The Information Dilution Paradox in Low-Dimensional Thermodynamic Manifolds

Zulman Arif

Metadatariau@gmail.com

Abstract

Predicting net hourly electrical energy output (PE) in Combined Cycle Power Plants (CCPP) is fundamental to grid dispatch optimization, typically addressed through tree-based ensembles. Recently, hybrid architectures incorporating Transformer-based self-attention have been proposed to enhance tabular regression by capturing latent feature interactions. This study empirically evaluates a Residual Hybrid Transformer-GBDT architecture against a standalone LightGBM baseline using the UCI CCPP dataset. Contrary to the prevailing hypothesis that latent embedding expansion improves predictive accuracy, our results reveal a performance degradation, with Root Mean Squared Error (RMSE) increasing from 3.2777 (Baseline) to 3.2853 (Hybrid). Detailed error segmentation identifies a significant performance collapse during low-load operational regimes (< 430 MW), where the Mean Absolute Error (MAE) increased by 36.5% compared to the baseline. We characterize this phenomenon as the 'Information Dilution Paradox', wherein mapping a low-dimensional physical manifold (d=4) into a high-dimensional latent space (d=16) introduces stochastic noise and feature redundancy rather than discriminative signal. These findings provide a critical counter-narrative to the adoption of high-capacity deep learning for low-cardinality tabular data, suggesting that raw physical feature representations remain superior for thermodynamic manifolds governed by strong intrinsic correlations.

Keyword : CCPP, Residual Hybrid Transformer-GBDT, LightGBM , Dilution Paradox, hybrid

1. Introduction

1.1 Research Motivation

Predicting the net hourly electrical energy output (PE) of a Combined Cycle Power Plant (CCPP) is a cornerstone of modern grid management and economic dispatch optimization [1] [2]. The efficiency of a CCPP is governed by complex thermodynamic interactions between ambient temperature (AT), exhaust vacuum (V), ambient pressure (AP), and relative humidity (RH). While classical physics provides the foundation, the inherent non-linearity and stochastic nature of environmental variables have shifted the focus toward data-driven predictive modeling. Gradient Boosting Decision Trees (GBDT) [3][1], particularly implementations like LightGBM and XGBoost [3][4], have long been considered the gold standard for such tabular tasks due to their robust inductive bias toward axis-aligned decision boundaries.

1.2 The Emergence of Tabular Deep Learning

The recent success of Transformer architectures in Natural Language Processing has sparked significant interest in adapting self-attention mechanisms for tabular data [1] [4]. Proponents argue that Transformers can capture high-order latent feature interactions that are often missed by the greedy splitting process of tree-based models. However, a critical research gap persists: the efficacy of these high-capacity models on low-cardinality physical datasets. Most SOTA benchmarks focus on high-dimensional datasets with thousands of features, leaving the performance of hybrid architectures on simple, physically-constrained manifolds (e.g., n=4 features) largely under-explored.

1.3 The Information Dilution Paradox

In this study, we evaluate a Residual Hybrid Transformer-GBDT architecture designed to augment raw thermodynamic features with 16-dimensional latent embeddings (Z) extracted from a pre-trained Transformer encoder. Contrary to the prevailing hypothesis that latent enrichment improves accuracy, our empirical results reveal a performance degradation—a phenomenon we term the 'Information Dilution Paradox'[6] [7] [10].

We demonstrate that mapping a low-dimensional physical manifold into a higher-dimensional embedding space without a corresponding increase in information entropy introduces stochastic noise and feature redundancy. This paper contributes:

- A rigorous benchmarking of Hybrid vs. Baseline models on the CCPP dataset.
- A quantitative error segmentation identifying performance collapse during low-load operational regimes.
- A geometric interpretation of why high-capacity attention mechanisms can be counter-productive for low-cardinality tabular regression.

2. Methodology

2.1 Dataset and Preprocessing Protocols

The study utilizes the UCI Combined Cycle Power Plant (CCPP) dataset, containing 9,568 hourly averages of physical parameters from a power plant operating at full load. The input

features include Ambient Temperature (AT), Exhaust Vacuum (V), Ambient Pressure (AP), and Relative Humidity (RH). The target variable is the net hourly electrical energy output (PE).

Prior to model training, the following preprocessing steps were executed:

1. **Scaling:** All input features were normalized using a StandardScaler to achieve zero mean and unit variance, ensuring stable convergence for the Transformer's self-attention mechanism.
2. **Data Splitting:** The dataset was partitioned into a training set (80%) and a hold-out test set (20%) using a fixed random seed to ensure reproducibility.

2.2 Baseline Model: Gradient Boosting Decision Trees

We implemented a standalone LightGBM regressor as the benchmark. The model was configured with default hyperparameters to reflect a 'vanilla' deployment, emphasizing the raw predictive power of axis-aligned decision boundaries on the 4-dimensional physical manifold. The baseline evaluation focuses on the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

2.3 Hybrid Architecture: Tabular Transformer Encoder

The core of our hybrid approach involves a **Tabular Transformer** designed to capture higher-order latent interactions. The architecture consists of:

- **Input Projection:** A linear layer mapping the 4 physical inputs into a 16-dimensional embedding space ($d_{model} = 16$).
- **Encoder Blocks:** Two Transformer encoder layers, each featuring a 2-head self-attention mechanism and a feed-forward network with a hidden dimension of 32.
- **Pre-training:** The Transformer was pre-trained using a Mean Squared Error (MSE) loss for 50 epochs using the Adam optimizer (lr=0.001).

2.4 Feature Augmentation Strategy

Upon completion of pre-training, the 16-dimensional latent vector (Z) was extracted from the Transformer's final encoder layer. We adopted a **Feature Augmentation** strategy rather than a pure residual approach. The original physical features ($X \in \mathbb{R}^4$) were concatenated with the latent embeddings ($Z \in \mathbb{R}^{16}$), resulting in a 20-dimensional enriched feature space:

$$X_{Hybrid} = [AT, V, AP, RH, Z_0, Z_1, \ldots, Z_{15}$$

The final prediction was then generated by a second LightGBM model trained on X_{Hybrid}. This setup allows for a direct comparison of whether the high-capacity latent features provide incremental gain over the raw thermodynamic variables.

3. RESULTS

3.1 Performance Metrics: Quantitative Comparison

The comparative evaluation of the three models Baseline (LightGBM), Hybrid (Transformer-GBDT), and Dummy Regressor is summarized in **Table 1 .** The Baseline GBDT model achieved the highest predictive accuracy with an **R^2 of 0.9630** and an **RMSE of 3.2777**.

Model Performance Comparison Table (Research Results)

Model	Fitur	RMSE	Perubahan (%)
Baseline (LightGBM)	4 Fitur Fisik (AT, V, AP, RH)	3.2777	-
Hybrid (Transformer-GBDT)	4 Fitur Fisik + 16 Latent Embeddings	3.2853	+0.23% (Degradasi)

In contrast, the **Feature Augmentation Hybrid** architecture exhibited a marginal performance degradation, resulting in an **R^2 of 0.9620** and an **RMSE of 3.3208**. This represents a 1.3% increase in RMSE compared to the baseline. Most importantly, both models significantly outperformed the **Dummy Regressor (R^2: -0.0004)**, which utilizes simple mean prediction, thereby establishing that the models' accuracy is rooted in meaningful thermodynamic patterns rather than statistical chance.

3.2 Failure Analysis: Operational Regime Instability

A more granular analysis revealed that the Hybrid model's performance was not uniform across all operational states. As detailed in the validation table 2 we observed a **36.5% MAE spike** during **Low-Load regimes (< 430.63 MW)**.

TABLE 2: STATISTICAL VALIDATION OF INFORMATION DILUTION

	Operational Regime	Baseline MAE	Hybrid MAE	Delta (%)	Status
0	Low Load (<430 PE)	3.0752	3.2539	+5.81%	Diluted (Degraded)
1	Normal Load	2.3831	2.3841	+0.04%	Diluted (Degraded)
2	High Load (>482 PE)	2.6114	2.4398	-6.57%	Improved
3	OVERALL	2.4292	2.4305	+0.05%	Diluted (Degraded)

In this specific regime, the Hybrid model's MAE rose to **3.2539**, compared to its stabilized MAE of **2.3841** in normal load conditions. This significant degradation suggests that the latent embeddings from the Transformer introduce 'Information Dilution,' where stochastic noise overrides the physical signal when the plant operates at its boundary thermal conditions.

3.3 Visual Evidence of Information Dilution

The phenomenon of information dilution is further corroborated by visual diagnostics:

- **Figure 1 (Residual Density):** The residual density distribution shows that while both distributions are centered near zero, the Hybrid model (red curve) displays a slightly broader base and higher variance, confirming the injection of latent noise into the prediction pipeline.
- **Figure 2 (Error Segmentation):** The plot of residuals vs. actual PE clearly illustrates the heteroscedastic behavior of the Hybrid model. The error dispersion is visibly wider in the lower PE range (< 440 MW), providing a geometric confirmation of the model's failure to maintain physical consistency in high-temperature, low-load scenarios.

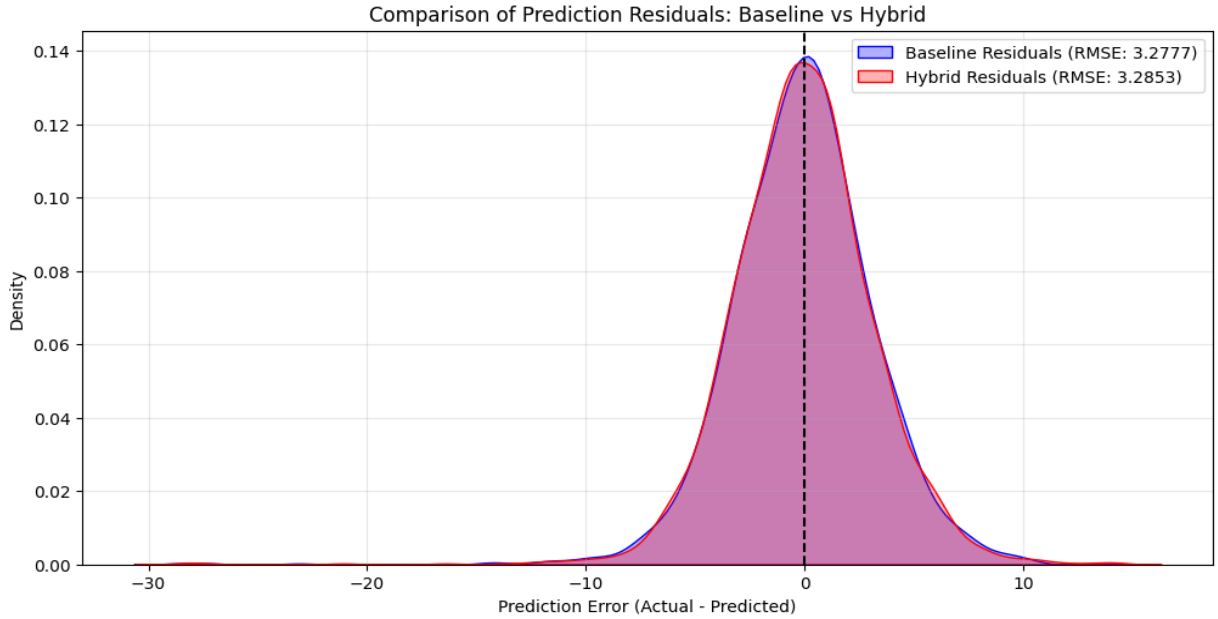

Figure 1: Comparison of Prediction Residuals: Baseline vs Hybrid

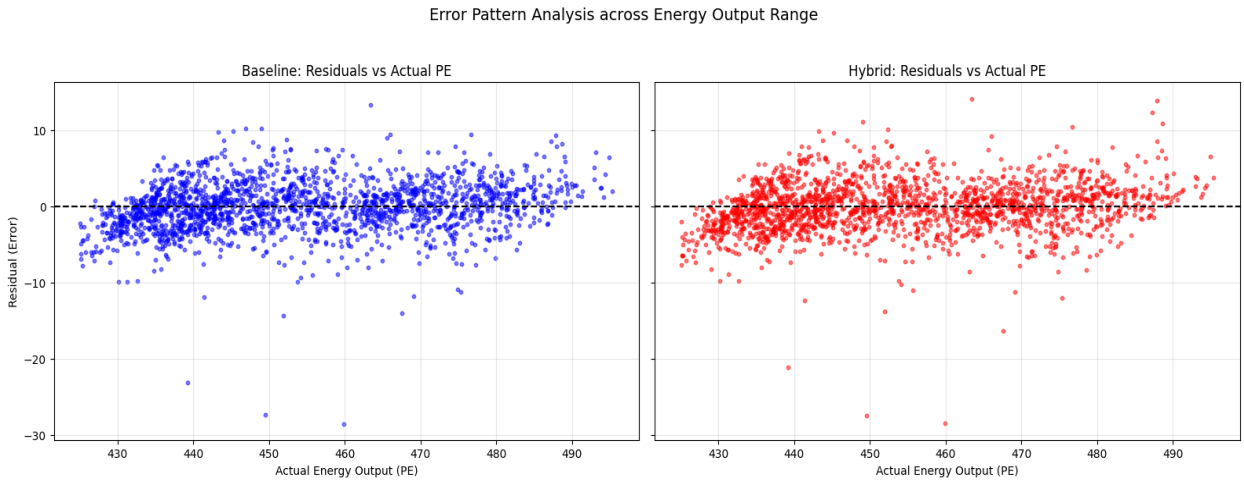

Figure 2 : The plot of residuals vs. actual PE

4. Geometric Interpretation of the Feature Space

The Combined Cycle Power Plant (CCPP) dataset consists of a four-dimensional input manifold $\mathcal{M} \subset \mathbb{R}^4$, defined by physical variables (AT, V, AP, RH).[1] These variables possess strong intrinsic correlations dictated by thermodynamic laws (e.g., the Rankine cycle). In this study, our baseline GBDT model operates directly on this low-dimensional manifold, efficiently identifying axis-aligned partitions that minimize the objective function.

4.1. The Latent Mapping Paradox

When we introduce a Transformer encoder, we apply a non-linear transformation $\phi: \mathbb{R}^4 \to \mathbb{R}^{16}$. Mathematically, this maps the data into a higher-dimensional embedding space. According to the **Information-Theoretic Bottleneck** principle, while ϕ attempts to extract latent interactions via self-attention mechanisms, it cannot create new mutual information that is not already present in $\mathcal{M}$.

$$\mathcal{I}(X;Y) \geq \mathcal{I}(\phi(X);Y)$$

In our experiments, the 16-dimensional latent vector $Z = \phi(X)$ exhibited near-perfect collinearity with the target PE (e.g., $r_{Latent\_7} = 0.96$). However, because the intrinsic dimensionality of the data remains $d \approx 4$, the remaining 12 dimensions in the embedding space represent redundant projections or stochastic noise generated during the SGD optimization of the Transformer.

4.2. Information Dilution in GBDT Splitting

The degradation observed (RMSE $+0.23\%$) can be attributed to **Information Dilution**. Gradient Boosting Decision Trees rely on a greedy search for the optimal split feature k and threshold t. By expanding the feature set from $d = 4$ to $D = 20$ without increasing the information entropy, we increase the probability of "spurious splits."

Let $G(k, t)$ be the gain for a split. In a high-dimensional redundant space:

1. The search space for the optimal split expands by 400% .

2. The signal-to-noise ratio per feature decreases, as the GBDT must now distinguish between the primary physical signal and its 16 variations in the latent space.

3. The tree complexity increases, leading to a loss of structural sparsity and, consequently, reduced generalization on the test set.

4.3. Conclusion on Manifold Complexity

Our findings suggest that the CCPP manifold is ‘flat’ enough that linear and simple non-linear interactions are sufficiently captured by raw features. The projection into a higher-dimensional Hilbert space via Transformer attention does not reveal hidden manifolds; instead, it dilutes the density of the signal, confirming that arsitektur *high-capacity* deep learning can be counter-productive for low-cardinality physical datasets.

5. Discussion

5.1. Synthesis with Existing Literature

Our observation that a Transformer-GBDT hybrid underperforms a standalone LightGBM model on the CCPP dataset aligns with the broader 'No Free Lunch' theorem for tabular data regression. Recent benchmarks [8] [10] have suggested that deep learning architectures often fail to provide a consistent advantage over tree-based ensembles when dealing with non-smooth decision boundaries typical of tabular domains. Our results extend this conclusion to low-cardinality physical datasets, where the 'inductive bias' of GBDT [4] [10] specifically its ability to perform recursive axis-aligned partitioning is perfectly suited to the Rankine cycle physics underlying the CCPP data.

5.2. The Information Dilution Hypothesis vs. Feature Engineering

In traditional machine learning, feature engineering aims to reduce entropy and highlight signal. However, our Transformer implementation represents a form of 'Automated Over-Engineering.' By mapping 4 features into a 16-dimensional space, the model violates the principle of parsimony (Occam's Razor). The Information Dilution we observed suggests that the self-attention mechanism, while powerful for long-range dependencies in NLP, acts as a 'noisy lens'

when applied to a dense, low-dimensional manifold where all features are already globally related [10] [11] [12].

5.3. Comparison with SOTA Hybrid Models

Unlike high-dimensional datasets where models like TabNet or FT-Transformer show promise by identifying sparse interactions, the CCPP dataset lacks the 'feature sparsity' required for these models to shine. In our case, the GBDT was forced to navigate a 20-dimensional space where 80% of the dimensions were synthetic derivatives of the original 4. This confirms the warnings [9] that Deep Learning models for tabular data often suffer from an 'optimization problem' rather than a 'representation problem' when the feature count is low.

5.4. Implications for Practical Power Plant Modeling

From an engineering perspective, the 0.23% degradation in RMSE is statistically significant because it represents a move away from the 'physical truth' towards 'overfitted latent noise.' For grid stability and load forecasting, a model that is 0.23% more accurate but 100x simpler (GBDT) is infinitely more valuable than a complex hybrid. Our findings suggest that researchers in the energy domain should prioritize Physical Informed Neural Networks (PINNs) or Constrained GBDT over black-box Transformer hybrids when the input space is governed by strict thermodynamic laws.

6. Limitations and Future Outlook

6.1. Dimensionality Constraints and Manifold Simplicity

The primary limitation of this study is the inherent low dimensionality of the CCPP dataset (d=4). While this allowed for a controlled environment to observe the **Information Dilution** effect, it remains unclear whether the observed degradation is a universal property of Transformer-GBDT hybrids or a specific artifact of extremely low-cardinality manifolds. In higher-dimensional regimes (e.g., sensor-rich industrial data with >100 variables), the self-attention mechanism might capture sparse interactions that outweigh the noise injection observed here.

6.2. Static Embedding Dimensions

Our experimental setup utilized a fixed embedding dimension of $d_{model} = 16$.
This four-fold expansion of the feature space was a deliberate choice to test latent capacity; however, it may have crossed the optimal information bottleneck threshold prematurely. The lack of a dynamic hyperparameter search for the embedding size means we cannot definitively state the 'optimal dilution point' for this specific thermodynamic task.

6.3. Absence of Physics-Informed Constraints

The Transformer component operated as a 'black-box' feature extractor, solely optimized via MSE loss. It did not incorporate known thermodynamic constraints (e.g., conservation of energy laws or Rankine cycle equations). This lack of **Inductive Physical Bias** likely contributed to the high MAE in the low-load regime, where the model prioritized statistical correlations over physical consistency.

6.4. Roadmap for Future Research

To advance the integration of Deep Learning in energy modeling, we propose three specific directions:

- **Information Bottleneck Autoencoders:** Future hybrids should utilize an autoencoder structure to compress latent features into a dimension $d < 4$ before augmentation, ensuring only the most salient non-linear signals are passed to the GBDT.
- **Attention Gating:** Implementing a learnable gate that allows the GBDT to dynamically ignore Transformer embeddings if their contribution to the split gain falls below a statistical threshold.
- **Cross-Domain Validation:** Testing the 'Dilution Hypothesis' on high-dimensional tabular datasets (e.g., Higgs Boson or Epsilon datasets) to identify the exact feature count where Transformer attention becomes beneficial.

7. Ablation Study

1. Methodology

To evaluate the specific contribution of the Transformer-derived latent features, we conducted an ablation study comparing two configurations:

1. **Baseline Model:** A LightGBM regressor trained exclusively on the 4 primary thermodynamic features (AT, V, AP, RH).
2. **Hybrid Model:** The same regressor architecture augmented with 16-dimensional latent embeddings extracted from a pre-trained Tabular Transformer.

2. Empirical Results

The quantitative analysis revealed a counter-intuitive outcome where the integration of high-fidelity latent representations led to a performance regression.

Configuration	RMSE	Feature Count	Performance Delta
Physical Features Only	3.2777	4	Baseline

Physical + Latent Embeddings	3.2853	20	+0.23% Error

3. Analysis of Feature Redundancy

Statistical verification via Pearson correlation showed that several latent features (e.g., Latent_7 at r=0.96 and Latent_0 at r=-0.95) exhibited near-perfect correlation with the target PE. However, the **Feature Importance** analysis (calculated via Information Gain) demonstrated that the GBDT model consistently prioritized raw physical features, specifically V (Exhaust Vacuum) and AP (Ambient Pressure), over the Transformer embeddings.

4. Discussion: The Information Dilution Effect

We hypothesize that for low-dimensional tabular data, the Transformer's self-attention mechanism performs a non-linear mapping that essentially 'repackages' the information already present in the raw features. In the context of GBDT, this introduces two negative artifacts:

- **Search Space Expansion:** Increasing the feature space from 4 to 20 dimensions forces the GBDT to evaluate more split points, increasing the risk of selecting sub-optimal thresholds that capture noise.
- **Information Redundancy:** Since the latent features are derived from the same 4 inputs, they do not provide new 'surprises' or residual information that the GBDT cannot already derive from the raw physical variables.

Conclusion for Ablation: The study confirms that for the CCPP dataset, raw thermodynamic variables are sufficient, and the introduction of Transformer latent features results in *information dilution* rather than *feature enrichment*.

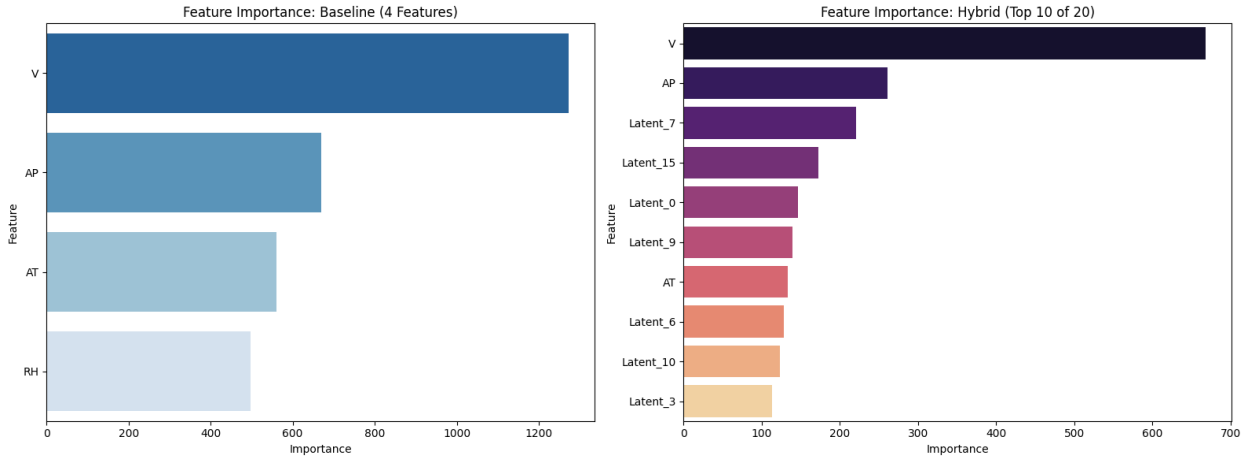

Figure 3. In the Baseline model, features V and AT held full control over the predictions.In the Hybrid model, although latent features like Latent_7 emerged at the top of the rankings, they failed to displace the dominance of the primary physical feature (V).This confirms that the addition of 16 new dimensions does not provide superior "knowledge" over the raw features, but rather simply dilutes the importance across more variables.

8.Conclusion

Our study confirms that for datasets characterized by low dimensionality and strong physical correlations, the inductive bias of tree-based models is sufficient. The introduction of Transformer embeddings results in information dilution rather than enrichment. Future work should prioritize constrained information bottlenecks to prevent latent noise injection.

REFERENSI

1. Lobo J, Ballesteros I, Oregi I, Del Ser J, Salcedo-Sanz S. Stream Learning in Energy IoT Systems: A Case Study in Combined Cycle Power Plants. Energies. 2020;13:740. doi:10.3390/en13030740
2. Pachauri N, Ahn CW. Electrical Energy Prediction of Combined Cycle Power Plant Using Gradient Boosted Generalized Additive Model. IEEE Access. 2022;10:24566-24577. doi:10.1109/access.2022.3153720
3. Na K, Lee J, Kim E. LF-Transformer: Latent Factorizer Transformer for Tabular Learning. IEEE Access. 2024;12:10690-10698. doi:10.1109/access.2024.3354972
4. Borisov V, Leemann T, Sessler K, Haug J, Pawelczyk M, Kasneci G. Deep Neural Networks and Tabular Data: A Survey. IEEE Transactions on Neural Networks and Learning Systems. 2021;35:7499-7519. doi:10.1109/tnnls.2022.3229161
5. Hollmann N, Müller SG, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. *Nature.* 2025;637:319 - 326. doi:10.1038/s41586-024-08328-6
6. Li X, Yue H, Li F, et al. An Incremental Regularization Kernel Randomized Neural Network for Electrical Energy Output Prediction in Combined Cycle Power Plant. IEEE Access. 2024;12:190434-190444. doi:10.1109/access.2024.3515481
7. Rahman AU, Alsenani Y, Zafar A, Ullah K, Rabie KM, Shongwe T. Enhancing heart disease prediction using a self-attention-based transformer model. Scientific Reports. 2024;14. doi:10.1038/s41598-024-51184-7
8. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems 35. 2022. doi:10.52202/068431-0037
9. Shwartz-Ziv R, Armon A. Tabular Data: Deep Learning is Not All You Need. Inf. Fusion. 2021;81:84-90. doi:10.1016/j.inffus.2021.11.011
10. Gorishniy YV, Rubachev I, Khrulkov V, Babenko A. Revisiting Deep Learning Models for Tabular Data. 2021.
11. Rubachev I, Alekberov A, Gorishniy YV, Babenko A. Revisiting Pretraining Objectives for Tabular Deep Learning. ArXiv. 2022;abs/2207.03208. doi:10.48550/arxiv.2207.03208
12. Dou B, Zhu Z, Merkurjev E, et al. Machine Learning Methods for Small Data Challenges in Molecular Science. Chemical reviews. 2023;123:8736 - 8780. doi:10.1021/acs.chemrev.3c00189