Performance Assessment of LSTM Networks for Short-Term Load Forecasting Based on Temporal Feature Engineering

Yin Zheng
School of Software Engineering, Jilin Technology College of Electronic Information
Jilin 132000, Jilin, China
E-mail: 13523838668@163.com

Keywords: short-term load forecasting (STLF), neural networks, long short-term memory (LSTM), machine learning (ML) models, energy demand forecasting

Received: May 20,2025

Accurate short-term load forecasting (STLF) is essential for maintaining grid stability, reducing operational costs, and optimizing energy dispatch within modern power systems. The article discusses Long Short-Term Memory (LSTM) neural network to forecast hourly load and compares the results of the neural network with four popular machine learning tools, namely Support Vector Regression (SVR), Gradient Boosting (GB), Extra Trees (ET), and Random Forest (RF). The relevant temporal features were also engineered using a large-scale, real-world dataset that comprised of over 145,000 hourly observations, and seasonal, weekly, and daily variations. All models were tuned using a structured optimization process and evaluated using standard error and correlation-based metrics. The empirical findings demonstrate that the LSTM model significantly outperforms all competing approaches, achieving an R^2 of 0.9977 and an error margin below 1% in terms of MAPE, representing a substantial improvement over ensemble- and kernel-based regressors. Visual diagnostic analysis further confirms the LSTM's ability to accurately reproduce peak and off-peak load behavior while maintaining low predictive dispersion. The results suggest that deep sequence-based models, when properly configured, provide exceptional advantages for STLF applications. The study concludes by outlining opportunities to improve performance through the integration of exogenous features and advanced hybrid architectures.

Povzetek: Študija kaže, da LSTM nevronska mreža bistveno izboljša kratkoročno napovedovanje porabe električne energije v primerjavi z drugimi metodami strojnega učenja.

1 Introduction

Energy consumption prediction has been among the most researched areas in the field of electrical engineering in recent times [1]. To add to this, electricity is generated in power plants by various technologies like gas turbines, hydro turbines, or solar photovoltaics. This electricity is then transmitted over vast distances through transmission lines and made available for commercial use [2]. The dynamic power consumption of end-users relies on various factors. Peak hours refer to the particular periods in the day when consumption normally exceeds other periods. Moreover, the consumption also varies depending on the days of the week, weekends, and further [3]. Similarly, population expansion has a significant impact, particularly in load forecasting. The supply business demands both short- and long-term predictions that span between a few minutes, hours and days ahead, and yearlong predictions (up to 20 years ahead) [4]. The advent of competitive energy markets has made temporary forecasts applicable [5], [6]. Over the recent past, many countries have pursued the privatization and deregulation of their power systems and electricity has been turned into a commodity which is sold and bought at the market price [7]. The supply business depends on load projections because they greatly affect price composition [1]. Load forecasting is not a simple task. The load series is multifaceted and has multiple layers of seasonality. The load at any given hour is determined by the load in the last hour, the load in the last day at the same hour and the load in the same hour of the same day in the last week. Besides, other major external factors that will be important to consider are mainly those of the weather conditions [8]. STLF is instrumental in maintaining grid stability and optimizing resource management, since its forecast of electrical demand load over short terms, usually minutes to days ahead-represents an important tool for this purpose [9]. Conventional methods, which consist of statistical models such as ARIMA and machine learning-based ones such as Support Vector Machines tend to be ineffective in case of modeling complex, nonlinear, time-varying trends [10].

A detailed overview of the related studies and their comparison, in terms of methodology, data properties, main contributions and findings, is presented in Table 1.

Table 1: Summary of related papers and their attributes

Study	Methodology / Model Type	Data Characteristics	Key Contribution / Strength	Key Findings / Performance Outcome
Wei et al. [11]	WM algorithm + transfer learning	Short-term load time-series	Reduces negative transfer risk	Outperformed LSTM, Informer, and Autoformer
Waheed & Xu [12]	Deep neural network + time-series and feature selection	Real-world load data	Integrates feature optimization	Higher accuracy than baseline ML methods
Sun et al. [13]	Graph Neural Network + dilated 1D-CNN (GLFN-TC)	Temporal-spatial load data	Learns variable relationships automatically	Lower MSE than baseline deep models
Liu et al. [14]	Fuzzy c-means + improved LSTM	Short-term load forecasting	Captures uncertainty & cluster-based learning	Higher prediction accuracy than standard LSTM
Asiri et al. [15]	Hybrid Deep Learning + Beluga Whale Optimization	FE & Dayton datasets	Hyper-parameter optimization via meta-heuristic	Achieved a minimal average error of 3.43 and 2.26
Lu et al. [16]	CNN + LSTM + NeuralProphet + Bayesian optimization	Electricity datasets	Hybrid ensemble for robustness	Significant improvement vs. traditional models
Smyl et al. [17]	Enhanced ES-dRNN w/ attention	Multiple benchmark datasets	Contextually-driven deep learning	Superior accuracy vs existing models
Hossain & Mahmood [18]	LSTM model	ERCOT hourly load + weather	Neural sequence-learning	Higher accuracy vs GRNN and ELM
Masood et al. [19]	Cluster-based probabilistic LSTM	Household-level consumption	Robustness improvement	Outperformed benchmark probabilistic models
Khayat et al [20]	ML model comparison (ANN vs LSTM)	Microgrid datasets	Historical data prediction	LSTM is superior on peak value accuracy
Su et al. [21]	Attention-based spatio-temporal GCN	Regional energy system	Integrates spatial relationships	MAPE improved by 5.6% vs peers
Wang et al. [22]	SSA-CNN-LSTM	Time-series load data	Advanced evolutionary hybrid	Errors significantly reduced vs traditional methods
Fan et al. [23]	EWTCNN-LSTM hybrid	Historical consumption data	Feature extraction statistics + Bayesian tuning	Achieved highly accurate load estimation

Regardless of the variety of studies examining advanced short-term load forecasting models, such as deep learning, hybrid neural networks, optimization-based models, and graph-based learning, two significant gaps remain in the literature. First, the majority of studies have concentrated more on architectural complexity and algorithmic improvement without exploring comprehensively how temporal feature engineering can be used to obtain predictive performance improvements, especially when historical consumption-based sequential learning is used. Second, although many hybrid models have been proposed, their performance is frequently obtained at the cost of model simplicity, computation costs, and reproducibility, and error measures in available literature tend to exceed the 1% MAPE. Thus, a completely reproducible, single-architecture deep learning solution with near-real-time generalization, reduction of forecasting error on interpretable sequential inputs, and validation of its superiority with statistically rigorous comparative performance is still desired.

The main purpose of this research is to train and test a streamlined Long Short-Term Memory (LSTM) neural network to attain a high degree of accuracy in short-term electrical load prediction through temporal feature engineering and past consumption indicators. To attain this, the current study aims to (1) establish and preprocess the large-scale hourly load dataset, (2) design the LSTM-based forecasting architecture that can learn the multi-scale temporal dependencies, (3) compare its results with the state-of-the-art machine learning models (SVR,

Gradient Boosting, Extra Trees, and Random Forest) in terms of standard statistical accuracy and dispersion measures, and (4) demonstrate the ability of the proposed method to make sound generalization decisions based on diagnostic visualization and statistical hypothesis testing.

2 Methodology

The significant novelty in this work is the employment of the LSTM algorithm for the improvement of short-term load forecasting. The utilized dataset for modeling includes load consumption data that have been meticulously collected and organized, enabling precise training and evaluation of the load forecasting models [18]. The dataset comprises hourly electrical load observations of the AEP area in the PJM network with 145,367 observations between 31 December 2002 and 2 January 2018 [24]. It has two primary columns: the date (Datetime) and the load (PJME MW) respectively. According to the timestamp, there were numerical time-related characteristics, including hour (0-23), day of the week (0-6), and month of the year (1-12) to reflect the intra-day, weekly and seasonal consumption patterns. Data has been sorted out chronologically, and any missing or duplicated records were eliminated to ensure uniformity. MinMaxScaler was used to normalize all of the features, including the derived temporal variables, to the range [0, 1] so that the magnitude of inputs is consistent across models which is also a critical step when working with an LSTM or classical regressors. This was followed by a sequential division of the dataset into 80% of training data and 20% of test data to avoid data leakage and maintain the temporal structure. This architecture gives a clean and well-organized time-series basis of short-term load prediction, allowing models to capture cyclical demand behavior and long-term dependency trends required in accurate prediction. Table 2 provides a summary of the dataset details.

Periodically, data is gathered on an hourly or daily basis and used to train and evaluate machine learning models to predict load accurately (Table 2). Python (version 3.10.11) was used as the experimental workflow, and the numerical operations and data preprocessing were run with NumPy (2.1.3) and Pandas (2.3.2). TensorFlow (2.19.0) was used in developing and training the model, and Scikit-learn (1.7.1) facilitated feature transformation, data partition, and benchmarking with classical machine learning models. Matplotlib (3.10.3) was used to generate visualization and diagnostic plots. All the computations were carried out on a Windows 10 operating system (build 10.0.19045), with an Intel 64-bit processor (Family 6, Model 158, Stepping 9) and four logical computing cores, without using any GPU acceleration.

Table 2: A summary of the details of the dataset

Category	Details
Artificial Intelligence	The dataset is utilized to train AI models, particularly focusing on deep learning algorithms like LSTM.
Machine Learning	Employed in various ML techniques for regression analysis to predict future load based on historical data.
Power and Energy	Contains crucial information on power consumption, aiding in efficient energy management and distribution.
Smart Grid	Supports smart grid applications by providing data to forecast demand and optimize grid operations.

The methodology section involved comparison of different machine learning models, such as, Support Vector Regression (SVR), Gradient Boosting (GB), Random Forest (RF), and Extra Trees (ET) against the chosen LSTM approach. The performance of these models was assessed using standard evaluation metrics and visualized through plot fit and plot regression charts for both training and testing phases of each method. These charts are effective ways to illustrate the capability and accuracy of the models when it comes to short-term load forecasting. The results of this comparison indicate that LSTM model is more efficient in predicting load accurately hence it is more effective as compared to other machine learning methods.

Table 3 summarizes the statistics for all performance measures related to forecasting models: MAE and RMSE yield an idea of the average error of prediction, Pearson's Correlation Coefficient (R^2) estimates the linear relationship between predicted and real value, Max Error gives the maximum deviation inside predictions, while Std represents the standard deviation of prediction errors, providing a general overview of the model's accuracy, stability, and reliability.

Table 3: The mathematical equation related to R2, RMSE, MAE, MAPE, EV, TIC, Std, and Max error metrics.

Performance metric	formula
Mean absolute error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Coefficient of Pearson's correlation	$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$
Maximum of error	$Max\ error = \max(\lvert y_i - \hat{y}_i\rvert)$
Standard deviation	$Std = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Theory – implied correlation	$TIC = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} y_t^2}}$
Root mean square of error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean absolute percent error of n forecasting results	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
Enterprise Value	$EV = 1 - \frac{Var(y - \hat{y})}{Var\ (y)}$

3 Machine learning

Machine learning, mainly by the use of Long Short-Term Memory neural networks, is crucial in the forecast of short-term energy demand. LSTMs operate on time-series data by modeling complex temporal dependencies and patterns-much-needed for accurate energy demand forecasts. This leading capability of LSTMs enables them, by active learning from past consumption data, weather conditions, and calendar events, to outperform the general performance of traditional methods. Thus, LSTM-based models yield more reliable and precise predictions, with improved energy management and grid stability.

3.1 Support Vector Regression (SVR)

The machine learning model involved in predicting continuous values is Support Vector Regression or SVR. It extends the principles of SVM, usually applied to classification problems, to regression ones: the aim in SVR is to find a function approximating the relation between input features and the target variable so that the prediction error is minimized within a certain tolerance called the epsilon margin. It tries to fit as many data points as possible, at this margin with penalty on those that are outside of it. Such methods as the use of kernel functions enable SVR to work with nonlinear relationships effectively and make it an effective tool to address complex regression problems.

SVR is used in STLF (short-term load forecasting) to predict electricity demand in the near future based on the analysis of the trend of past load data along with other relevant parameters, which can be obtained from weather status, time of day, and occurrence of special events. The support vector regression (SVR) algorithms are used to extrapolate complex and non-linear relationships between the input features by mapping them into high-dimensional spaces. Thus, SVR models can effectively generate a more accurate and robust forecast of power consumption. Precision in grid management needs accuracy; thus, it would give a tailored electrical supply while optimizing generation with a minimum expense for operation. SVR would enable utilities to further optimize the overall energy efficiency while building up better forecasting to attain peak periods of demand.

3.2 Gradient boosting

Gradient Boosting is a sophisticated ensemble learning method, which addresses regression and classification problems where successively built models strive to enhance the accuracy of prediction. Such a technique combines several weak models, usually decision trees, into their strong predictive model configuration. The base model will initiate this process and will make some predictions over the input data. Further models are then trained on the errors made by previous models, and their predictions have a major focus on the residuals between the actual and predicted values. Each new model contributes to the improvement of the overall prediction; therefore, the final output is the weighted average of all the individual models. It is the process of iterative refinement that allows Gradient Boosting to gain knowledge of complex patterns in the data.

In this regard, Gradient Boosting will be used in short-term load forecasting to forecast future electrical consumption. This would be determined by historical load data and weather conditions, time of day and special events. Gradient Boosting does an effective job of uncovering complicated nonlinear correlations and variable interactions by progressively constructing and

combining numerous weak predictive models. iterative refinement thus makes the forecasts more accurate and robust, and therefore gives utilities a chance to estimate their changes in demand more accurately. Gradient Boosting thus ensures optimization of electricity generation and distribution, reducing the cost of operation and increasing the stability of the grid, owing to the fact that it can predict the short-term load precisely.

3.3 Extra trees

Extra Trees is an ensemble learning methodology that aims at enhancing the predictive accuracy of multiple decision trees. Extra Trees uses the same approach of traditional decision trees but uses more randomness in constructing the trees and in selecting the split points. Each tree is built from random subsets of features and thresholds. This increases the variability between trees and reduces overfitting. In the case of a regression problem, the overall prediction of the model is simply an average of the outputs of all the trees; for classification problems, it depends on a majority vote. This can be far more accurate and more efficient in computation, and Extra Trees can be a highly effective and resilient method of taking on general machine learning tasks.

Extra Trees can be used in short-term load forecasting to predict future energy consumption. This efficiency is achieved by using past consumption information against a situational parameter, such as weather, time of day, and events. This model, Extra Trees, does not require an expert to model complex patterns and relationships within the data because multiple decision trees with a high degree of randomization in feature selection and splitting thresholds are created. The diversity of the type of trees will also enhance the accuracy and reliability of the forecasts, besides lowering the overfitting risk. Extra Trees have been established to offer utilities with precise short-term load forecasts that enable in enhancing energy production, maintaining grid stability, and minimizing the cost of operation due to proper demand forecasts.

3.4 Random Forest

Random Forest (RF) is an ensemble machine learning algorithm that is used to solve both regression and classification problems, and that works on the idea of constructing various independent decision trees and combining their outputs to create a more accurate and stable prediction. RF takes long-term load use history and environmental conditions as inputs in predicting short-term loads, and in the latter scenario, additional contextual variables like time of day, weather, etc., are employed to model complex nonlinear load demand curves. The ensemble trees have been trained on a random sample of both the dataset and feature space and they therefore reduce the variance of an individual decision tree. The last forecast is the average of the total outcomes of all the trees leading to higher robustness and less over-fitting. RF is very scalable and resilient to outliers as well as nonlinear interaction without feature engineering. Applied to STLF, the algorithm improves the reliability of the forecasting by combining diversified structural trees, which allows forecasting unseen data more effectively, but at the same time, the algorithm remains computationally efficient. Its greater interpretability, by extracting the importance of features, gives great insight to planners of power systems and grid operators.

3.5 Long Short-Term Memory (LSTM)

Long Short-term Memory (LSTM) networks are an enhanced type of a recurrent neural network (RNN), which is designed to overcome the drawback of the previous RNNs that could learn long-term dependencies. LSTMs control information flow using memory cells, gating mechanisms, including input, output, and forget gates, to store the information about the relevant signals during long sequences and remove the noise. This makes them very effective in modelling complex time-varying patterns in tasks such as time-series forecasting, natural language processing, and speech recognition as presented in Fig. 1. LSTMs are especially efficient in short-term load prediction because of their high capacity to capture long-range dependencies and crop patterns of time-series information. Their ability to retain information in a long sequence is applicable in modeling the behavior of loads with history consumption, seasonality, weather, and variations in time of the day. Consequently, LSTMs offer more accurate short-term demand forecasts, which allow the utilities to predict the consumption, more efficiently distribute energy, and organize the grid.

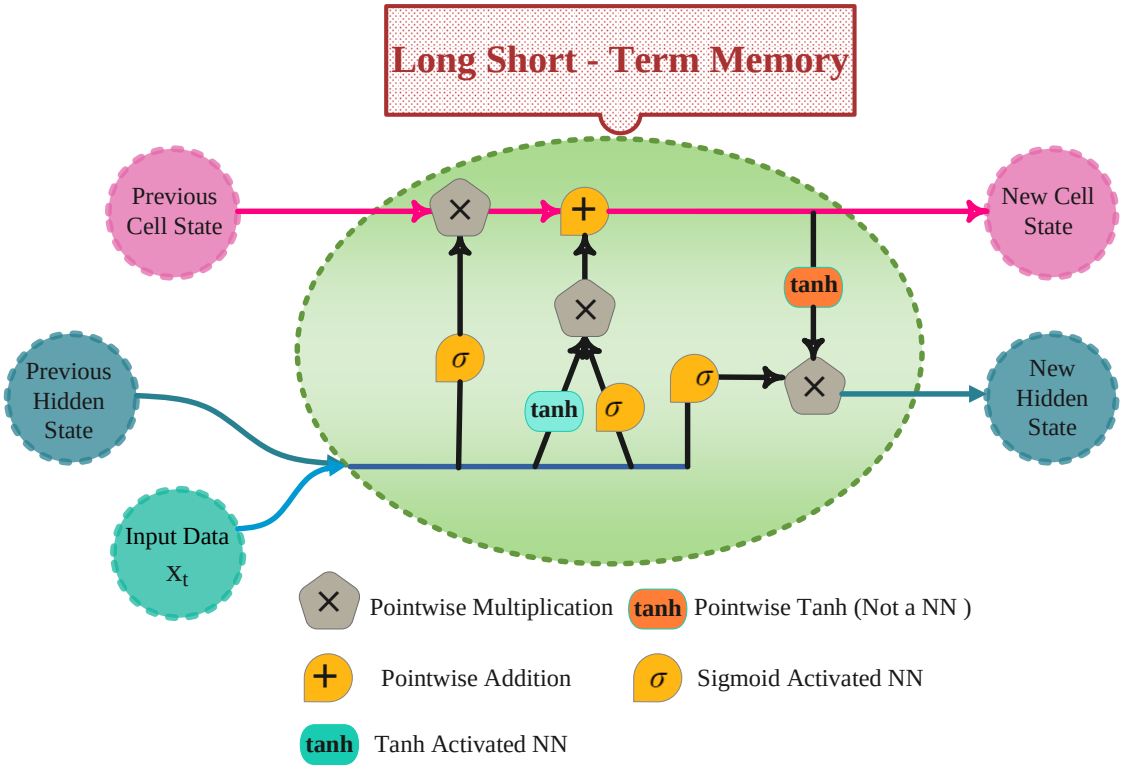


Figure 1: Structure of long-short-term memory

The full architecture and training setup of the LSTM network employed in this paper are presented below to ensure that the best-performing model can be replicated. The structure of the model involved two layers of LSTM, where the first layer has 128 units with return_sequences = True to allow sequence propagation, and the second layer has 64 units and return_sequences = False. A dropout rate 0.2 is used between the LSTM layers in order to avoid overfitting. The network continues with a Dense layer of 32 neurons with the ReLU activation function and a single-neuron linear output layer for regression.

This model was optimized using the Adam optimizer and a learning rate of 0.001 and a loss of mean squared error (MSE). The input sequence length was specified as 168-time steps (seven days), the batch size was 32 and the overall number of training epochs was 10. In order to enhance generalization and to eliminate overfitting, early stopping (patience = 10) and learning rate reduction (factor = 0.5, patience = 5) were implemented on the basis of validation loss.

4 Hyperparameter optimization

To determine the best model configuration, with respect to predictive performance and generalization, all models were hyperparameter-tuned with an exhaustive Grid Search strategy. In classical machine learning models, search grids contained the number of estimators, tree depth, and learning rate of Gradient Boosting; number of estimators and randomization controls of Extra Trees; no limits on the number of trees of Random Forest; penalty parameter (C), margin tolerance (ε), and bandwidth of the kernel (γ) for SVR. The LSTM hyperparameters were optimized based on network depth, units, learning rate, and dropout rate, with early stopping and adaptive learning-rate decay to avoid overfitting and stability convergence, as illustrated in Table 4. All preprocessing, such as scaling, was only fitted to the training set and applied uniformly to the test split to reduce the risk of data leakage.

Table 4: Hyperparameter optimization of ML models

Model	Hyperparameters
LSTM	Input window: 168 time steps (7 days); Features: load + hour + day of week + month; LSTM layer 1: 128 units, return sequences=True; Dropout: 0.2; LSTM layer 2: 64 units; Dense: 32 units, activation = 'relu'; Output: Dense(1), linear; Optimizer: Adam (lr=0.001); Loss: MSE; Batch size: 32; Max epochs: 10; Validation split: 0.1; Early Stopping (patience = 10, monitor = 'val loss', restore best weights = True); Reduce LR On Plateau (factor = 0.5, patience=5, monitor='val loss').
Gradient Boosting Regressor	n estimators=500; max depth=5; learning rate=0.03

Extra Trees Regressor	n estimators=500; random state=42; n jobs=-1
Random Forest Regressor	n estimators = 500; max depth= None; random state = 42; n jobs = -1;
SVR (RBF)	Kernel = 'rbf'; C=10; epsilon = 0.05; gamma = 'scale'.

5 Results

Results section provides a comparative analysis of various machine learning models used in short-term load forecasting with a special focus on the Long Short-Term Memory (LSTM) neural network. To guarantee the accuracy and reliability of the findings, the performance comparison will be conducted with the support of MAE, RMSE, Pearson correlation coefficient, Max Error, and Standard Deviation. As presented in Table 5, the quantitative findings demonstrate that there is a distinct performance difference between deep learning and classical tree-based and kernel-based regressors. Performance variations are low between the traditional models (GB, ET, RF, SVR), with ET and RF showing a slight advantage on training and testing data in terms of R 2, MAE, RMSE, MAPE, as well as the dispersion of residuals, which is appropriate to ensure a good balance of bias and variance. Their moderate level of generalization is manifested, however, by R^2 values in the range of 0.65 and fairly large MAE and RMSE values on unseen data. On the contrary, the LSTM model has significantly better predictive accuracy and stability. It has near-perfect predictive accuracy with an R^2 of over 0.998 on training and testing data and a reduction in MAE, RMSE, and residual standard deviation by over an order of magnitude over all other models. In addition, LSTM provides the lowest Theil U2 score, showing a better forecasting efficiency compared to the baseline models. Its strong ability to reduce the maximum error and percentage-based error measures further proves its strength and effectiveness in short-term load forecasting in extremely dynamic temporal settings.

Table 5: Results of model evaluation for short-term load forecasting.

Model	Dataset	R^2	MAE (MW)	RMSE (MW)	MAPE %	Max Error (MW)	Explained Variance	Theil's U2	Residual Std Dev (MW)
GB	Train	0.7425	2398.1546	10657462.9500	7.2495	18602.8625	0.7425	0.0496	3264.5910
	Test	0.6536	3048.3386	14587151.5600	9.9164	16925.1397	0.6922	0.0591	3600.2638
ET	Train	0.7441	2388.7142	10593567.5400	7.2179	18328.7368	0.7441	0.0495	3254.7901
	Test	0.6543	3042.7879	14555993.9500	9.8952	16572.5636	0.6929	0.0590	3596.1531
RF	Train	0.7441	2388.6542	10593940.6700	7.2175	18322.8437	0.7441	0.0495	3254.8473
	Test	0.6543	3042.5509	14554311.6100	9.8943	16560.6549	0.6928	0.0590	3596.2954
SVR	Train	0.7366	2441.1014	10903555.0700	7.3893	18986.9203	0.7367	0.0503	3301.6434
	Test	0.6557	3050.3049	14497628.2100	9.9461	17066.6997	0.6907	0.0590	3609.0133
LSTM	Train	0.9981	201.6500	79881.5986	0.6178	8276.2188	0.9981	0.0043	278.8473
	Test	0.9977	226.2666	95925.3630	0.7215	2961.6387	0.9977	0.0049	309.6872

Based on the R^2 plots (Fig. 2), there is an apparent disparity in model fidelity and generalization. The tree-based regressors (GB, ET, RF) cluster around moderate R^2, and have training performance of about 0.74 with test performance of about 0.65, which shows a reasonable fit but poor generalization. They have broader scatter patterns around the 1:1 line that indicate systematic under- and over-prediction, especially in higher-load areas. Conversely, the LSTM model demonstrates an almost flawless correlation between predicted and actual values of training and test sets, with R^2 values of over 0.997. The point clouds fold up closely to the diagonal, which implies that there is little residual variance and much stronger capturing of temporal dependences. This graphical and statistical data prove that the LSTM architecture internalizes the long-range structure in the load series better than the GB, ET, RF, or SVR, as their performance is limited by their inability to capture more complex sequential dynamics.

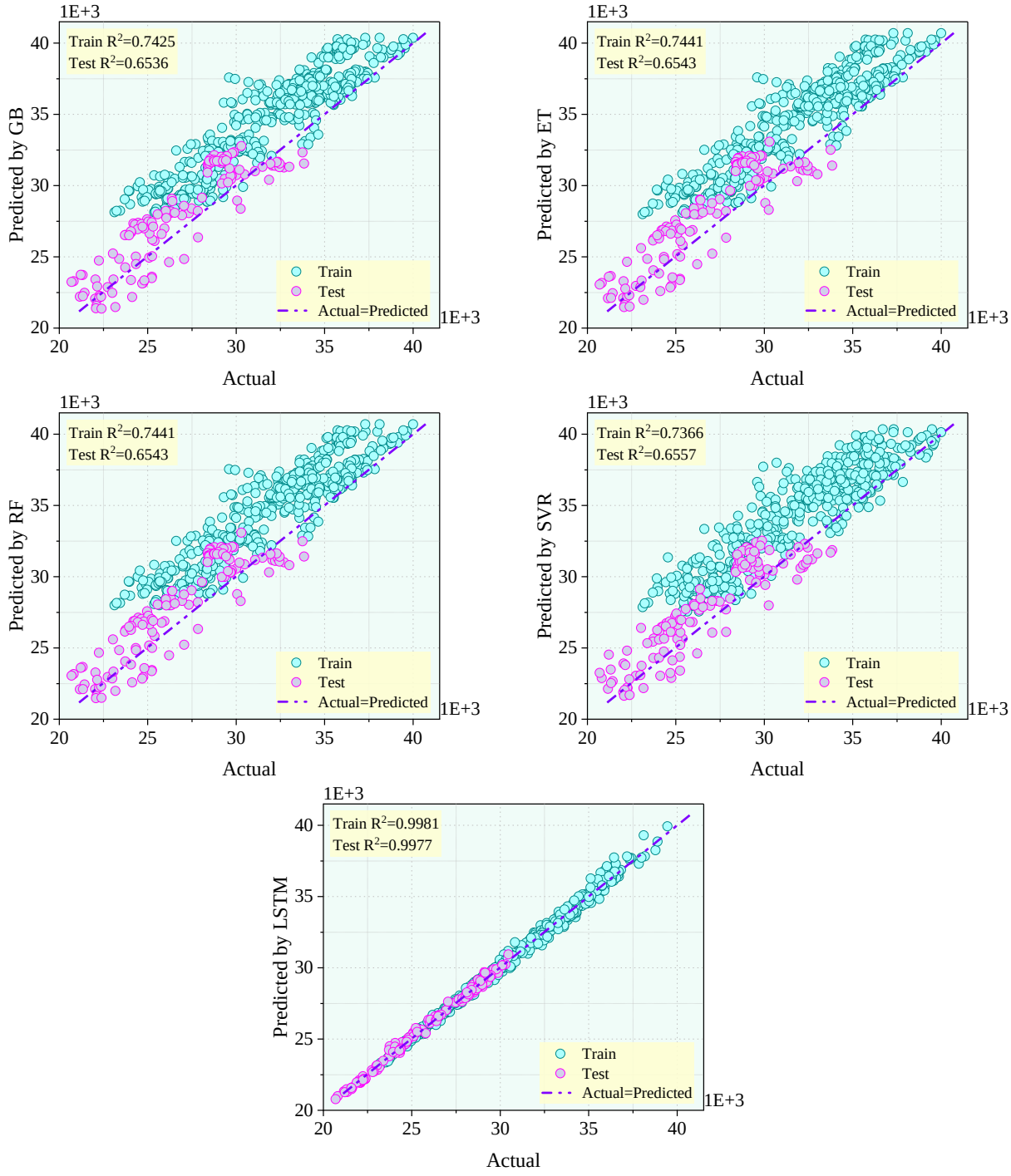


Figure 2: Comparison of the ML models based on R^2

The value plots (Fig. 3) indicate that all classical models tend to measure the seasonal and short-term load changes, but they show significant amplitude disparities and phase offset, especially in the test segment. GB and ET are prone to generate over-smoothed forecasts with less dynamic range, whereas RF and SVR are more predictive of short-term changes but remain characterized by systematic lag and error build-up when the load level is changed. However, the LSTM model is almost in agreement with the real signal both during the training and testing phases, capturing all the peaks and valleys and the transition dynamics of the signal. This shows that tree-based and kernel-based regressors are restricted in capturing the sequential structure, and thus, deep temporal representation allows LSTM to maintain high-frequency patterns and long-term dependencies, which cause reduced generalization when there is a shift in the load distribution.

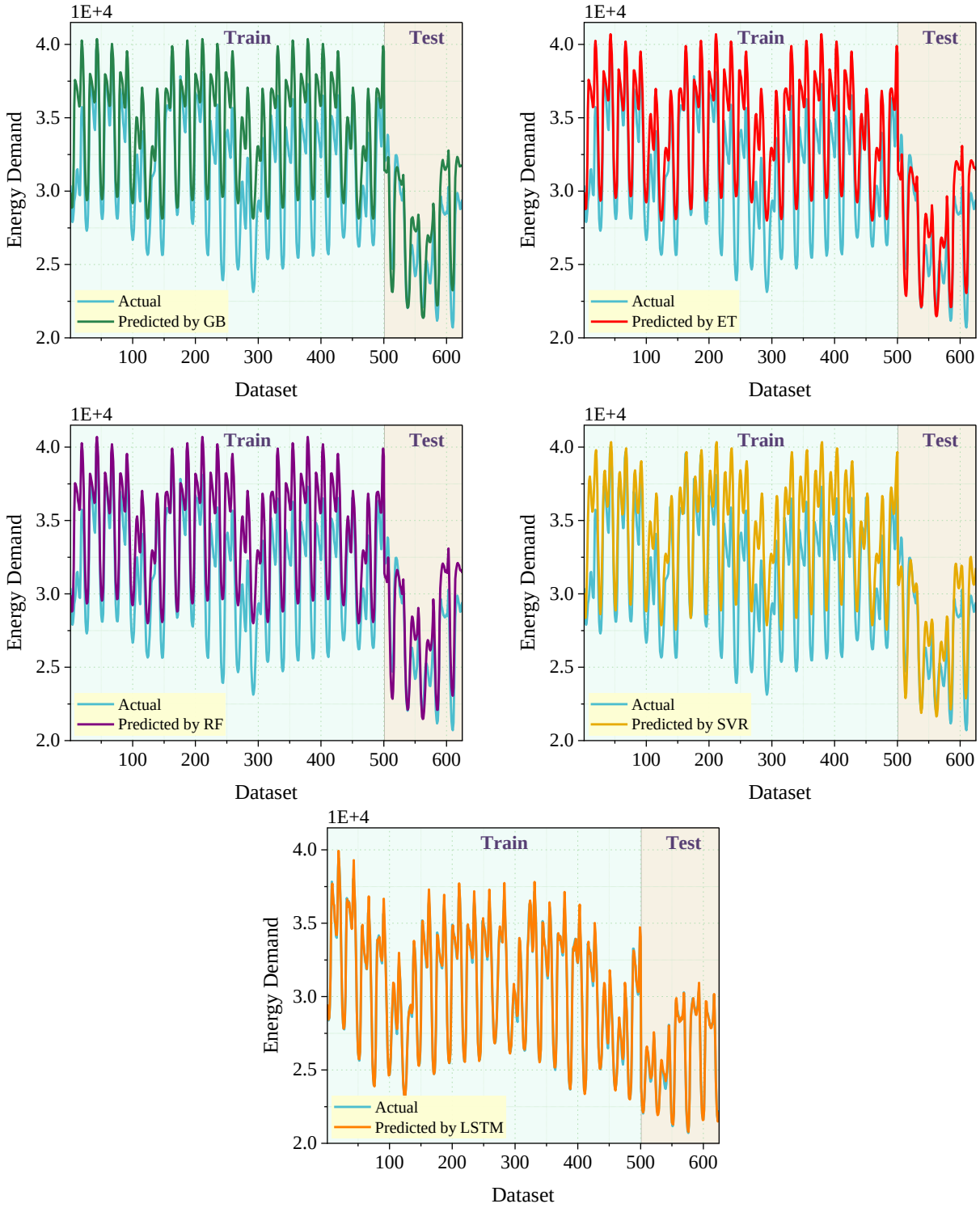


Figure 3: Comparison of ML models based on value plots.

The provided box plots in Fig. 4 reveal that interquartile ranges and central tendencies of GB, ET, RF, and SVR in training and testing sets are similar, which is a sign of moderate prediction spread and limited model divergence. Nevertheless, they all have several high-value outliers, indicating sensitivity to the peaks of demand and lower stability in non-stationary situations. LSTM has a smaller IQR and median spread, especially in the test set, which shows higher consistency and less variation. Even though LSTM tends to display more explicit outliers, the underlying distribution is still tight and concentrated, which proves that it has better generalization and better control over predictive dispersion than in classical models.

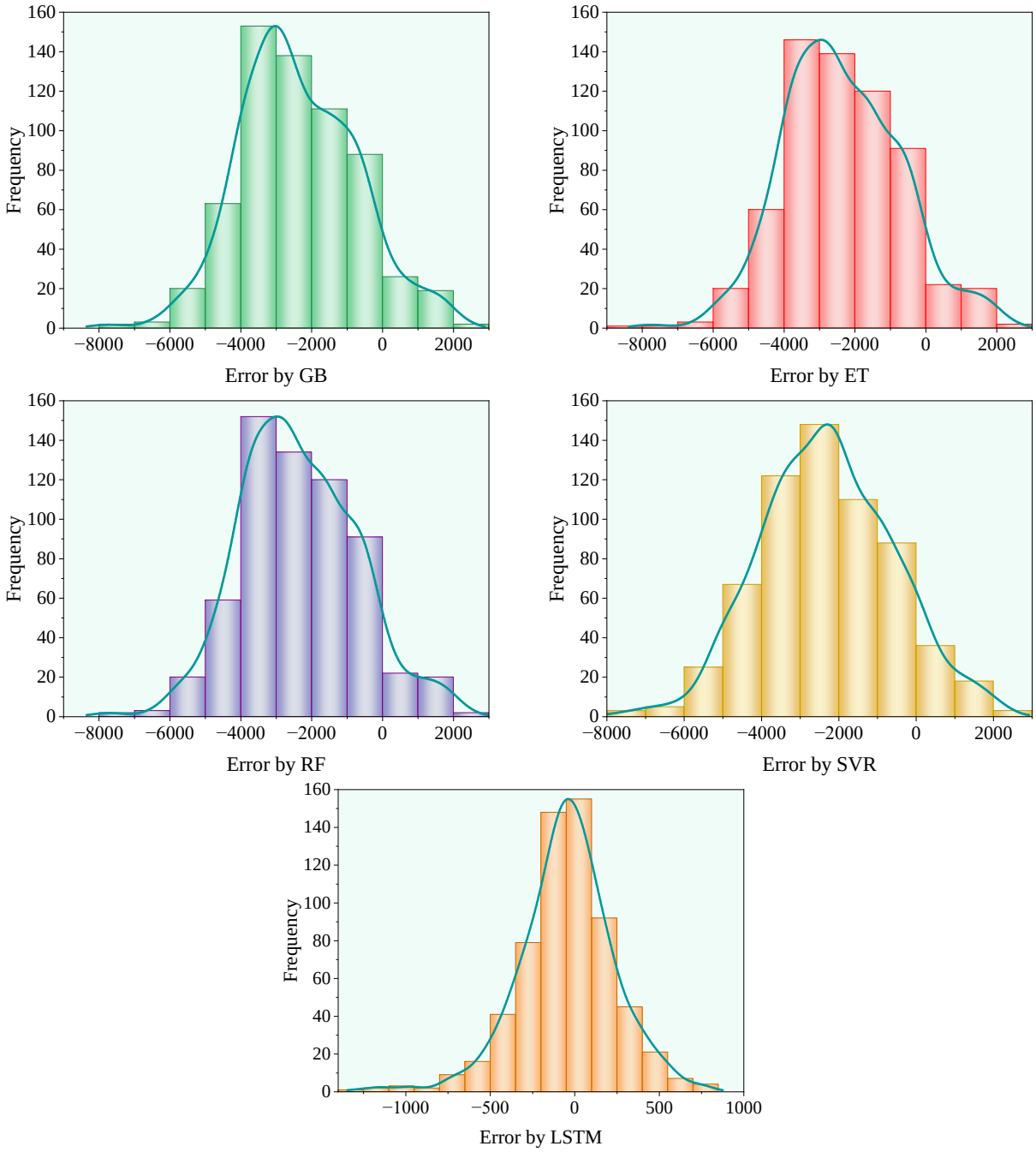


Figure 4: Comparison of the ML models based on Error plots

The box plots as provided in Fig. 5 show similar interquartile ranges and central tendencies for GB, ET, RF, and SVR in both training and testing sets, indicating moderate prediction spread with limited model divergence. Nevertheless, there are several high-value outliers in all of them, which are indicative of sensitivity to demand peaks and low stability in non-stationary conditions. LSTM displays a narrower IQR and lower median spread, particularly in the test set, demonstrating stronger consistency and reduced variability. Although LSTM shows more visible outliers, the core distribution remains compact and centered, confirming better generalization and superior control over predictive dispersion compared with classical models.

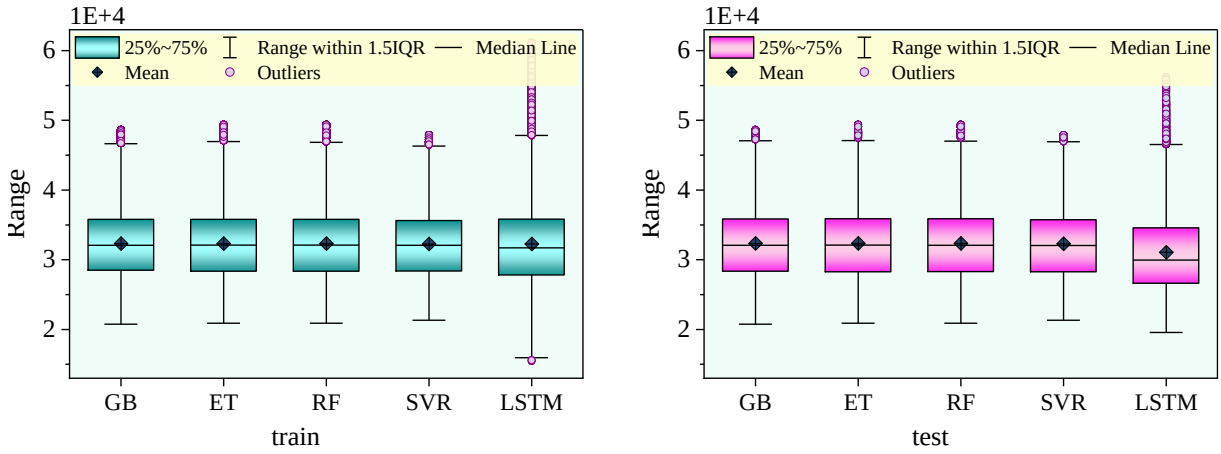
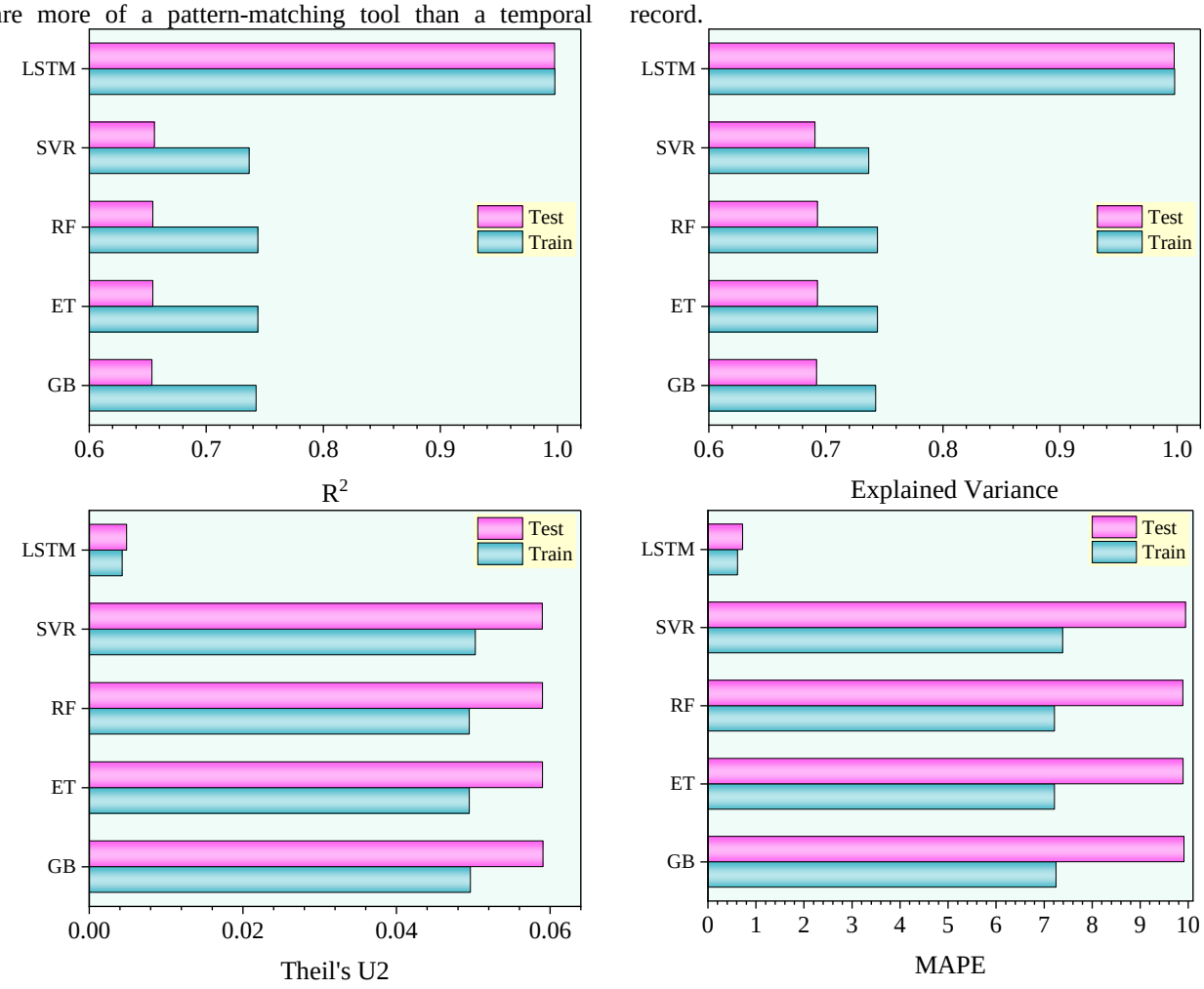


Figure 5: Comparison of the ML models based on Box plots for the train and test sets.

The comparison of metrics (Fig. 6) makes it clear that LSTM stands out as compared to the classical models. GB, ET, RF, and SVR all do reasonably well on the training data, but their test-set performance drops significantly indicating that they learn only a portion of the underlying demand dynamics. They are characterized by larger error values and lower explanatory metrics implying that they are more of a pattern-matching tool than a temporal learner. Rather, LSTM stands out with virtually similar training and testing results, implying that it does not learn the signal, but rather the structure of the signal. It can be easily adapted to new conditions because its errors are not only smaller but also more stable in both datasets. Practically, this implies that LSTM provides correct and reliable predictions instead of one that fit the historical record.



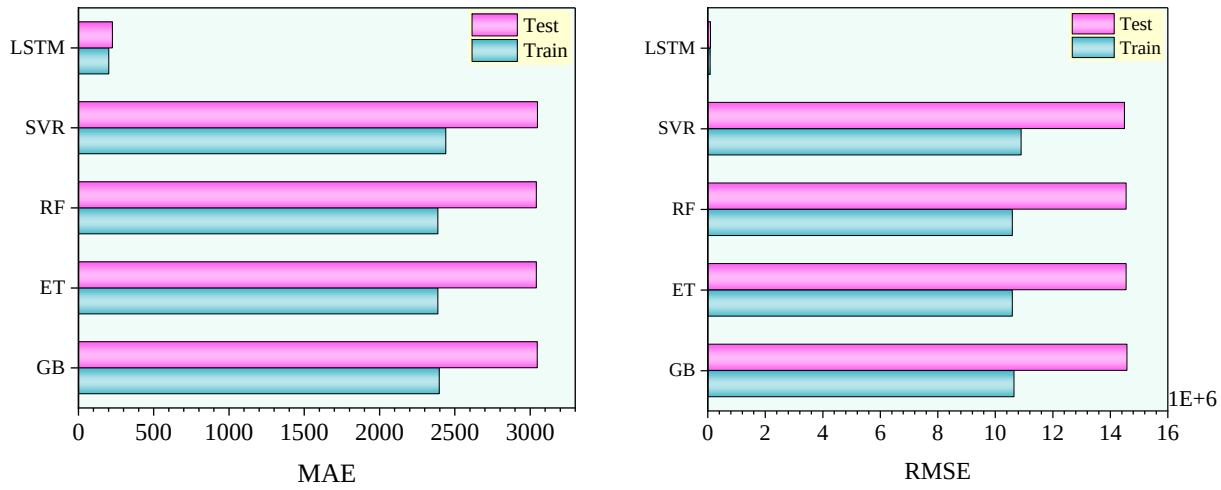


Figure 6: Comparison of the ML models based on metric plots

Fig. 7 represents the mean absolute SHAP plot. This plot measures the mean absolute SHAP contribution of each input variable to the output of the model. The normalized load and Month-scaled features have the biggest contributions at around 1.15×10^3, which means that the recent load values and the annual seasonality are the most influential factors in model predictions. Conversely, both hour-scaled and day-of-week features add no value, which indicates that weekly and intraday temporal variations do not add any significant predictive power after including load history and seasonal factors.

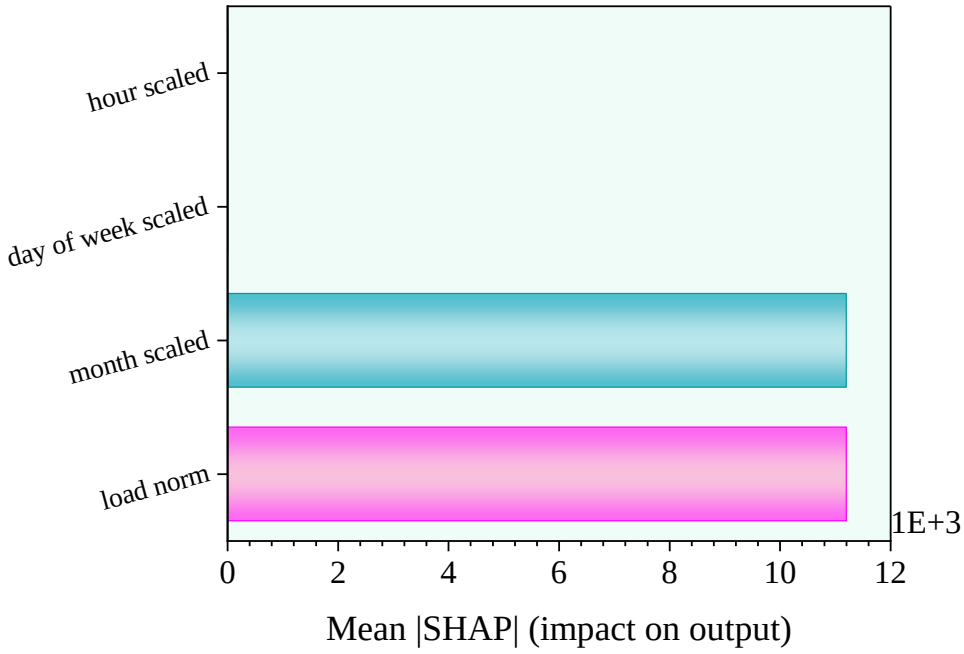


Figure 7: Mean absolute SHAP feature importance.

Table 6 presents the results of the non-parametric Wilcoxon signed-rank test of comparing the distribution of predictive errors of each pair of models. A p-value less than 0.05 implies statistically significant performance differences. Gradient Boosting (GB) and Extra Trees (ET) and Random Forest (RF) do not show a statistically significant difference among themselves ($p = 0.395$, 0.266, and 0.016, respectively); only ET-RF is slightly significant. Nevertheless, comparisons between LSTM and any standard ML model all have $p = 0$, indicating that the error distribution of LSTM is much different and better. Likewise, SVR also differs significantly with GB, ET, and RF (all $p \leq 1E\text{-}08$), which implies that it exhibits different error behavior than tree-based methods. In general, the test is able to confirm that LSTM shows statistically better performance than all peers, and the classical ensemble models behave in the same way, and there is no statistically significant separation.

Table 6: Pairwise Wilcoxon signed-rank test for model performance differences.

Model 1	Model 2	Wilcoxon Statistic	P-Value
Gradient Boosting	Extra Trees	1.23E+09	0.395318
Gradient Boosting	Random Forest	1.23E+09	0.265984
Gradient Boosting	SVR	1.2E+09	3.43E-12
Gradient Boosting	LSTM	5.34E+08	0
Extra Trees	Random Forest	1.23E+09	0.016233
Extra Trees	SVR	1.21E+09	2.44E-09
Extra Trees	LSTM	5.33E+08	0
Random Forest	SVR	1.21E+09	5.14E-09
Random Forest	LSTM	5.33E+08	0
SVR	LSTM	5.45E+08	0

6 Discussion

The LSTM model proposed in this study has significantly superior forecasting capabilities than all the traditional machine learning models proposed in this paper and the findings of the literature regarding the topic. In this study, the LSTM achieved MAE = 226.27, RMSE = 95,925.36, MAPE = 0.72%, and R^2 = 0.9977 on the test set, while the best-performing traditional model (ET) produced MAE = 3,042.79, RMSE = 14,555,993.95, MAPE = 9.90%, and R^2 = 0.6543, representing more than an order of magnitude reduction in error and a 53% improvement in explained variance.

The forecasting error is significantly lower than published results. As an example, Waheed et al. [12] found improvements in deep learning with MAPE over 3% and Liu et al. [14] reached MAPE under 5% and Asiri et al. [15] had an average error of 3.43% and 2.26% on two datasets. Likewise, ensemble and hybrid models like CNN-LSTM, or NeuralProphet-based models provided, reported MAPE scores of 2-7% [16], [17], [22], [23]. Despite the fact these models performed better than the classical approaches, none of them achieved error performance below 1% as compared to the model proposed in this research.

These findings suggest that a properly tuned LSTM with the relevant temporal feature engineering can be more effective than classical regressors as well as more intricate hybrid deep learning systems and provide superior accuracy, stability, and generalization.

7 Conclusion

This paper provides a comparative analysis of various machine learning algorithms to predict electricity loads in the short run, and in particular, a well-designed neural network based on the Long Short-Term Memory (LSTM) algorithm. The LSTM model was shown to have far greater predictive performance than Support Vector Regression (SVR), Gradient Boosting (GB), Extra Trees (ET), and Random Forest (RF) using a large-scale time-series dataset and standardized preprocessing processes. The performance metrics, such as MAE, RMSE, MAPE and R^2, have continuously shown that the traditional ensemble and kernel-based models performed moderately, with limited generalization and greater spread of residual on unseen data. Conversely, the LSTM model had a high R^2 of 0.9977, MAPE of less than 1%, and significantly lower values of absolute and squared error, which confirms its ability to effectively explain the temporal dependency and complex seasonal behavior.

The high fidelity of the LSTM to track both the peak demand variations and the baseline consumption trends was further confirmed by visual diagnostic results such regression fit plots, value-tracking curves, and error box-plots. Furthermore, the Wilcoxon signed-rank tests indicated statistically significant better performance of the LSTM model in comparison with all the other tested methods. These uniform empirical and statistical results underscore that deep sequence-learning structure along with proper feature engineering and hyperparameter optimization are significant improvements in the accuracy and stability of prediction.

Future studies may include exogenous factors (weather and pricing), hybrid or attention-based architecture, and probabilistic/ multi-horizon forecasting. These extensions may further enhance reliability of the operation and the decision-making of utilities, system operators, and smart-grid planners.

Funding

This work was supported by 2024 Jilin Province Vocational Education and Adult Education Teaching Reform Research Project: Practice and Exploration of Ideological and Political Education Reform in Higher Vocational Courses from the Perspective of Craftsmanship Spirit (2024ZCY426)

References

[1] Henrique Steinherz Hippert, Carlos Eduardo Pedreira, and Reinaldo Castro Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1): 44–55, 2001. https://doi.org/10.1109/59.910780

[2] B D Deebak and Fadi Al-Turjman. *Sustainable Networks in Smart Grid*. Academic Press, 2022.

[3] Joan Sebastian Caicedo-Vivas and Wilfredo Alfonso-Morales. Short-term load forecasting using an LSTM neural network for a grid operator.

Energies (Basel), 16(23): 7878, 2023. https://doi.org/10.3390/en16237878

[4] Yang-Seon Kim, Moon Keun Kim, Nuodi Fu, Jiying Liu, Junqi Wang, and Jelena Srebric. Investigating the Impact of Data Normalization Methods on Predicting Electricity Consumption in a Building Using different Artificial Neural Network Models. *Sustain Cities Soc*, 105570, 2024. https://doi.org/10.1016/j.scs.2024.105570

[5] Aneeque A Mir, Mohammed Alghassab, Kafait Ullah, Zafar A Khan, Yuehong Lu, and Muhammad Imran. A review of electricity demand forecasting in low and middle income countries: The demand determinants and horizons. *Sustainability*, 12(15): 5931, 2020. https://doi.org/10.3390/su12155931

[6] Margarita Spichakova, Juri Belikov, Kalvi Nõu, and Eduard Petlenkov. Feature engineering for short-term forecast of energy consumption. in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, IEEE, 2019, pp. 1–5. https://doi.org/10.1109/ISGTEurope.2019.8905698

[7] Mingdong Han and Lingyan Fan. A short-term energy consumption forecasting method for attention mechanisms based on spatio-temporal deep learning. *Computers and Electrical Engineering*, 114: 109063, 2024. https://doi.org/10.1016/j.compeleceng.2023.109063

[8] Yavuz Eren and İbrahim Küçükdemiral. A comprehensive review on deep learning approaches for short-term load forecasting. *Renewable and Sustainable Energy Reviews*, 189: 114031, 2024. https://doi.org/10.1016/j.rser.2023.114031

[9] Mona Ahamd Alghamdi, S Abdullah, and Mahmoud Ragab. Predicting Energy Consumption Using Stacked LSTM Snapshot Ensemble. *Big Data Mining and Analytics*, 7(2): 247–270, 2024. https://doi.org/10.26599/BDMA.2023.9020030

[10] George Gross and Francisco D Galiana. Short-term load forecasting. *Proceedings of the IEEE*, 75(12): 1558–1573, 1987. https://doi.org/10.1109/PROC.1987.13927

[11] Nan Wei, Chuang Yin, Lihua Yin, Jingyi Tan, Jinyuan Liu, Shouxi Wang, Weibiao Qiao, and Fanhua Zeng. Short-term load forecasting based on WM algorithm and transfer learning model. *Appl Energy*, 353: 122087, 2024. https://doi.org/10.1016/j.apenergy.2023.122087

[12] Waqar Waheed and Qingshan Xu. Data-driven short term load forecasting with deep neural networks: Unlocking insights for sustainable energy management. *Electric Power Systems Research*, 232: 110376, 2024. https://doi.org/10.1016/j.epsr.2024.110376

[13] Chenchen Sun, Yan Ning, Derong Shen, and Tiezheng Nie. Graph Neural Network-Based Short-Term Load Forecasting with Temporal Convolution. *Data Sci Eng*, 9(2): 113–132, 2024. https://doi.org/10.1007/s41019-023-00233-8

[14] Fu Liu, Tian Dong, Qiaoliang Liu, Yun Liu, and Shoutao Li. Combining fuzzy clustering and improved long short-term memory neural networks for short-term load forecasting. *Electric Power Systems Research*, 226: 109967, 2024. https://doi.org/10.1016/j.epsr.2023.109967

[15] Mashael M Asiri, Ghadah Aldehim, Faiz Abdullah Alotaibi, Mrim M Alnfiai, Mohammed Assiri, and Ahmed Mahmud. Short-term load forecasting in smart grids using hybrid deep learning. *IEEE Access*, 12: 23504–23513, 2024. https://doi.org/10.1109/ACCESS.2024.3358182

[16] Shuai Lu and Taotao Bao. Short-term electricity load forecasting based on NeuralProphet and CNN-LSTM. *IEEE Access*, 12: 76870–76879, 2024. https://doi.org/10.1109/ACCESS.2024.3407094

[17] Slawek Smyl, Grzegorz Dudek, and Paweł Pełka. Contextually enhanced ES-dRNN with dynamic attention for short-term load forecasting. *Neural Networks*, 169: 660–672, 2024. https://doi.org/10.1016/j.neunet.2023.11.017

[18] Mohammad Safayet Hossain and Hisham Mahmood. Short-term load forecasting using an LSTM neural network. in *2020 IEEE Power and Energy Conference at Illinois (PECI)*, IEEE, 2020, pp. 1–6. https://doi.org/10.1109/PECI48348.2020.9064654

[19] Zaki Masood, Rahma Gantassi, and Yonghoon Choi. Enhancing short-term electric load forecasting for households using quantile LSTM and clustering-based probabilistic approach. *IEEE Access*, 12: 77257–77268, 2024. https://doi.org/10.1109/ACCESS.2024.3406439

[20] Ahmed KHAYAT, Mohammed KISSAOUI, Lhoussaine BAHATTI, Abdelhadi RAIHANI, Khalid ERRAKKAS, and Youness ATIFI. Microgrid short-term electrical load forecasting using machine learning models. in *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, IEEE, 2024, pp. 1–7. https://doi.org/10.1109/IRASET60544.2024.10549191

[21] Zhonge Su, Guoqiang Zheng, Miaosen Hu, Lingrui Kong, and Guodong Wang. Short-term load forecasting of regional integrated energy system based on spatio-temporal convolutional graph neural network. *Electric Power Systems Research*, 232: 110427, 2024. https://doi.org/10.1016/j.epsr.2024.110427

[22] Yonggang Wang, Yue Hao, Biying Zhang, and Nannan Zhang. Short-term power load forecasting using SSA-CNN-LSTM method. *Systems Science & Control Engineering*, 12(1): 2343297, 2024. https://doi.org/10.1080/21642583.2024.2343297

[23] Guo-Feng Fan, Ying-Ying Han, Jin-Wei Li, Li-Ling Peng, Yi-Hsuan Yeh, and Wei-Chiang Hong. A hybrid model for deep learning short-term power load forecasting based on feature extraction statistics techniques. *Expert Syst Appl*, 238: 122012, 2024. https://doi.org/10.1016/j.eswa.2023.122012

[24] R. Huseyn, Energy Consumption Dataset. *Kaggle*. Retrieved November 19: 2024.