carbon intensity data, using a CNN-LSTM model with 1-D CNN layers for short-term patterns and an LSTM layer for long-term dependencies to predict hourly carbon intensity.

To evaluate CarbonCast's carbon intensity forecasting, we evaluate the 11 regional grids as listed in Table 1 [1]. Methodology details are discussed in §4.1. Prediction accuracy is measured using mean absolute percentage error (MAPE), where lower values indicate better performance. Figure 2 shows the averaged hourly prediction accuracy results of CarbonCast across 11 grids up to 4 days into the future. The results exhibit significant inconsistency in prediction accuracy across grids. Notably, ES recorded the highest average MAPE of 21.03%, with values of 23.86% and 25.81% on the latter days. CISO (13.27%), ERCOT (11.1%), DE (17.18%), NL (10.3%), and SE (12.5%) have moderate average MAPE values. PJM (5.21%), EPE (1.73%), ISNE (6.2%), MISO (7.27%), and PL (4.99%) have the lowest MAPE values. This inter-regional variability in prediction accuracy raises questions about the CarbonCast's reliability and generalizability, which we will investigate next.

2.2 Why does CarbonCast not perform well consistently across grids?

To understand why highly accurate prediction is hard and CarbonCast performs inconsistently across grids, we examine the energy source mix for each grid. Figure 3 shows the breakdown of energy

Table 1: 11 grids in this paper: their geographical coverage, average CI in 2024 (gCO_2e/kWh), and electricity sources.

Grid	State/Country	Avg. CI	Primary electricity sources
CISO	CA	230	Natural gas, solar, hydro
PJM	IL, MI, IN, OH, PA, NJ, DE, DC, KY, WV, VA, NC	398	Natural gas, nuclear, coal
EPE	TX, NM	540	Natural gas, solar
ISNE	ME, VT, NH, MA, CT, RI	298	Natural gas, nuclear, hydro
ERCOT	TX	371	Natural gas, wind, coal
MISO	MN, WI, IA, IL, IN, MI, AR, MS, LA, KY, ND, TX, MO	485	Natural gas, coal, wind
DE	Germany	333	Wind, coal, solar
SE	Sweden	23	Hydro, nuclear, wind
ES	Spain	125	Wind, nuclear, solar
NL	Netherlands	263	Natural gas, wind, solar
PL	Poland	704	Coal, wind, natural gas

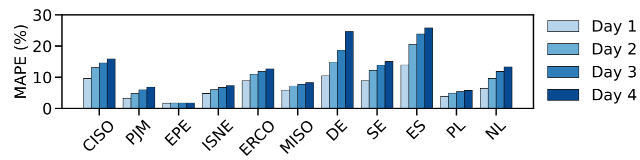

Figure 2: Averaged hourly prediction accuracy of CarbonCast for 11 regional grids over a 4-day horizon.

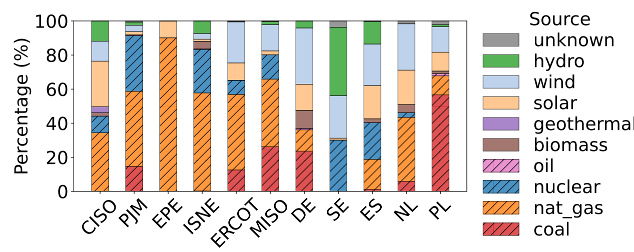

Figure 3: Energy source mixes for 11 regional grids in 2024.

sources for 11 regional grids in 2024 [1], revealing significant regional variations in both types and proportions. Analyzing these alongside prediction accuracy, we derive the following insights.

Grids with high levels of renewable energy integration, such as wind and solar, face considerable challenges in accurate forecasting due to their inherent variability. For instance, in Spain (ES), where 24.36% of energy is derived from wind, 13.17% from hydro, and 19.55% from solar, the MAPE ranges from 13.93% to 25.81%. The variability in output from these sources is driven by complex environmental factors, including topography, surface roughness, temperature inversions, and cloud cover, all of which introduce significant uncertainty into predictions [11]. Conversely, grids highly relying on fossil fuels tend to achieve higher forecasting accuracy because energy production remains relatively unaffected by weather fluctuations. In the case of EPE, where 90.12% of energy is generated from natural gas, the MAPE of 1–4 days prediction is remarkably low, ranging between 1.68% and 1.76%.

2.3 Can other methods surpass CarbonCast?

To address CarbonCast's inconsistent performance across grids, we explore alternative models that may perform better. We hypothesize

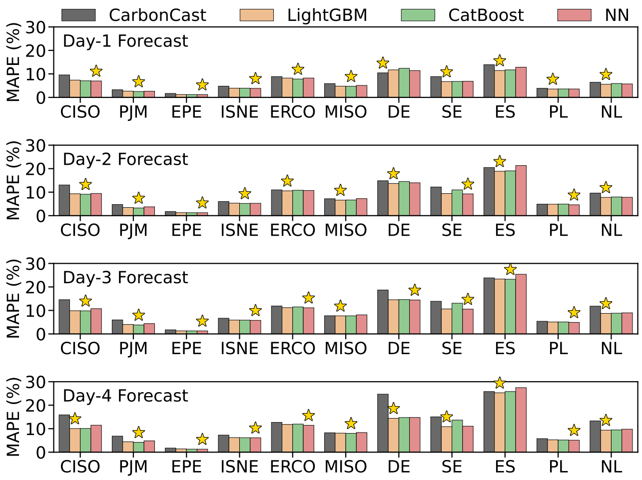

Figure 4: Prediction accuracy of different methods. The best method for each grid on each day is marked with a star.

that region-specific factors require tailored modeling. To test this hypothesis, we evaluated three high-performing methods: LightGBM [24], CatBoost [14], and one neural network model [22]. While we also tested models like XGBoost [12] and Random Forest [7], these three outperformed the rest.

- **LightGBM** is a tree-based gradient boosting framework designed for faster training, reduced memory usage, and higher accuracy [24]. We select it for its outstanding performance in time series predictions, such as greenhouse temperature prediction [9].
- **CatBoost** is a tree-based gradient boosting framework optimized for categorical features [14]. We select it as it automatically handles categorical features while reducing overfitting and bias [15].
- **NN** is a standard neural network implemented using the FastAI Tabular Model [22]. It includes an input layer followed by three fully connected layers. The first two layers apply ReLU activation, batch normalization, and dropout for regularization, with 200 and 100 neurons, respectively; the final layer outputs a single prediction value.

Figure 4 shows that LightGBM, CatBoost, and NN outperform CarbonCast in 10 grids, with average MAPE improvements of 17.57%, 16.49%, and 15.84%, respectively. Their MAPE standard deviations (0.051, 0.0528, and 0.0566) are also lower than CarbonCast's (0.0574), indicating more consistent performance.

Takeaways. First, no single predictor outperforms across all grids due to each grid's unique characteristics. Second, the tested predictors mostly outperform CarbonCast in accuracy and provide more consistent results across grids.

3 EnsembleCI

Our previous analysis shows that designing a global model (i.e., CarbonCast in this study) for all grids does not achieve the best prediction accuracy for each grid due to the unique characteristics of each region, such as energy source mix and weather conditions. This highlights the necessity of an adaptive approach capable of selecting highly accurate methods tailored to each grid. To address

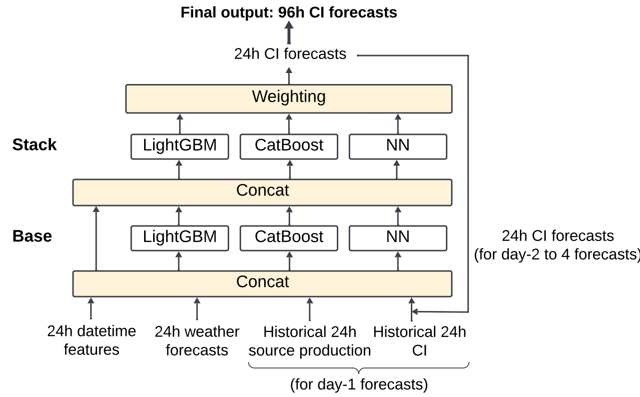

Figure 5: The architecture design of EnsembleCI.

this challenge, we present EnsembleCI, an end-to-end ensemble learning approach that leverages a pool of highly accurate sublearners and adaptively combines them with different weights to deliver the best predictions. The EnsembleCI design includes two key components: subleaner selection and ensembling strategy.

Sublearner selection. EnsembleCI is a general framework that can incorporate any machine learning models as sublearners. In this paper, we select LightGBM, CatBoost, and NN evaluated in §2.3 as they have proved their ability to outperform CarbonCast in most cases. Each sublearner is trained independently in the usual fashion before being combined in the ensemble. We do not include CarbonCast as a sublearner because it has a two-tier implementation where each tier is trained independently, being unable to seamlessly integrate into the end-to-end EnsembleCI design. We leave a uni-tier version of CarbonCast as future work.

Ensembling strategy. Various ensembling strategies, including bagging [6], boosting [19], stacking [33], and weighted combinations [23], have been introduced. Inspired by recent AutoML frameworks like `AutoGluon` [17], we adopt multi-layer stacking for EnsembleCI due to its ability to capture interactions between sublearners, enhancing final predictions [30]. Stacking trains a model using the aggregated predictions of the base models as its input features. Figure 5 illustrates the multi-layer stacking ensembling architecture of EnsembleCI, which includes two stages: base and stacking. In the base stage, EnsembleCI trains each sublearner individually using the raw features as input and outputs their corresponding predictions. In the stacking stage, each sublearner takes as input a concatenation of the raw features and the predictions generated by all sublearners in the base stage. To produce the final predictions, EnsembleCI uses ensemble selection [10] to aggregate the predictions from the stacking stage in a weighted manner.

Model Adaptability. The weights in EnsembleCI are optimized during training and remain static post-deployment. To adapt to evolving grid conditions, the model can be retrained periodically, ensuring that the weights reflect the latest energy source dynamics and weather patterns. This periodic retraining ensures that EnsembleCI remains adaptable to changing grid conditions, such as fluctuations in renewable energy generation, thereby improving the robustness of long-term predictions.

Implementation. We implement EnsembleCI on top of the `AutoGluon` [17] framework as it is a widely used AutoML framework to automate machine learning workflows. We implement data preprocessing and evaluation metrics with the `scikit-learn` package [29]. As shown in Figure 5, two separate models are trained for a 24-hour forecast horizon: one for the initial 24 hours (day 1) and another for the subsequent 24-hour periods (day 2–4). Each model consists of 24 sub-models, with each sub-model predicting the carbon intensity for one hour. The outputs from the Day 1 model, along with future weather forecasts and datetime features, serve as inputs to the Day 2–4 model. This model is applied recursively across three iterations to generate forecasts for the next 72 hours. Combining the outputs of both models provides a complete 96-hour carbon intensity forecast.

4 Evaluation

In this section, we outline our evaluation methodology, compare EnsembleCI and CarbonCast prediction results, and discuss key features influencing the final predictions.

4.1 Methodology

We evaluate a total of 11 regional grids, as detailed in Table 1, including 6 in the US and 5 in the EU. For a fair comparison, both CarbonCast and EnsembleCI are evaluated using CarbonCast v3.0 dataset. For US grids, training uses data from 2019 through the first half of 2022 and testing uses the second half of 2022. For EU grids, training uses data from 2019 through the first half of 2021, with testing on the second half of 2021. The input features include the historical 24-hour groundtruth carbon intensity values, historical 24-hour energy source production data (used only for day-1 forecasts), future 24-hour datetime features (sin/cos transformations of hour of day, day of year, and day of week), and 24-hour weather forecasts. As the goal is to evaluate the prediction accuracy of the models themselves, we remove the one-day source production predictions from external sources like OASIS [8] and ENTSO-E [16] for both CarbonCast and EnsembleCI. We use MAPE as the evaluation metric for prediction accuracy and take the average of five runs.

4.2 Prediction Results

Table 2 shows the prediction results for EnsembleCI and CarbonCast up to 4 days blue into the future across 11 grids. Overall, EnsembleCI achieves the lowest MAPE in almost all grids, except day-1 in DE. Averaged over all grids, EnsembleCI achieves lower MAPEs of 5.92%, 8.16%, 9.08%, and 9.56% for day-1 through day-4, respectively, compared to CarbonCast's averages of 7.05%, 9.62%, 11.12%, and 12.48%. This reflects relative improvements of 18.1%, 17.13%, 19.69%, and 23.4% across the respective horizons. This trend is evident in individual grids, such as CISO, where EnsembleCI achieves a MAPE of 6.67% on day-1, reducing the error by 31.32% compared to CarbonCast's 9.57%. Similarly, in the PJM grid, EnsembleCI achieves a 38.88% improvement on day-4, with a MAPE of 4.19% compared to CarbonCast's 6.85%.

Additionally, EnsembleCI demonstrates greater robustness for long-term forecasting. For instance, in DE, EnsembleCI's MAPE increases from 11.57% on day-1 to 14.23% on day-4, yielding a total increase of 2.66%. In contrast, CarbonCast's MAPE increases more sharply, rising from 10.45% on day-1 to 24.7% on day-4, yielding a total increase of 14.25%, approximately five times greater than that of EnsembleCI. This pattern holds across other grids, where the average MAPE increase for EnsembleCI is 3.63%, compared to 5.43% for CarbonCast. These results demonstrate that EnsembleCI is more reliable than CarbonCast for long-term forecasting.

Table 2: MAPE (%) of EnsembleCI and CarbonCast.

	EnsembleCI				CarbonCast [26, 27]			
Grid	day-1	day-2	day-3	day-4	day-1	day-2	day-3	day-4
CISO	**6.67**	**8.89**	**9.73**	**10.15**	9.57	13.07	14.58	15.87
PJM	**2.43**	**3.24**	**3.78**	**4.19**	3.25	4.76	5.96	6.85
EPE	**1.13**	**1.19**	**1.22**	**1.22**	1.68	1.74	1.75	1.76
ISNE	**3.74**	**5.1**	**5.63**	**5.94**	4.8	6.02	6.69	7.28
ERCOT	**7.69**	**10.3**	**10.86**	**11.23**	8.87	10.98	11.88	12.67
MISO	**4.61**	**6.64**	**7.65**	**7.91**	5.87	7.2	7.75	8.26
DE	11.57	**13.7**	**13.93**	**14.23**	**10.45**	14.86	18.69	24.7
SE	**6.78**	**9.17**	**10.2**	**10.37**	8.87	12.2	13.9	15.03
ES	**11.41**	**19.15**	**23.32**	**25.55**	13.93	20.05	23.86	25.81
NL	**5.68**	**7.73**	**8.68**	**9.33**	6.44	9.63	11.82	13.31
PL	**3.43**	**4.67**	**4.87**	**4.99**	3.87	4.9	5.39	5.78
Average MAPE	**5.92**	**8.16**	**9.08**	**9.56**	7.05	9.62	11.12	12.48
# Lowest MAPE	10	11	11	11	1	0	0	0

Table 3: Top-3 features for three grids with the highest average ranks for three grids. Green dots mark those also in the top-3 for individual sublearners.

	CISO			MISO			DE		
LightGBM	●	●		●	●		●	●	●
CatBoost	●	●		●	●	●	●	●	●
NN	●	●	●	●	●		●	●	
Top-3 Features	hist_solar	forecast _dswrf	hist_CI	hist_CI	hist_coal	hist_wind	hist_CI	hist_solar	forecast _dswrf

4.3 Interpreting Features

We calculate feature importance for EnsembleCI using permutation importance [5]. This method measures each feature's importance by assessing how prediction error changes when its values are randomly shuffled. We showcase three grids – CISO, MISO, and DE – that have unique energy sources. Table 3 lists the top-3 features for each sublearner in these grids. We observe strong alignment among sublearners in selecting the most important features, highlighting their robustness. In CISO and DE, "hist_solar" and "forecast_dswrf" dominate, underscoring the prevalence of solar energy and its temporal variability in these grids. Meanwhile, MISO prioritizes "hist_coal" and "hist_wind", reflecting its reliance on these energy sources. These results shed light on regional energy dynamics by analyzing models' feature importance.

5 Conclusion

We present EnsembleCI, an adaptive, end-to-end ensemble learning-based approach that outperforms CarbonCast in both accuracy and robustness for carbon intensity forecasting. By addressing regional variability through dynamic ensemble design, EnsembleCI offers a scalable, reliable solution for sustainable energy management.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Undergraduate Research Assistantship (URA) program of the Cheriton School of Computer Science at the University of Waterloo.

References

[1] [n. d.]. Electricity Maps. https://app.electricitymaps.com/map/.

[2] [n. d.]. WattTime. https://watttime.org/.

[3] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Udit Gupta, Manoj Chakkaravarthy, David Brooks, and Carole-Jean Wu. 2023. Carbon explorer: a holistic framework for designing carbon aware datacenters. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Volume 2*.

[4] US Energy Information Administration. [n. d.]. Real-time Operating Grid. https://www.eia.gov/electricity/gridmonitor/dashboard/electric_overview/US48/US48.

[5] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* (2010).

[6] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24 (1996), 123–140.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[8] California Independent System Operator. 2025. Open Access Same-Time Information System (OASIS). https://oasis.caiso.com/mrioasis/logon.do.

[9] Qiong Cao, Yihang Wu, Jia Yang, and Jing Yin. 2023. Greenhouse Temperature Prediction Based on Time-Series Features and LightGBM. (2023).

[10] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. 2004. Ensemble selection from libraries of models. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*.

[11] D. Chaturvedi and Isha Singh. 2016. Solar Power Forecasting: A Review. (07 2016).

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[13] Yi Ding and Tianyao Shi. 2024. Sustainable LLM Serving: Environmental Implications, Challenges, and Opportunities. In *2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC)*.

[14] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).

[15] Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev. 2017. Fighting biases with dynamic boosting. *arXiv preprint arXiv:1706.09516* (2017).

[16] ENTSO-E. 2015. ENTSO-E Transparency Platform. https://transparency.entsoe.eu/.

[17] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-Tabular: Robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505* (2020).

[18] Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Chukwunyere Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. LLMCarbon: Modeling the End-to-End Carbon Footprint of Large Language Models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

[19] Yoav Freund, Robert E Schapire, et al. 1996. Experiments with a new boosting algorithm. In *icml*, Vol. 96. Citeseer, 148–156.

[20] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture (ISCA)*.

[21] Walid A Hanafy, Qianlin Liang, Noman Bashir, David Irwin, and Prashant Shenoy. 2023. CarbonScaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)* (2023).

[22] Jeremy Howard and Sylvain Gugger. 2020. Fastai: a layered API for deep learning. *Information* (2020).

[23] D. Jimenez. 1998. Dynamically weighted ensemble neural networks for classification. In *IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, Vol. 1. 753–756 vol.1.

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.).

[25] Amy Li, Sihang Liu, and Yi Ding. 2024. Uncertainty-Aware Decarbonization for Datacenters. In *Workshop on Sustainable Computer Systems (HotCarbon)*.

[26] Diptyaroop Maji, Prashant Shenoy, and Ramesh K Sitaraman. 2023. Multi-Day Forecasting of Electric Grid Carbon Intensity Using Machine Learning. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys)*.

[27] Diptyaroop Maji, Prashant Shenoy, and Ramesh K Sitaraman. 2023. Multi-day forecasting of electric grid carbon intensity using machine learning. *ACM SIGENERGY Energy Informatics Review* 3, 2 (2023), 19–33.

[28] Sophia Nguyen, Beihao Zhou, Yi Ding, and Sihang Liu. 2024. Towards Sustainable Large Language Model Serving. In *The 3rd Workshop on Sustainable Computer Systems (HotCarbon)*.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research* (2011).

[30] Saeed Shafieian and Mohammad Zulkernine. 2023. Multi-layer stacking ensemble learners for low footprint network intrusion detection. *Complex & Intelligent Systems* (2023).

[31] Tianyao Shi, Yanran Wu, Sihang Liu, and Yi Ding. 2024. GreenLLM: Disaggregating Large Language Model Serving on Heterogeneous GPUs for Lower Carbon Emissions. *arXiv preprint arXiv:2412.20322* (2024).

[32] Abel Souza, Shruti Jasoria, Basundhara Chakrabarty, Alexander Bridgwater, Axel Lundberg, Filip Skogh, Ahmed Ali-Eldin, David Irwin, and Prashant Shenoy. 2023. CASPER: Carbon-Aware Scheduling and Provisioning for Distributed Web Services. In *Proceedings of the 14th International Green and Sustainable Computing Conference (IGSC)*. 67–73.

[33] Kai Ming Ting and Ian H. Witten. 1997. Stacking Bagged and Dagged Models. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*.

[34] U.S. Energy Information Administration. 2023. How much of U.S. carbon dioxide emissions are associated with electricity generation? https://www.eia.gov/tools/faqs/faq.php?id=77&t=11.

[35] World Nuclear Association. 2024. Carbon Dioxide Emissions From Electricity. https://world-nuclear.org/information-library/energy-and-the-environment/carbon-dioxide-emissions-from-electricity.

[36] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems (MLSys)* (2022).

[37] Yanran Wu, Inez Hua, and Yi Ding. 2025. Unveiling environmental impacts of large language model serving: A functional unit view. *arXiv preprint arXiv:2502.11256* (2025).