Predictive Modeling of Transformer Oil Temperature: A Comparative Analysis of Statistical, Machine Learning, and Deep Learning Approaches

immediate

Transformer oil temperature is a vital indicator of power distribution system health. This paper examines a reduced subset of the Electricity Transformer Temperature (ETT) dataset (ETT-small-h1), which consists of 726 daily records of oil temperature and six external load features (HUFL, HULL, MUFL, MULL, LUFL, LULL). We benchmark three categories of forecasting methods—traditional statistical models (Linear Regression, VAR), machine learning techniques (SVR, Decision Tree Regression, Random Forest Regression, and K-means-based clustering), and deep learning architectures (MLP, Gradient Boosting, KalmanForecaster, and LSTM)—under both simple and rolling forecasting strategies. Empirical results show that advanced ensemble and neural network models, notably Gradient Boosting and MLP, consistently outperform other methods in predictive accuracy. Analyses of correlation, Granger causality, and transfer entropy highlight HUFL, MULL, and MUFL as the most influential load features. These findings underscore the value of sophisticated nonlinear approaches for timely and accurate oil temperature predictions, enabling improved predictive maintenance and resource optimization in power systems.

Background

Power transformers are a key link in modern power distribution networks, ensuring reliable transmission of energy from power plants to end users. Transformer oil temperature is an important indicator of its operating health, directly reflecting the thermal stress and potential aging risks of the equipment. Excessive oil temperature may indicate overload operation or early degradation of insulation materials. If these problems are not solved in time, they may lead to costly equipment failures and large-scale power outages.

This study uses the Electricity Transformer Temperature (ETT) dataset([1]), which collects high-frequency and detailed records of oil temperature and various load characteristics of power transformers in parts of China. Through in-depth analysis of this dataset, the study aims to reveal the complex thermodynamic laws in transformer operation. Accurate prediction of oil temperature can not only enable grid operators to implement preventive maintenance strategies in advance and reduce the risk of unexpected downtime, but also optimize energy utilization efficiency, thereby ensuring the safe and stable operation of the entire power system.

This study is of great significance because it targets a core challenge in power system management - how to predict transformer health in a timely and accurate manner. By achieving accurate oil temperature prediction, it can not only improve the reliability of equipment operation, but also significantly reduce operating costs and energy losses, while extending equipment life, bringing significant environmental and economic benefits.

Data

The data used in this research come from a reduced version of the ETT dataset, referred to as ETT-small-h1. Building on the original one-hour sampling frequency, additional consideration was given to computational resource constraints and the large overall data volume, resulting in only the 0:00 record for each day being retained. The data cover the period from July 1, 2016, to June 26, 2018, yielding a total of 726 valid records. Each record is arranged chronologically, facilitating analysis of how the temperature of the transformer oil changes over time.

In this study, each data point contains eight dimensions: the recorded date, six external load values, and the target variable "oil temperature." Among these, the external load values include High Useful Load (HUFL), High UseLess Load (HULL), Middle Useful Load (MUFL), Middle UseLess Load (MULL), Low Useful Load (LUFL), and Low UseLess Load (LULL). These load values collectively reflect the operating conditions of the transformer under various load levels and provide essential input features for oil temperature prediction. Meanwhile, "oil temperature (OT)" serves as the only target variable in the dataset, recording the thermal state of the transformer and directly indicating its health status.

This data subset is particularly well-suited as the primary data source for the present research because it offers a relatively long yet continuous time series while substantially reducing the data volume and the computational burden by retaining only a single record each day. This balance enables a focused examination of the overall temporal evolution of oil temperature, as well as a multivariate prediction and analysis of oil temperature changes, thereby providing a reliable data foundation for subsequent experimental design and model validation.

Research Questions

In this study, our primary research questions are designed to guide our analysis and model comparisons. We begin by analyzing the data through exploratory methods, then compare the performance of different predictive approaches. Specifically, our research questions include:

- Is it possible that the selected data subset (ETT-small-h1) sampled at 0:00 each day (yielding 726 data points) captures the long-term trends in transformer oil temperature under sparse sampling conditions?
- What is the relative contribution of the six external load features (HUFL, HULL, MUFL, MULL, LUFL, LULL) to the prediction of oil temperature (OT)?

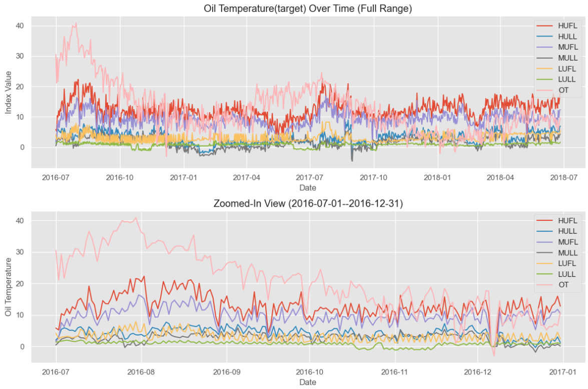

Fig. 1. Time-series visualization of the oil temperature (OT) and six external load features over the entire observational window (top), along with a zoomed-in segment (2016-07-01 to 2016-12-31) to highlight short-term fluctuations.

- How do traditional statistical methods (e.g., Linear Regression(2) and VAR(3)), machine learning models (e.g., SVR(4), Decision Tree Regression(5), Random Forest Regression(6), and K-means-based feature clustering(7)), and deep learning models (e.g., MLP Regression(8), Gradient Boosting Regression(9), KalmanForecaster(10), and LSTM(11)) compare in terms of prediction accuracy using both simple and rolling forecasting strategies?
- Based on the analysis and comparisons, can we identify a robust and effective predictive solution that balances computational efficiency and high accuracy for transformer oil temperature forecasting, and is the rolling approach actually superior?

Exploratory Data Analysis

Figure 1 shows the process of data visualization and figure 2 showcases correlation analysis and causation analysis. The exploratory analysis commenced with depicting overall trends in both the target variable (oil temperature, OT) and the six external load features (HUFL, HULL, MUFL, MULL, LUFL, LULL) over the entire observation window. In the upper panel of the time series figure, each variable is plotted across the complete span of available dates, revealing that HUFL and MULL in particular experience a wider range of fluctuations, while LUFL and LULL remain comparatively lower in magnitude. This global overview highlights the substantial volatility in oil temperature during certain intervals (e.g., mid-2016 through early 2017), which may indicate periods of heightened system dynamics or environmental impacts. A zoomed-in panel (covering 2016-07-01 to 2016-12-31) further illustrates how specific load indices change in tandem with OT, bringing out short-term correlations or phase shifts that might go unnoticed in the broader view.

Following the initial visual inspection, a correlation analysis was conducted, with results displayed in a heatmap. Positive correlations are shown in deeper shades of blue, while weaker correlations appear lighter. The matrix suggests that HUFL exhibits notably high correlation coefficients with OT, and MULL also shows moderate to strong associations. MUFL aligns closely with HUFL and exhibits a nontrivial correlation to OT as well. Meanwhile, some features (e.g., LUFL and LULL) appear only weakly correlated to OT in a purely linear sense. Such insights guide subsequent steps by pinpointing variables that may serve as stronger predictors in regression or forecasting models.

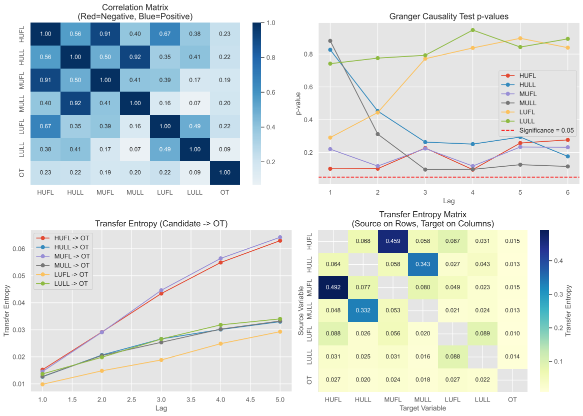

Fig. 2. Correlation heatmap, Granger causality p-value plot, and transfer entropy analysis among the OT and load feature variables, illustrating pairwise correlations, potential lead–lag relationships, and directional information flows.

To explore directionality and potential causation, two complementary methods were applied. First, Granger Causality Tests (plotted top-right) examine whether past values of each external load feature significantly enhance prediction of OT. Notably, HUFL and MULL show relatively low p-values in multiple lag configurations, suggesting that incorporating their historical values into a model for OT can statistically improve forecast accuracy. However, certain features, such as LUFL or LULL, maintain higher p-values across lags and thus may carry less explanatory power from a linear lead–lag perspective. Second, Transfer Entropy (TE), visualized both as a matrix (bottom-right) and as line plots for varying lags (bottom-left), quantifies potential nonlinear and directional information flows. In these plots, HUFL → OT and MULL → OT curves increase substantially with lag, indicating that these variables carry considerable predictive information over time, while LUFL and LULL remain comparatively low. This dual assessment helps confirm that HUFL, MULL, and MUFL are consistently associated with a higher informational or causal contribution to changes in OT, even once nonlinear dependencies are considered.

Answering the question of the relative contribution of the six external load features to the prediction of oil temperature, the combined evidence (time series visualization, correlation heatmap, Granger causality p-values, and Transfer Entropy scores) points to HUFL, MULL, and MUFL as having the greatest impact. HUFL in particular shows a strong correlation with OT and yields high Transfer Entropy, suggesting it transmits significant predictive signals to OT over multiple time lags. MULL likewise displays robust causal and entropic evidence of influencing OT. MUFL ranks next, showing moderate correlation and a measurable TE curve. Conversely, HULL, LUFL, and LULL appear to play more secondary roles: although they may still offer incremental predictive power under specific modeling setups, their overall contribution to OT forecasting is smaller relative to the more influential load indices. These findings not only clarify which inputs may be most valuable for short-term modeling of oil temperature but also highlight the potential for advanced nonlinear methods to capture otherwise overlooked relationships.

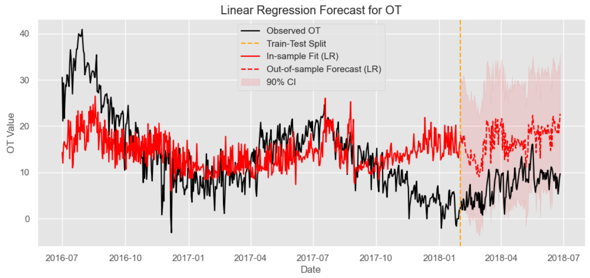

Fig. 3. Linear Regression forecast for transformer oil temperature, depicting both in-sample fitting (training data) and out-of-sample predictions (test data), along with a 90% confidence interval.

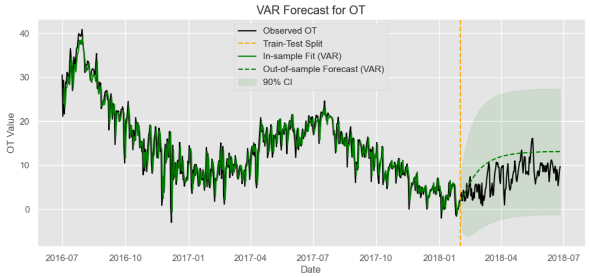

Fig. 4. Vector Autoregression (VAR) forecast for transformer oil temperature, showing how the model leverages mutual dependencies among multiple time series to produce in-sample and out-of-sample predictions.

Methodologies

This study compares three categories of predictive approaches—traditional statistical methods, machine learning models, and deep learning models—to forecast transformer oil temperature. The dataset was divided chronologically into a training portion (approximately 80% of the observations) and a test portion (approximately 20%) in order to measure out-of-sample performance. Two strategies were employed: a simple (one-shot) approach that uses the model trained on the initial period to forecast all future points, and a rolling approach that periodically re-trains or updates the model with new information.

Traditional Statistical Methods

Figure 3 depicts the forecasting curve of Linear Regression. This baseline approach considers oil temperature as a linear function of explanatory variables, with parameters estimated by minimizing squared errors on the training subset. Once fitted, the model was applied to the test data to generate forecasts. Although conceptually straightforward, linear assumptions often fail to capture nonlinear patterns inherent in transformer oil temperature sequences, leading to potential underfitting.

Figure 4 then shows the the prediction curve for Vector Autoregression (VAR). Here, all relevant series—including the oil temperature and any auxiliary factors—were modeled jointly, allowing each variable to depend on its own lagged values and those of the other variables. This system-based perspective can reveal underlying interdependencies, though VAR still relies on linear relationships and may require stationarity adjustments to handle trends or shifts in operating conditions.

Machine Learning Models

Figure 5 illustrates Support Vector Regression (SVR) using different kernel functions, such as linear, RBF, or polynomial kernels, to learn nonlinear mappings from historical data to future temperatures. The choice of hyperparameters, including the kernel's regularization and bandwidth settings, can markedly influence forecasting accuracy, and multiple kernel variations were tested for optimal performance.

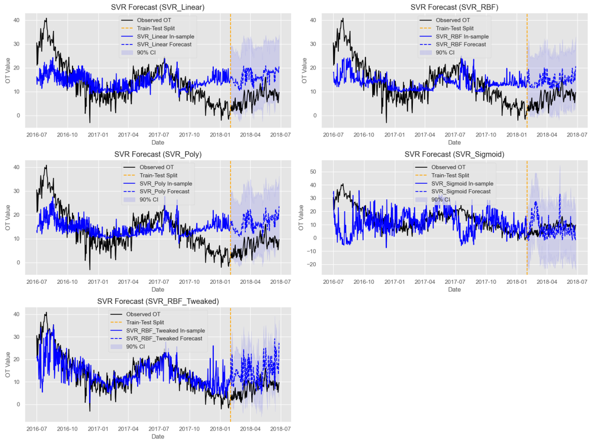

Fig. 5. Support Vector Regression (SVR) forecasts for transformer oil temperature using various kernel functions (e.g., linear, polynomial, RBF), demonstrating the impact of kernel choice on predictive performance.

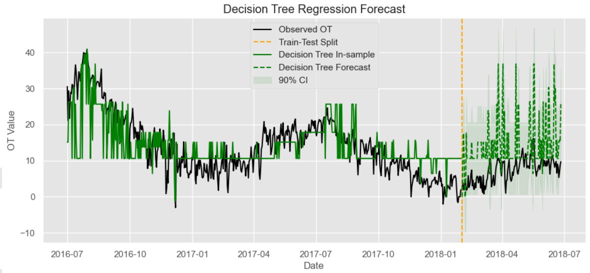

Fig. 6. Decision Tree Regression forecast for transformer oil temperature, illustrating piecewise constant predictions, the train–test split, and the model's sensitivity to branching depth.

Figure 6 presents the curve of Decision Tree Regression, where the feature space is recursively divided into regions that generate piecewise constant predictions. Such a tree can isolate important feature interactions, though it may also overfit unless pruning or depth controls are imposed.

Figure 7 highlights the Random Forest Regression approach, which mitigates high variance by averaging predictions from many decision trees trained on bootstrap samples of the original data. While each tree alone may be prone to noise, the ensemble effect typically yields more stable forecasts.

Figure 8 demonstrates the forecasting when a K-means-based feature clustering technique was applied to discover latent groupings or patterns in the predictor space. After clustering, the resulting cluster memberships served as additional inputs to a predictive model, aiming to enhance forecast accuracy by revealing subtler structures in the data.

Deep Learning Models

Figure 9 shows the prediction of a Multilayer Perceptron (MLP) network, which uses multiple hidden layers of neurons to transform input features into oil temperature predictions. Trained via backpropagation, the MLP can approximate nonlinear relationships more effectively than linear methods, though it may require careful

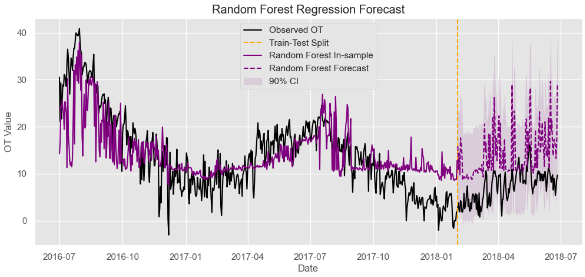

Fig. 7. Random Forest Regression forecast for transformer oil temperature, highlighting the ensemble approach's capacity to reduce variance through multiple decision trees aggregated by bagging.

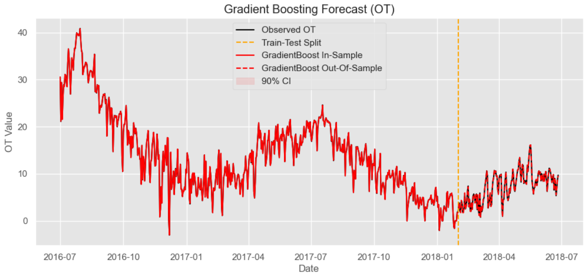

Fig. 10. Gradient Boosting Regression forecast for transformer oil temperature, showcasing iterative ensemble refinement where new weak learners are added to correct errors from previous ones.

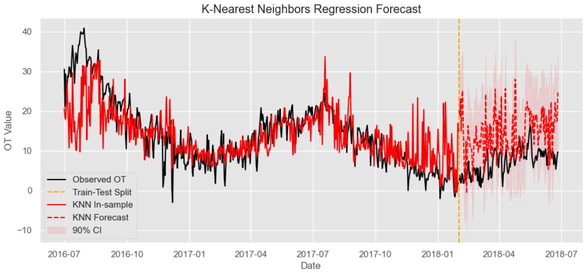

Fig. 8. K-Nearest Neighbors (KNN) Regression forecast for transformer oil temperature, visualizing how predictions emerge from the average of the most similar training observations in feature space.

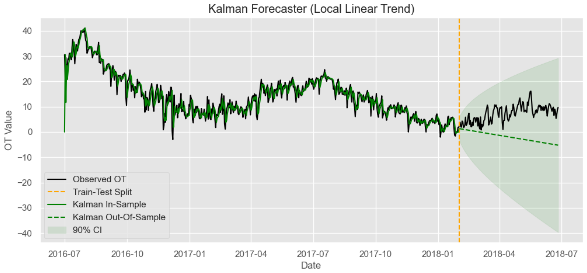

Fig. 11. Kalman Forecaster (Local Linear Trend) forecast for transformer oil temperature, highlighting the model's sequential state estimates and its application in tracking gradual changes over time.

tuning of layer sizes, learning rates, and regularization parameters to avoid overfitting.

Figure 10 introduces the forecasting Gradient Boosting, an iterative ensemble method that fits sequences of weak learners to residual errors, progressively refining the forecast. By selectively targeting misclassified or high-error instances at each step, Gradient Boosting often achieves high accuracy across many regression tasks, including oil temperature forecasting.

Figure 11 incorporates a KalmanForecaster, which interprets the system through a state-space lens, updating estimates of hidden states as new observations arrive. Although well-suited for tracking gradual, time-varying processes, this approach may be limited by its predominantly linear-Gaussian assumptions if the data exhibit more complex behaviors.

Figure 12 applies a Long Short-Term Memory (LSTM) network to address potential long-range dependencies in the time series. By maintaining memory cells that can store information over extended sequences, LSTMs can capture temporal patterns that simpler feed-forward networks might overlook, though they also demand significant computational resources and meticulous hyperparameter tuning.

In addition to implementing each of these forecasting methods, a small grid search was performed to optimize key hyperparameters using three-fold cross-validation on the training set, followed by evaluation on the same test set. For instance, SVR required tuning C, γ, ϵ, and kernel type; KNN benefited from adjusting the number of neighbors; Decision Trees and Random Forests needed parameters like depth constraints and the number of estimators. By systematically exploring each model's parameter space, the best-performing configurations were selected and then used to generate out-of-sample forecasts. This process demonstrated the impact of careful tuning in balancing predictive accuracy and overfitting risks.

Across all these figures and modeling techniques, training was consistently performed on the initial 80% of the data, with out-of-sample forecasts evaluated on the remaining 20%. Models were

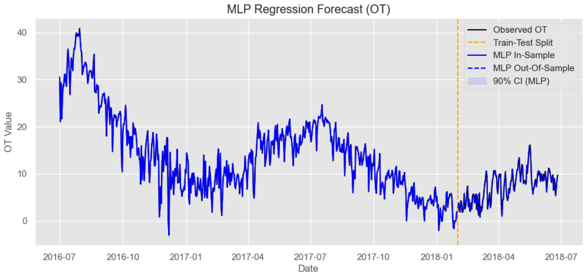

Fig. 9. Multilayer Perceptron (MLP) Regression forecast for transformer oil temperature, demonstrating the ability of a feed-forward neural network to capture nonlinear relationships in both in-sample and out-of-sample predictions.

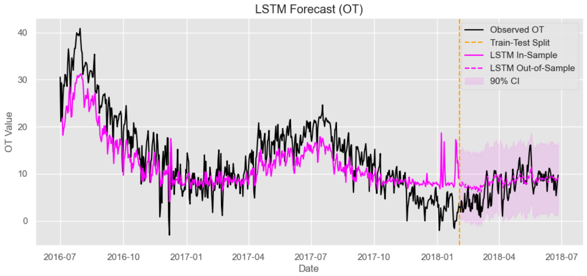

Fig. 12. LSTM neural network forecast for transformer oil temperature, emphasizing how memory cells capture long-range dependencies and temporal patterns in the data for in-sample and out-of-sample forecasts.

compared under both simple and rolling forecasting schemes to observe how continuously updating the model affects accuracy. This methodological setup enables a fair assessment of whether more advanced and potentially computationally intensive models can indeed provide higher-quality predictions than simpler linear or classical machine learning methods for the problem of transformer oil temperature forecasting.

Results

In the table1 comparing the predictive performance of various methods under both simple (one-shot) and rolling (iteratively updated) forecasting strategies, several patterns emerge that highlight how traditional statistical methods, classical machine learning models, and more advanced deep learning or ensemble approaches differ in terms of predictive accuracy. First, we note that linear methods such as Linear Regression and Vector Autoregression (VAR) exhibit relatively high errors under simple forecasting, particularly in their mean absolute percentage error (MAPE). This indicates that purely linear assumptions cannot capture the more complex, nonstationary dynamics of transformer oil temperature. However, VAR does show noticeable improvement when switching to rolling updates, suggesting that frequent retraining can partially mitigate model mismatch over time.

In contrast, the ensemble and neural network models—namely Random Forest, Gradient Boosting, and MLP—stand out as substantially more accurate overall, with much lower MAPE, MAE, and RMSE values in both simple and rolling modes. Specifically, Gradient Boosting and MLP maintain error metrics close to or even below 1–2% in MAPE, indicating robust performance and suggesting they are better able to capture nonlinear temperature dynamics. The rolling strategy confers additional benefits for most models, although MLP and Gradient Boosting already achieve strong results under simple forecasting, leaving less room for improvement. Meanwhile, the LSTM and Kalman approaches—both of which are designed for sequential data—yield mixed outcomes. Kalman filtering alone cannot accommodate the evidently complex dynamics, leading to relatively high errors, whereas the LSTM's simple-forecasting result shows moderate performance but remains behind MLP and Gradient Boosting in this particular setup. These differences in predictive accuracy also have implications for computational overhead: while frequent rolling updates can refine models in real time, the choice of algorithm should balance retraining costs against gains in forecast accuracy. From this figure, MLP and Gradient Boosting appear to be the most promising candidates for transformer oil temperature prediction, achieving strong accuracy while still retaining reasonable levels of computational efficiency.

Conclusions

In examining forecasting performance under both simple (one-shot) and rolling (incrementally updated) strategies, several clear trends emerge. Linear models such as Linear Regression and VAR generally yield larger errors in the simple approach—especially in terms of MAPE—indicating that they struggle to capture the inherent nonlinear and time-varying patterns of transformer oil temperature. However, when retrained regularly through rolling updates, VAR in particular demonstrates improved accuracy, highlighting that incremental learning can help these models adapt to changing system conditions.

By contrast, ensemble methods (Random Forest, Gradient Boosting) and the MLP neural network deliver consistently superior results, characterized by comparatively lower MAPE, MAE, and RMSE. Among them, MLP and Gradient Boosting achieve error rates often below 1–2%, suggesting a strong ability to capture complex temperature dynamics. The rolling method further enhances their already robust performance, though these models already exhibit strong predictive power even in simple mode. Meanwhile, both LSTM and KalmanForecaster—despite their suitability for sequential data—produce mixed outcomes. KalmanForecaster appears to lag in handling the pronounced nonlinearities, while LSTM's accuracy in simple mode remains notably behind MLP and Gradient Boosting, although it may improve with more extensive hyperparameter tuning or larger datasets.

These discrepancies in prediction accuracy also underscore the importance of weighing computational costs against accuracy gains. Although rolling updates can refine model performance over time, the overhead of frequent retraining may be nontrivial. Overall, MLP and Gradient Boosting emerge as the leading contenders for transformer oil temperature prediction, achieving strong accuracy while offering a favorable balance between complexity and effectiveness.

Limitations

Sparse Sampling: Retaining only one daily record (0:00) may overlook short-term fluctuations in oil temperature, potentially diminishing model granularity.
Data Source Specificity: The study relies exclusively on the ETT-small-h1 subset, and broader generalizability to other transformers or operating conditions remains unverified.
Feature Scope: Only six external load variables are utilized, excluding other potentially relevant factors such as ambient temperature and meteorological influences.
Hyperparameter Sensitivity: Advanced models like neural networks and Gradient Boosting require careful tuning and can be computationally intensive, especially for real-time or large-scale deployment.

[1] H Zhou, et al., Informer: Beyond efficient transformer for long sequence time-series forecasting in *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference.* (AAAI Press), Vol. 35, pp. 11106–11115 (2021).
[2] DC Montgomery, EA Peck, GG Vining, *Introduction to Linear Regression Analysis.* (John Wiley & Sons), (2012).
[3] CA Sims, Macroeconomics and reality: model evaluation, policy, and forecasting. *Econometrica* **48**, 1–48 (1980).
[4] AJ Smola, B Schölkopf, A tutorial on support vector regression. *Stat. computing* **14**, 199–222 (2004).
[5] L Breiman, JH Friedman, RA Olshen, CJ Stone, *Classification and Regression Trees.* (CRC Press), (1984).
[6] L Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
[7] J MacQueen, Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. on Math. Stat. Probab.* **1**, 281–297 (1967).
[8] DE Rumelhart, GE Hinton, RJ Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
[9] JH Friedman, Greedy function approximation: A gradient boosting machine. *Annals Stat.* **29**, 1189–1232 (2001).
[10] RE Kalman, A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**, 35–45 (1960).
[11] S Hochreiter, J Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

Table 1. Comparison of forecasting results across different models (Simple vs. Rolling)

Method	Strategy	max_error	MAE	MSE	RMSE	MAPE	MASE
LinearReg	Simple	17.935899	9.392129	98.840340	9.941848	200.043562	4.363714
LinearReg	Rolling	16.636338	7.826856	72.357478	8.506320	178.384910	3.636466
VAR	Simple	11.159230	3.907238	20.359628	4.512164	90.957057	1.815357
VAR	Rolling	7.433545	1.556315	3.937513	1.984317	35.800529	0.723085
SVR_RBF	Simple	15.693519	7.257191	63.639642	7.977446	170.444719	3.371792
SVR_RBF	Rolling	15.693519	5.552888	43.700182	6.610611	145.856738	2.579949
KNN	Simple	21.736999	7.532729	82.252800	9.069333	167.405756	3.499811
KNN	Rolling	18.712800	4.964751	43.780243	6.616664	121.417701	2.306692
DecisionTree	Simple	27.562600	7.629228	98.151569	9.907147	158.718049	3.544646
DecisionTree	Rolling	18.558750	4.381771	40.022677	6.326348	100.195387	2.035832
RandomForest	Simple	20.440523	6.827421	64.995221	8.061961	146.338382	3.172115
RandomForest	Rolling	17.681690	4.858292	35.476737	5.956235	114.317367	2.257230
MLP	Simple	0.383629	0.103343	0.018463	0.135880	1.929249	0.048015
MLP	Rolling	0.335375	0.079684	0.010870	0.104259	1.468502	0.037022
GradientBoost	Simple	0.233683	0.061597	0.006119	0.078222	1.296185	0.028619
GradientBoost	Rolling	0.208226	0.048250	0.004081	0.063880	1.066578	0.022417
Kalman	Simple	19.544506	9.273856	107.928023	10.388841	121.961652	4.308762
LSTM	Simple	6.864547	2.106423	7.596138	2.756109	61.702971	0.994722