JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 11 Issue 02
2026
Research Article
https://doi.org/10.56979/1102/2026/1477

Bin-Wise Quantized CatBoost for Energy-Efficient Drug-Review Rating Prediction in Biomedical E-Commerce

Muhammad Ahmad Zia[1*], Mahnoor[1], Ali Hussain[1], Mohammed Abual-Rub[2], Najla Abdulaziz Almousa[2], and Ashraf Al-Shaikh Khalil[3]

[1]Department of Computer Science & IT, The University of Lahore, Lahore, 54600, Pakistan.
[2]Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia.
[3]Deanship of E-learning and Information Technology, King Faisal University, Al-Ahsa 31982, Saudi Arabia.
[*]Corresponding Author: Muhammad Ahmad Zia. Email: muhammad.zia@cs.uol.edu.pk

Received: March 30, 2026 **Accepted:** July 18, 2026

Abstract: Artificial intelligence is increasingly applied on biomedical e-commerce platforms such as online drug-review aggregators, and machine learning provides most of its practical capabilities. Ma-chine learning models for predicting patient-generated ratings on such platforms are typically optimized for accuracy alone, while the energy and carbon cost of training is rarely reported. This study examines whether bin-wise quantization in CatBoost. reducing the feature bin count. can lower the computational and carbon cost of training such models without degrading predictive accuracy. Two public drug-review datasets are used: the UCI Drugs.com corpus (50,000 reviews, rating 1–10) and the WebMD corpus (50,000 reviews, satisfaction 1–5). Each review is represented by a 384-dimensional sentence embedding of its text, eight text-length and style statistics, a sentiment score, and two tabular features. CatBoost is used as the primary model because it produced markedly more stable generalization than the other libraries, and its bin count is directly exposed to the practitioner. Five CatBoost-based models are first compared per dataset under five-fold cross-validation: Ridge regression, standard CatBoost, and CatBoost with the feature bin count fixed at 64, 128, and 255. Hyperparameters are held constant across CatBoost variants to isolate the effect of the bin count. Accuracy equivalence is assessed with two one-sided tests (TOST) at a predefined margin of ±0.01 R^2, and training time and CO_2 emissions are tracked with CodeCarbon. The test R^2 values of the bin-reduced and standard CatBoost are statistically equivalent across both datasets and all three bin counts (TOST, all $p < 0.001$), while 64-bin reduction lowers training time and CO_2 emissions by around 48%. To test whether the effect is specific to CatBoost, LightGBM and XGBoost are evaluated under the same protocol; bin reduction lowers their training cost by roughly 25–60% while remaining statistically equivalent in accuracy (TOST, all $p < 0.002$), con-firming that the effect is general to histogram-based gradient boosting. SHAP analysis shows the review-text embedding is the dominant predictor, with sentiment and engagement as secondary signals. The work is presented as a sustainability-oriented empirical study: explicit control of the bin-count parameter exposes a favorable accuracy–efficiency trade-off applicable to biomedical review marketplaces.

Keywords: Artificial Intelligence; Green Machine Learning; Gradient Boosting; CatBoost; Feature Bin Reduction; Carbon Footprint; Drug Review Prediction; Sentence Embeddings; Equivalence Testing; Biomedical E-Commerce

1. Introduction

Artificial intelligence is increasingly integrated into biomedical e-commerce, with machine learning providing most of its practical capabilities. Patient choice and assessment of medicines is more and more

being influenced by biomedical e-commerce, where machine learning models power the underlying recommendation and ranking systems. Drug-review aggregators like Drugs.com and WebMD gather millions of patient reviews and numerical satisfaction scores used to fuel features like recommendations, safety monitoring and quality rating. In this context, a representative supervised learning task is predicting ratings for a review from its content and associated metadata—and gradient-boosted decision trees are some of the best model families for tabular and mixed feature representations.

Training machine-learning models has become a known problem when it comes to computing costs. Strubell et al. [1] have demonstrated that the carbon emissions from training large natural-language models are comparable to the carbon produced by several cars during their entire lifespan, while Schwartz et al. [2] have introduced the concept of Green AI, which places equal importance on computational and ecological efficiency as well as accuracy. Patterson et al. [3] had quantified the carbon cost for large models, and Henderson et al. [4] had proposed a systematic approach for reporting energy and carbon footprints. This has been followed by a series of practical instrumentation, such as Lacoste et al. impact calculator [5], CodeCarbon by Schmidt et al. [6] and CarbonTracker by Anthony et al. [7]. In recent reviews [8, 9] an emerging field is described, which optimizes model design concurrently for accuracy and ecological cost.

Applied analytics studies have been slower to adopt this perspective. In the majority of rating-prediction studies and in the majority of healthcare applications of gradient boosting only the accuracy is reported and the energy or carbon cost of training is not made public. There is a gap in the case of tree-based gradient boosting, which is the dominant model in tabular machine learning, and is a significant proportion of production training. Feature quantization — discretizing continuous inputs into a finite number of bins before split finding — is a standard acceleration technique in histogram-based boosting and is performed internally by CatBoost [10, 11], LightGBM [12], and XGBoost [13]. The bin count is rarely exposed to practitioners as a deliberate control for managing the accuracy–energy trade-off, and its effect on cross-validated accuracy and tracked emissions has not been systematically characterized for biomedical review prediction.

This study investigates whether reducing the feature bin count lowers training time and carbon emissions without degrading predictive accuracy on drug-review rating prediction. Two drug-review datasets are used, each review represented by a dense sentence embedding of its text together with engineered text statistics and tabular features. CatBoost is adopted as the primary model and is examined in depth across bin counts, because it produced far more stable generalization than the other gradient-boosting libraries and because its bin count is directly exposed as a parameter. Accuracy equivalence between the bin-reduced and standard CatBoost is assessed with a formal equivalence test rather than by the absence of a significant difference. LightGBM and XGBoost are then evaluated under the same protocol to test whether the bin-count effect generalizes beyond CatBoost.

The contributions are as follows. The first is a cross-validated, leakage-controlled evaluation of feature bin reduction in CatBoost on two biomedical drug-review datasets, jointly reporting accuracy, training time, and tracked CO_2 emissions. The second is formal evidence of accuracy equivalence: two one-sided tests confirm that bin-reduced CatBoost is equivalent to standard CatBoost in test R^2 within a predefined ± 0.01 margin on all comparisons ($p < 0.001$), while 64-bin reduction lowers training time and CO_2 emissions by approximately 48%. The third is a cross-library experiment showing that the same bin-count reduction lowers the training cost of LightGBM and XGBoost by roughly 25–60% with no meaningful accuracy loss, so the effect is general to histogram-based gradient boosting rather than specific to CatBoost. The fourth is an interpretability analysis using SHAP that identifies the review-text embedding as the dominant predictor, with sentiment and engagement as secondary signals. The fifth is the explicit framing of the contribution as a sustainability-oriented empirical study: by externalizing the bin-count parameter, the work exposes a favorable accuracy–efficiency trade-off rather than proposing a new learning algorithm.

The remainder of the paper is organized as follows. Section 2 reviews related work in green machine learning, gradient boosting, and drug-review prediction. Section 3 describes the datasets, feature construction, models, evaluation protocol, and emission tracking. Section 4 reports results on accuracy equivalence, efficiency, and feature importance. Section 5 discusses implications, and Section 6 states limitations. Section 7 concludes.

2. Related Work

2.1. Green Machine Learning and Carbon Reporting

The awareness of environmental costs of machine learning was initiated by the study of Strubell et al. [1] who estimated carbon emissions from the training of large language models, a concern later broadened by Bender et al. [14] to encompass the wider environmental and societal costs of increasingly large models. Schwartz et al. [2] defined Green AI as research that quantifies efficiency in addition to accuracy, and García-Martín et al. [15] surveyed methods for estimating the energy consumption of machine-learning workloads; Patterson et al. [3] improved the calculation of emissions from the training of large neural networks. Henderson et al. [4] advocated in their published research for a standardized reporting of energy and carbon. It has now become possible with the help of tooling: Lacoste et al. [5] developed an emissions calculator, Schmidt et al. [6] created CodeCarbon to track emissions in-process, Anthony et al. [7] produced CarbonTracker for use during training, and Lannelongue et al. [16] released the Green Algorithms framework for standardized carbon-footprint estimation. Recent surveys [8,9] by Verdecchia and by Bolón-Canedo, respectively, bring together the field. This approach is used in the present work for tree-based gradient boosting, a model that has been less explored than large neural models, but is widely used in tabular and biomedical analytics.

2.2. Gradient Boosting and Feature Quantization

Building on the gradient boosting framework introduced by Friedman [17], gradient boosted decision trees continue to be the most powerful general solution for tabular data. The popular XGBoost was introduced by Chen and Guestrin [13] and LightGBM by Ke et al. [12] which adopted the histogram split finding method, and CatBoost by Prokhorenkova et al. [10] that used ordered boosting and native categorical features handling; the use of CatBoost across domains has been reviewed by Hancock and Khoshgoftaar [18]. The methods based on histograms are based on dividing each continuous feature into a small number of bins prior to the split finding, thus narrowing the search space per feature and speeding up training, CatBoost uses a default of 254 bins [11]. To conclude, Shi et al. [19] demonstrated that gradient values can be quantized in a few bits without causing much loss of accuracy during training. The results presented confirm that quantization is a known efficiency technique, but the number of bins is rarely considered as a control of the accuracy–energy trade-off that can be directly used by the practitioner, which is the subject of this work.

2.3. Drug-Review Rating Prediction

The Drugs.com review corpus introduced by Gräßer et al. [20] has become a standard benchmark for sentiment and rating analysis of patient drug reviews, and later work such as Al-Hadhrami et al. [21] and Colón-Ruiz and Segura-Bedmar [22] applied deep-learning models to the same task. Basiri et al. [23] further advanced this line by fusing deep and traditional classifiers for drug-review sentiment. Reported accuracy is typically modest and depends heavily on the review text, since numeric ratings of subjective patient experience are inherently noisy. CatBoost and related gradient-boosting methods have been ap-plied across biomedical informatics, including clinical risk modeling [24] and cardiovascular prediction [25], confirming their suitability for biomedical tabular tasks. The present study differs in objective: rather than maximizing rating-prediction accuracy, it characterizes the accuracy–energy trade-off of quantization on this representative biomedical e-commerce task.

3. Materials and Methods

The methodology follows a controlled experimental design in which Ridge regression and four CatBoost configurations, one standard and three with reduced bin counts, are compared on two biomedical drug-review datasets, under shared five-fold cross-validation with per-fit carbon tracking, and LightGBM and XGBoost are then evaluated under the same protocol to test whether the bin-count effect generalizes across libraries. The subsections below describe each component of the pipeline in turn: the datasets, the preprocessing and filtering steps, the dense feature representation, the models and hyperparameters, the evaluation protocol, equivalence and significance testing, and carbon-footprint measurement.

3.1. Datasets

Two public datasets for drug reviews are used. The Drugs.com corpus introduced by Gräßer et al. [20] and hosted at the UCI Machine Learning Repository includes patient reviews of drugs, each comprising a

free-text review, the medical condition they treat, a usefulness count, a date when the review was posted, and a numerical rating on a 1–10 scale. The WebMD corpus, available on Kaggle, consists of patient reviews along with a free-text review, condition, demographic fields, and satisfaction ratings from 1 to 5, the latter of which is the target for prediction. A third dataset, Druglib.com, was assessed during the development of the model but not included in the main analysis because of its small size, which created a large train-test gap and unstable cross-validation estimates; retaining it would have diminished rather than enhanced the evidence. Table 1 summarizes the two datasets used.

Table 1. Summary of the two biomedical drug-review datasets.

Dataset	Source	Reviews	Rating Scale	Target
Drugs.com	UCI ML Repository	50,000	1–10	rating
WebMD	Kaggle	50,000	1–5	satisfaction

3.2. Data Preprocessing and Filtering

Every corpus is loaded from its public release and transformed to a common schema of five fields: review text, condition, usefulness count, rating, and the declared rating scale. Each dataset is then sampled at random to 50,000 reviews with the same random seed, to keep runtimes within reasonable limits and to make the two corpora the same size for the analysis. Since the sentence-embedding model works on raw sentences without any preprocessing applied to them, no stemming, lemmatization, or stop-word removal is used. Reviews are not filtered by length, and no rows are dropped, so the sample remains representative of what the platforms contain.

The two corpora differ in a number of ways that are important for modeling. Drugs.com reviews are longer, with a median word count of 84 and a median of 453 characters, and all of the sampled reviews include text. WebMD reviews are shorter, with a median of 39 words, and 5,666 of the 50,000 sampled rows (11.3%) are missing a review body; these rows are retained so that the sample is not biased towards more expressive patients, although they carry little textual signal. The usefulness count is present for the majority of rows in the Drugs.com dataset, at 96.1%, with an average of 28 votes and a maximum of 949, while it is absent throughout the WebMD sample, where it is constant at zero and therefore has no impact. Drugs.com covers 693 distinct conditions, the most common being Birth Control at 17.9%, while WebMD covers a larger number at 1,199, the most common being the unspecified category Other at 13.4%. Two WebMD rows carry ratings outside the declared 1–5 range and are retained as recorded, since at 0.004% of the sample they cannot alter the results. These disparities are reported because they relate directly to the accuracy differences found between the two datasets.

3.3. Feature Construction

Each review is represented by dense continuous features. The review text is encoded into a 384-dimensional sentence embedding using the all-MiniLM-L6-v2 sentence-transformer model, applied to the raw review text. Eight text-length and style statistics are added: character count, word count, sentence count, average word length, exclamation count, question-mark count, uppercase ratio, and words per sentence. A sentiment polarity score and two tabular features complete the representation: the usefulness count and a frequency encoding of the medical condition. This gives a 395-dimensional vector per review (384 embedding, 8 text statistics, 1 sentiment, 2 tabular). The representation is dense and high-cardinality, which is the regime in which the bin-count parameter has a measurable effect; sparse text representations such as term-frequency vectors are near-discrete and largely insensitive to the bin count. The two tabular features behave differently across datasets: the usefulness count is informative on Drugs.com but constant on WebMD, so on WebMD it contributes nothing and the model relies on the remaining features. Sentence embeddings are computed once per review and are deterministic, so they introduce no dependence between cross-validation folds.

3.4. Models and Hyperparameters

The detailed bin-count analysis is carried out with CatBoost, and five models are tested on each dataset. Ridge regression is a linear baseline. The default quantization implemented in standard CatBoost is based on a histogram. Three bin-reduced versions explicitly set the CatBoost bin count to 64, 128, and 255, respectively. Strictly, CatBoost performs feature quantization internally with a default of 254 bins; the

present work investigates the effect of varying this bin count, so the three variants are referred to as bin-reduced CatBoost configurations throughout. CatBoost is chosen as the primary model not because it reaches the highest absolute accuracy, which it does not, but for three practical reasons. Its border_count parameter exposes the bin count directly to the practitioner, which is the exact control this study examines. Its ordered boosting and native categorical handling produced much tighter generalization than the other libraries in these experiments, with a train-to-test R^2 gap near 0.11 against roughly 0.33 to 0.39 for LightGBM and XGBoost on the same features. And it is already widely adopted in biomedical informatics. The only difference in the hyperparameters across any of the CatBoost variants is that of the bin count; they are all otherwise the same: 600 iterations, a depth of 6, learning rate of 0.05, L2 leaf regularization of 6.0, an early stopping round of 40 on a 10% internal validation split, and the same random seed. Per-dataset hyperparameter tuning is intentionally omitted to preserve this experimental control, as tuning could change the results for production deployments, but the relative ordering of the bin-reduced and standard variants is expected to remain the same. To confirm that the bin-count effect is not specific to CatBoost, LightGBM and XGBoost are evaluated under the same protocol in Section 4.5, each at its default bin setting and at 64 bins, with matched depth, learning rate, and iteration budget.

3.5. Evaluation Protocol

Five-fold cross validation is done for each dataset for the five models with shared fold split thus allowing paired comparison analysis. The StandardScaler is only trained on the training fold, and applied to the held-out fold, and embeddings are deterministic, hence no information leaks from test to train. There is a training-fold R^2, training-fold RMSE and training-fold MAE recorded for each fold and model, and a corresponding test-fold R^2, test-fold RMSE and test-fold MAE is always displayed, meaning that the generalization gap is visible at all times. Data is presented as mean ± SD (95% CI) across the folds.

3.6. Equivalence and Significance Testing

The accuracy claim is equivalence rather than improvement; it is assessed with two one-sided tests (TOST) on per-fold test R^2, comparing each bin-reduced variant against standard CatBoost. The equivalence margin is predetermined at ±0.01 R^2. This choice is justified on three grounds. It corresponds to one percentage point of explained variance, which is negligible for a rating-prediction task where the best models explain only 30 to 47 percent of the variance. It is smaller than the fold-to-fold standard deviation of test R^2 that is normally observed, which sits between 0.004 and 0.008 in these experiments, so the margin is stricter than the natural run-to-run variation. And it is fixed before the tests are run, so it does not depend on the observed differences. Paired t-tests and Wilcoxon signed-rank tests are also conducted and reported as ancillary tests; they are not used in the claim for accuracy, because the folds are not independent.

3.7. Carbon Footprint Measurement

The training time and CO_2 emissions for each of the per-fold model fits are recorded with CodeCarbon [6]. The tracker monitors power usage of the CPU and memory, as well as the duration of the process, to calculate the energy used, and then applies the carbon-intensity factor for the region to calculate the CO_2 emissions. The experiments were performed in one Google Colab CPU environment with no GPU, on an Intel Xeon processor with two virtual cores and about 13 GB of RAM, in Python 3 with CatBoost, LightGBM, XGBoost, scikit-learn, and CodeCarbon at their current stable releases. In order to keep the timing comparison between the different libraries fair, all three were measured in the same session on the same instance, as their time and emission figures depend on the hardware resources allocated. A default sampling interval was used to monitor emissions per fit. Since no energy counters are available at the hardware level, CodeCarbon resorts to a processor thermal design power estimation approach. Absolute mission values are therefore only approximate and are referred to as comparative values obtained under the same conditions rather than as precise physical measurements; relative reductions are treated as the major efficiency indicator. A summary of the methodology pipeline can be found in Figure 1.

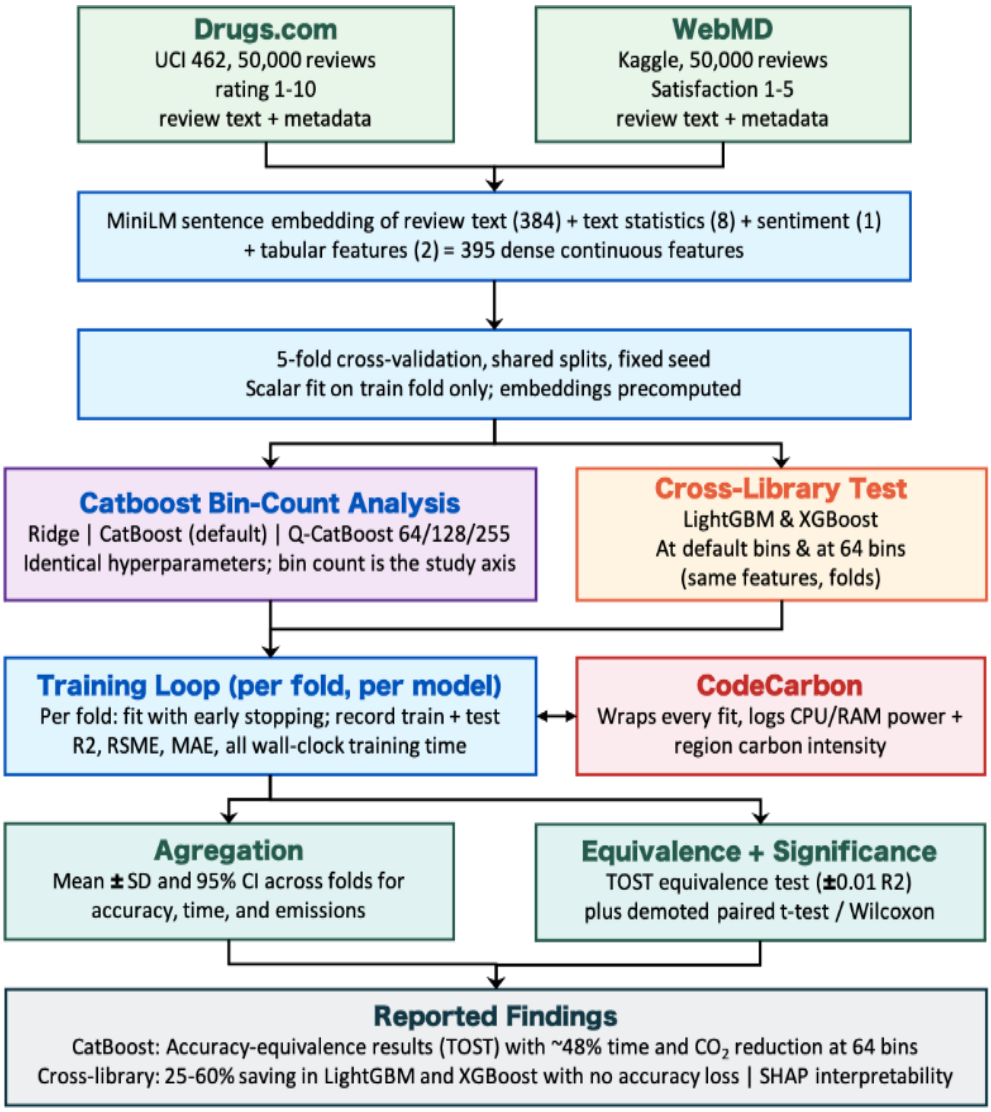

Figure 1. End-to-end methodology pipeline from data sources through dense feature construction, cross-validation, model training with carbon tracking, equivalence and significance testing, and reported findings.

4. Results

4.1. Predictive Accuracy

The values of cross-validated R^2 for the training set, cross-validated R^2 for the test set, the generalization gap, test-set RMSE, and test-set MAE for all five models are presented in Table 2 on both datasets. The ordering of the models remains the same throughout the datasets, with Ridge regression being the poorest of the lot, standard CatBoost providing a strong baseline, and the three quantized ones following it almost perfectly. On Drugs.com, standard CatBoost attains test $R^2 = 0.407$ (95% CI 0.401–0.413), and the 64-bin variant attains 0.408 (0.402–0.414). On WebMD, standard CatBoost attains 0.314 (0.304–0.324) and the 64-bin variant 0.315 (0.305–0.325). The generalization gap is around 0.11 on Drugs.com and 0.12 on WebMD for all CatBoost variants, which means that the over-fitting is moderate and controlled. The absolute R^2 scores remain relatively low, in line with the challenges of predicting subjective patient ratings even with review-text embeddings.

Table 2. Cross-validated accuracy (mean across folds) for all models on both datasets.

Dataset	Model	Train R^2	Test R^2	Gap	Test RMSE	Test MAE
Drugs.com	Ridge	0.381	0.368	0.013	2.601	2.104
Drugs.com	CatBoost	0.515	0.407	0.108	2.519	2.003
Drugs.com	Q-CB 64	0.516	0.408	0.108	2.517	2.000
Drugs.com	Q-CB 128	0.515	0.407	0.109	2.520	2.003

Drugs.com	Q-CB 255	0.515	0.406	0.109	2.522	2.005
WebMD	Ridge	0.319	0.304	0.015	1.346	1.127
WebMD	CatBoost	0.433	0.314	0.119	1.337	1.126
WebMD	Q-CB 64	0.433	0.315	0.118	1.336	1.125
WebMD	Q-CB 128	0.433	0.315	0.118	1.336	1.126
WebMD	Q-CB 255	0.432	0.314	0.118	1.337	1.126

4.2. Accuracy Equivalence

Table 3 reports the two one-sided tests comparing each quantized variant against standard CatBoost on per-fold test R^2. All six comparisons fall within the predefined ±0.01 R^2 equivalence margin with $p < 0.001$, so quantized CatBoost is statistically equivalent to standard CatBoost in predictive accuracy across both datasets and all three bin counts. The observed mean differences are between −0.0013 and +0.0009 R^2, an order of magnitude smaller than the equivalence margin. The supplementary paired t-tests and Wilcoxon tests are mostly non-significant, consistent with the absence of any practically meaningful accuracy difference; given the dependence between cross-validation folds, these tests are reported only as context.

Table 3. Two one-sided tests (TOST) for accuracy equivalence on per-fold test R^2.

Dataset	Comparison	Mean ΔR^2	Equivalence Margin	TOST p	Equivalent
Drugs.com	Q-CB 64 vs CatBoost	+0.0009	±0.01	<0.001	Yes
Drugs.com	Q-CB 128 vs CatBoost	−0.0004	±0.01	<0.001	Yes
Drugs.com	Q-CB 255 vs CatBoost	−0.0013	±0.01	<0.001	Yes
WebMD	Q-CB 64 vs CatBoost	+0.0007	±0.01	<0.001	Yes
WebMD	Q-CB 128 vs CatBoost	+0.0007	±0.01	<0.001	Yes
WebMD	Q-CB 255 vs CatBoost	+0.0002	±0.01	<0.001	Yes

4.3. Training Time and Carbon Emissions

Table 4 reports mean training time and CO_2 emissions with the relative reduction of each quantized variant against standard CatBoost. The 64-bin variant reduces training time by 48.4% (95% CI ±1.1) and emissions by 48.0% (±1.2) on Drugs.com, and by 48.0% (±1.5) and 47.7% (±1.5) respectively on WebMD. The 128-bin variant reduces both quantities by approximately 32% on both datasets, and the 255-bin variant is within 1% of standard CatBoost, as expected since it approaches the default bin count. The effect is therefore monotonic in the bin count and highly consistent across the two datasets. Figures 2 and 3 show the training-time and emission profiles per dataset, and Figures 4 and 5 show that test R^2 is flat across bin counts while the training cost falls.

4.4. Feature Importance

Figure 6 displays the SHAP feature-group importance for the Drugs.com CatBoost model. The review-text embedding is the major contributor by a wide margin, ahead of the sentiment score and usefulness count, while the condition frequency and the text-length statistics contribute least. The ordering is intuitive: the content of the review explains most of the predictable variation in the rating, while engagement and sentiment serve as secondary signals. The embedding block's aggregated absolute SHAP contribution is 6.44, which is an order of magnitude larger than the largest single engineered feature. Figure 7 shows the engineered features in more detail using dependence plots. The sentiment score is the strongest engineered feature, with a mean absolute SHAP of 0.60, and follows a smooth monotonic curve: strongly negative sentiment pushes the predicted rating down by up to two points, while positive sentiment pushes it up. The usefulness count (0.59) suggests a saturation curve, where each increase in helpfulness votes raises the predicted rating up to about fifty votes and then flattens out. The condition frequency (0.21) contributes more for rarer conditions, and the review word count (0.13) raises the predicted rating for reviews longer than approximately 130 words. This interpretation supports the validity of the model and explains why dense embeddings, rather than sparse text counts, are appropriate for this task. The usefulness count is constant on WebMD, so this engagement signal is not available there, which is one explanation for the slightly higher accuracy of the Drugs.com models.

Table 4. Training time and CO_2 emissions with relative reductions against standard CatBoost.

Dataset	Model	Time (s)	CO_2 (kg)	Δ Time	Δ CO_2
Drugs.com	CatBoost	100.1	2.2×10^{-4}	—	—
Drugs.com	Q-CB 64	51.6	1.2×10^{-4}	−48.4%	−48.0%
Drugs.com	Q-CB 128	67.9	1.5×10^{-4}	−32.2%	−31.9%
Drugs.com	Q-CB 255	99.4	2.2×10^{-4}	−0.7%	−0.8%
WebMD	CatBoost	99.9	2.2×10^{-4}	—	—
WebMD	Q-CB 64	51.9	1.2×10^{-4}	−48.0%	−47.7%
WebMD	Q-CB 128	67.9	1.5×10^{-4}	−32.0%	−31.8%
WebMD	Q-CB 255	99.3	2.2×10^{-4}	−0.6%	−0.6%

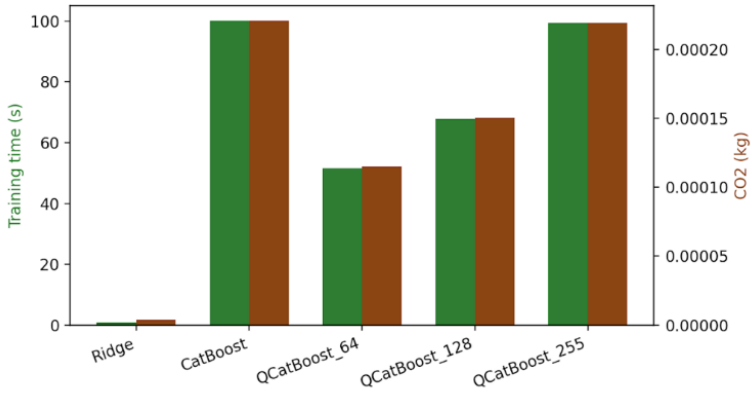

Figure 2. Drugs.com training time and CO_2 emissions across models.

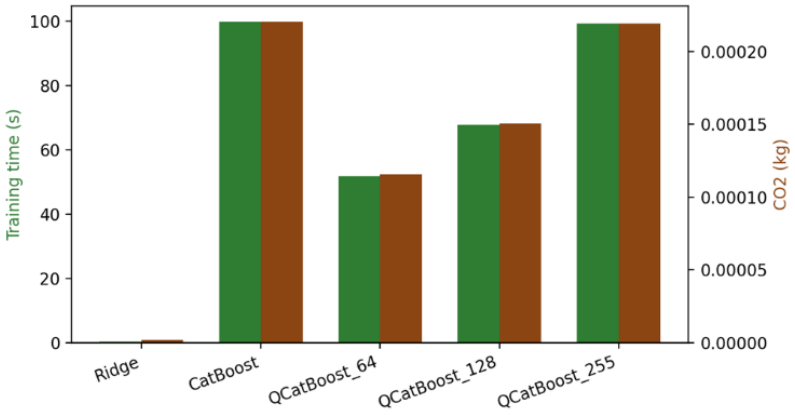

Figure 3. WebMD training time and CO_2 emissions across models.

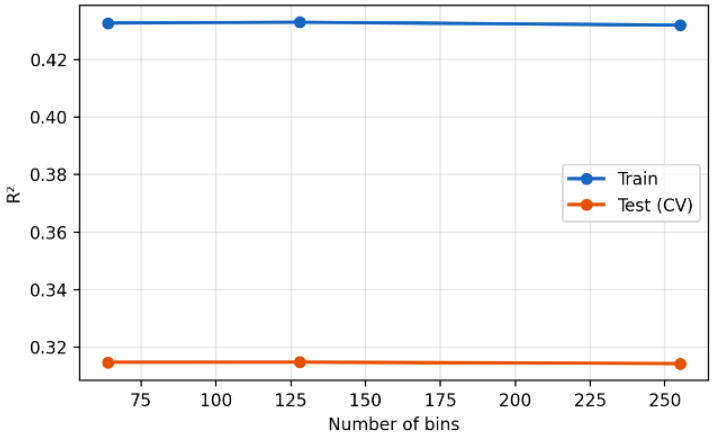

Figure 4. Drugs.com training and test R^2 across bin counts.

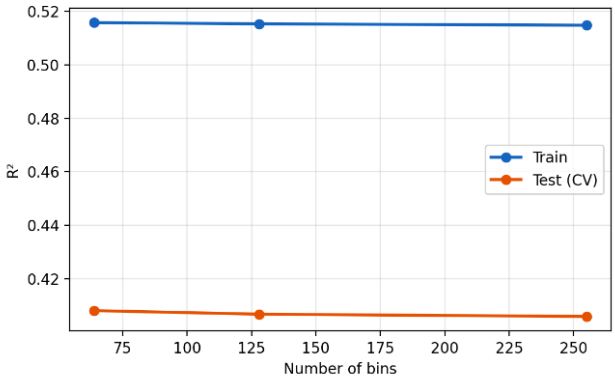

Figure 5. WebMD training and test R^2 across bin counts

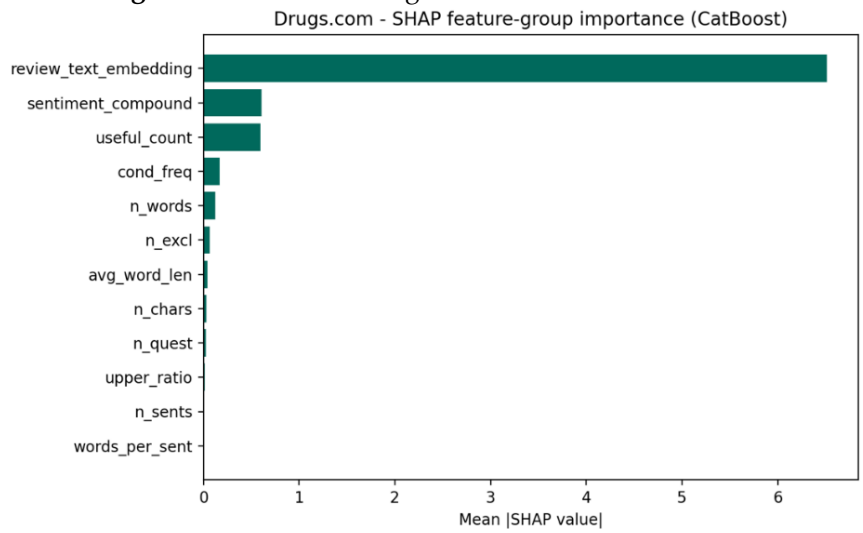

Figure 6. SHAP feature-group importance for the Drugs.com CatBoost model.

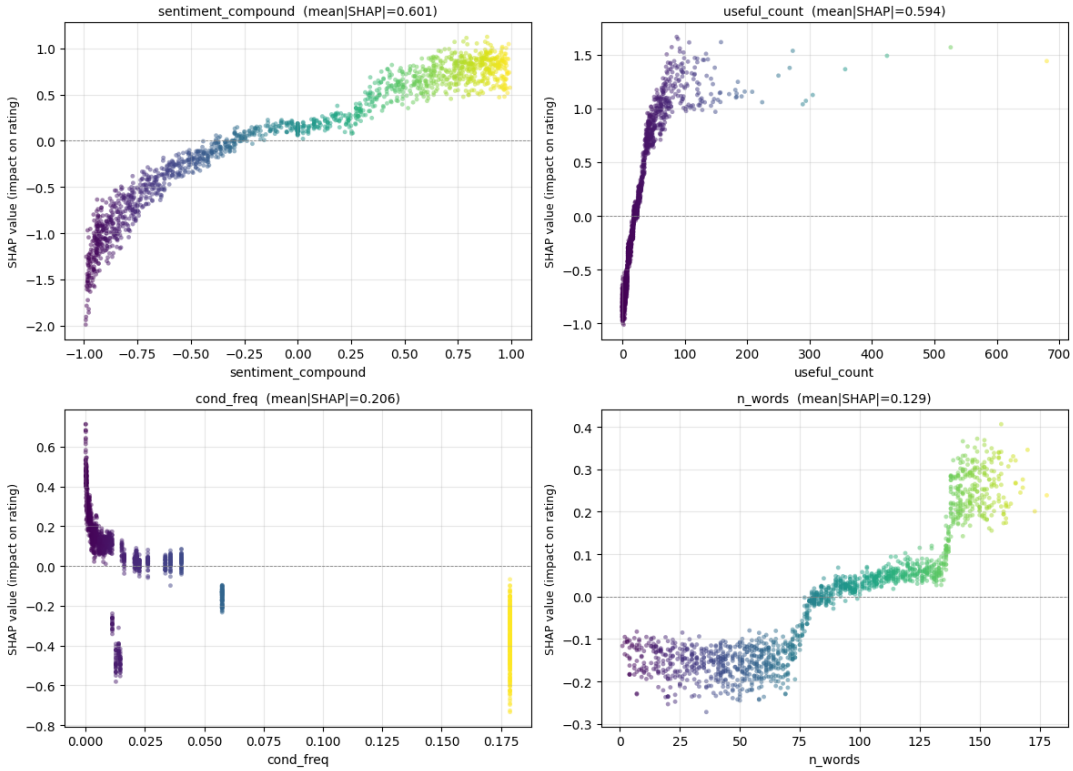

Figure 7. SHAP dependence plots for the top engineered features of the Drugs.com CatBoost model, showing how sentiment, usefulness count, condition frequency, and review length affect the predicted rating.

4.5. Cross-Library Generalization of the Bin-Count Effect

To verify if the accuracy–efficiency trade-off also applies to other libraries, LightGBM and XGBoost were also tested on the exact same dense features, on the same five folds, and with the same carbon tracking, and in the same session to ensure the timing and emission figures are comparable. The three libraries shown in Table 5 are at their default bin settings, as is the Ridge baseline. There are two patterns to be observed. First, LightGBM and XGBoost have a higher absolute test R^2 than CatBoost on both datasets, by roughly 0.05 to 0.06. The second is that this increased accuracy is accompanied by a much weaker level of generalization: the difference between the R^2 of the train and test sets for LightGBM and XGBoost is around 0.33 to 0.39, three times that of CatBoost (around 0.11 to 0.12). CatBoost therefore trades a modest amount of absolute accuracy for substantially more stable behavior, and that is the property that motivates its adoption as the key model.

Table 5. Cross-library comparison at default bin settings under identical five-fold cross-validation.

Dataset	Model	Test R^2	Test RMSE	Test MAE	Overfit Gap	Time (s)	CO_2 (kg)
Drugs.com	Ridge	0.368	2.601	2.104	0.013	0.3	1.0×10^{-6}
Drugs.com	CatBoost	0.410	2.514	1.998	0.111	101.9	1.1×10^{-4}
Drugs.com	LightGBM	0.467	2.388	1.862	0.327	85.8	9.2×10^{-5}
Drugs.com	XGBoost	0.470	2.383	1.855	0.338	175.0	1.9×10^{-4}
WebMD	Ridge	0.304	1.346	1.127	0.015	0.2	1.0×10^{-6}
WebMD	CatBoost	0.317	1.334	1.124	0.123	100.5	1.1×10^{-4}
WebMD	LightGBM	0.353	1.298	1.076	0.369	83.8	8.9×10^{-5}
WebMD	XGBoost	0.354	1.297	1.075	0.384	171.4	1.8×10^{-4}

Table 6 reports the effect of reducing the bin count to 64 in each library. The reduction lowers training time and emissions in every case: by about 46% for CatBoost, 27 to 29% for LightGBM, and 59 to 60% for XGBoost, with the differences reflecting how each library builds its histograms. Accuracy is unaffected in every case. Applying the same two one-sided tests used for CatBoost, all six library–dataset combinations are statistically equivalent to their default-bin counterparts within the ±0.01 R^2 margin (all $p < 0.002$), and in five of the six the reduction produced a small improvement of up to 0.004 R^2. The bin-count reduction is therefore not a CatBoost-specific artefact but a general lever for lowering the training cost of histogram-based gradient boosting, at no measurable ac-curacy cost on this class of task.

Table 6. Effect of reducing the feature bin count to 64 on training time, emissions, and test R^2 per library, with two one-sided tests (TOST) for accuracy equivalence at a ±0.01 R^2 margin.

Dataset	Library	Time (default→64)	Time Saving	CO_2 Saving	Δ Test R^2	TOST p	Equivalent
Drugs.com	CatBoost	101.9→55.1 s	45.9%	45.9%	−0.000	5.6×10^{-5}	Yes
Drugs.com	LightGBM	85.8→62.4 s	27.3%	25.0%	+0.002	1.2×10^{-3}	Yes
Drugs.com	XGBoost	175.0→70.3 s	59.8%	59.5%	+0.002	4.4×10^{-4}	Yes
WebMD	CatBoost	100.5→55.4 s	44.9%	44.9%	+0.001	2.0×10^{-4}	Yes
WebMD	LightGBM	83.8→59.2 s	29.4%	28.1%	+0.004	6.7×10^{-4}	Yes
WebMD	XGBoost	171.4→70.0 s	59.1%	58.6%	+0.003	6.1×10^{-4}	Yes

5. Discussion

The findings consistently indicate that, for biomedical drug-review rating prediction, reducing the CatBoost bin count to 64 halves training time and carbon emissions without materially affecting predictive accuracy. The accuracy claim does not rely on the absence of a significant difference, but on the formal equivalence test, which is the proper test for demonstrating that two configurations are equivalent and thus not significantly different in performance. The six bin-reduced versus standard comparisons all fall within the ±0.01 R^2 margin at $p < 0.001$.

Efficiency improvements decrease as the bin count increases, and the pattern repeats on two in-de-pendent datasets. The 64-bin configuration provides the best accuracy–efficiency trade-off among those tested, the 128-bin configuration sits in the middle, and the 255-bin configuration is not significantly

different from default CatBoost, as expected since it is close to the default bin count. The practical implication is that the bin count is already provided in all CatBoost installations and can be used as a direct control on the accuracy–energy tradeoff, while there is significant saving on this class of task with no measurable impact on accuracy.

The contribution is empirical and not algorithmic. This study does not propose a new method, as CatBoost, LightGBM, and XGBoost all quantize features internally [10, 12, 13]. Its importance is that the trade-off is established with formal equivalence testing under controlled and cross-validated conditions, and demonstrated not only for CatBoost but for all three libraries. This is made concrete in Section 4.5, where for every library and dataset the bin-reduced model lowered training time and emissions, by about 46% for CatBoost, 27 to 29% for LightGBM, and 59 to 60% for XGBoost, while remaining statistically equivalent in accuracy under the same equivalence test (all $p < 0.002$). The differences in savings are due to the differing ways that the libraries build and cache their histograms. In addition, there is a trade-off between accuracy and stability across the libraries: LightGBM and XGBoost reach a higher absolute test R^2 than CatBoost by about 0.05 to 0.06, but their gap between train and test R^2 is about three times larger than CatBoost's, which is why CatBoost is the primary model. These dynamics are not visible in default installations. The framing is located in an applied biomedical context, which is a lesser studied area within the broader body of knowledge on Green AI [1, 2, 8, 9].

SHAP analysis improves interpretability, which is especially valuable for biomedical informatics applications. The review-text embedding is the most prevalent indicator, ahead of sentiment and engagement, which suggests the model is based on review text rather than on spurious correlations. The dependence plots provide an interpretable clinical reading: the more positive the sentiment, the more helpfulness votes, and the longer the review, the more the predicted rating goes up. This further strengthens confidence in the equivalence result, as the dominant signal is the same in both models.

6. Limitations

There were a number of restrictions on the interpretation of these findings. The study is done on two datasets in one domain, drug reviews and ratings in the online marketplace, and the results cannot be directly generalized to other review sites such as RateMDs, Healthgrades, or forum-based patient communities, because of the differences in the length of reviews, the rating scales, and the presence or absence of engagement signals. The WebMD sample also shows this sensitivity, since roughly eleven percent of its reviews are empty and its usefulness count is absent, which probably explains why it is less accurate. A third dataset, Druglib.com, was excluded due to the volatility of the estimates on its smaller sample. There are five folds for cross-validation, leading to a limited resolution of uncertainty estimation; the confidence intervals should be considered as indicative only, and repeated cross-validation would narrow them. The equivalence test is formal, but it takes into account the results of the folds, which are not fully independent. Absolute emission values are not true physical measurements; they are reported relatively and rely on the cloud region and on thermal-design-power estimation without energy counters on the hardware. The hyperparameters remain fixed to isolate the bin-count effect, and tuned deployments could have different absolute accuracy, but the equivalence between bin-reduced and standard variants continues to be expected. Lastly, the cross-library experiment covers CatBoost, LightGBM, and XGBoost, but only with a single common feature representation and a single common hyperparameter budget; more comprehensive experiments with different representations and optimized settings are left to future research.

7. Conclusions

In this study, the authors investigated whether reducing the feature bin count can decrease the computational and carbon footprint of training gradient-boosting models for drug-rating prediction while preserving accuracy. On two biomedical drug-review datasets, bin-reduced CatBoost was statistically equivalent to standard CatBoost in test R^2 within a ± 0.01 margin defined by two one-sided tests (all $p < 0.001$), while 64-bin reduction lowered training time and CO_2 emissions by about 48%. The same reduction lowered the training cost of LightGBM and XGBoost by roughly 25 to 60% while remaining statistically equivalent in accuracy, showing that the effect is general to histogram-based gradient boosting rather than specific to CatBoost. Among the libraries, CatBoost gave the most stable generalization, which is why it

served as the primary case study, while LightGBM and XGBoost reached somewhat higher absolute accuracy at the cost of a much larger train-to-test gap. The bin count therefore acts as an effective and readily available control on the accuracy-energy trade-off, with the 64-bin configuration the most favorable among those tested. The work argues for routine joint reporting of accuracy and carbon cost in biomedical analytics. Future studies should extend to other biomedical domains, other tuned-hyperparameter regimes, and additional gradient-boosting libraries, and should use repeated cross-validation to further reduce uncertainties.

Funding: This article has been funded by the Deanship of Scientific Research in King Faisal University with Grant Number KFU263549.

Acknowledgments: The authors are thankful to King Faisal University for funding and supporting this research. The authors also acknowledge the developers of CodeCarbon and the sentence-transformers library, and the maintainers of the UCI Machine Learning Repository.

Data Availability Statement: The Drugs.com dataset is publicly available from the UCI Machine Learning Repository as introduced by Gräßer et al. [20]; the WebMD dataset is publicly hosted on Kaggle. All code, parameter settings, per-fold result tables, and figures supporting this study, including the cross-library experiment, are publicly available in the project repository [26] at https://doi.org/10.5281/zenodo.21455740. The repository contains the preprocessing, feature-building, cross-validation, cross-library, and equivalence-testing scripts, together with the per-fold result tables and figures, and will remain publicly accessible upon acceptance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; ACL: Florence, Italy, 2019; pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355
2. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. Communications of the ACM 2020, 63, 54–63. https://doi.org/10.1145/3381831
3. Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon Emissions and Large Neural Network Training. arXiv preprint 2021, arXiv:2104.10350. https://doi.org/10.48550/arXiv.2104.10350
4. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research 2020, 21, 1–43.
5. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the Carbon Emissions of Machine Learning. In NeurIPS Workshop on Tackling Climate Change with Machine Learning; 2019. arXiv:1910.09700. https://doi.org/10.48550/arXiv.1910.09700
6. Schmidt, V.; Goyal, K.; Joshi, A.; Feld, B.; Conell, L.; Laskaris, N.; Blank, D.; Wilson, J.; Friedler, S.; Luccioni, S. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing. Zenodo, 2021. https://doi.org/10.5281/zenodo.4658424
7. Anthony, L.F.W.; Kanding, B.; Selvan, R. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. arXiv preprint 2020, arXiv:2007.03051. https://doi.org/10.48550/arXiv.2007.03051
8. Verdecchia, R.; Sallou, J.; Cruz, L. A Systematic Review of Green AI. WIREs Data Mining and Knowledge Discovery 2023, 13, e1507. https://doi.org/10.1002/widm.1507
9. Bolón-Canedo, V.; Morán-Fernández, L.; Cancela, B.; Alonso-Betanzos, A. A Review of Green Artificial Intelligence: Towards a More Sustainable Future. Neurocomputing 2024, 599, 128096. https://doi.org/10.1016/j.neucom.2024.128096
10. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In Advances in Neural Information Processing Systems 31; Curran Associates: 2018; pp. 6638–6648. https://doi.org/10.48550/arXiv.1706.09516
11. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. arXiv preprint 2018, arXiv:1810.11363. https://doi.org/10.48550/arXiv.1810.11363
12. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30; Curran Associates: 2017; pp. 3146–3154.
13. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785
14. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21); ACM: Virtual Event, Canada, 2021; pp. 610–623. https://doi.org/10.1145/3442188.3445922
15. García-Martín, E.; Rodrigues, C.F.; Riley, G.; Grahn, H. Estimation of Energy Consumption in Machine Learning. Journal of Parallel and Distributed Computing 2019, 134, 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007
16. Lannelongue, L.; Grealey, J.; Inouye, M. Green Algorithms: Quantifying the Carbon Footprint of Computation. Advanced Science 2021, 8, 2100707. https://doi.org/10.1002/advs.202100707
17. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 2001, 29, 1189–1232. https://doi.org/10.1214/aos/1013203451
18. Hancock, J.T.; Khoshgoftaar, T.M. CatBoost for Big Data: An Interdisciplinary Review. Journal of Big Data 2020, 7, 94. https://doi.org/10.1186/s40537-020-00369-8
19. Shi, Y.; Ke, G.; Chen, Z.; Zheng, S.; Liu, T. Quantized Training of Gradient Boosting Decision Trees. In Advances in Neural Information Processing Systems 35; Curran Associates: 2022; pp. 18822–18833.
20. Gräßer, F.; Kallumadi, S.; Malberg, H.; Zaunseder, S. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health; ACM: 2018; pp. 121–125. https://doi.org/10.1145/3194658.3194677
21. Al-Hadhrami, S.; Vinko, T.; Al-Hadhrami, T.; Saeed, F.; Qasem, S.N. Deep Learning-Based Method for Sentiment Analysis for Patients' Drug Reviews. PeerJ Computer Science 2024, 10, e1976. https://doi.org/10.7717/peerj-cs.1976

22. Colón-Ruiz, C.; Segura-Bedmar, I. Comparing Deep Learning Architectures for Sentiment Analysis on Drug Reviews. Journal of Biomedical Informatics 2020, 110, 103539. https://doi.org/10.1016/j.jbi.2020.103539
23. Basiri, M.E.; Abdar, M.; Cifci, M.A.; Nemati, S.; Acharya, U.R. A Novel Method for Sentiment Classification of Drug Reviews Using Fusion of Deep and Machine Learning Techniques. Knowledge-Based Systems 2020, 198, 105949. https://doi.org/10.1016/j.knosys.2020.105949
24. Zhang, S.-Z.; Ding, H.-Y.; Shen, Y.-M.; et al. Harness Machine Learning for Multiple Prognoses Prediction in Sepsis Patients: Evidence from the MIMIC-IV Database. BMC Medical Informatics and Decision Making 2025, 25, 152. https://doi.org/10.1186/s12911-025-02976-y
25. Hamid, M.; Hajjej, F.; Alluhaidan, A.S.; bin Mannie, N.W. Fine tuned CatBoost Machine Learning Approach for Early Detection of Cardiovascular Disease through Predictive Modeling. Scientific Reports 2025, 15, 31199. https://doi.org/10.1038/s41598-025-13790-x
26. Zia, M.A.; Mahnoor; Hussain, A.; Abual-Rub, M.; Al-mousa, N.A.; Al-Shaikh Khalil, A. Bin-Wise Quantized CatBoost for Energy-Efficient Drug-Review Rating Predic-tion in Biomedical E-Commerce: Reproducibility Repository. Zenodo, 2026. https://doi.org/10.5281/zenodo.21455740