European Journal of Computer Science and Information Technology, 14(5),30-45, 2026
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

SDRL for Hourly Electricity Load Forecasting: A Prophet–Gradient-Boosting Hybrid with Statistically Validated Accuracy

Md Salman, [1] Md Shahin Ali,[2] Shaykat Purokayisto, [3]*
[1,2,3] *School of Artificial Intelligence, Nanjing University of Information Science and Technology*
***Correspondence Email**: salman25121998@gmail.com | shaykatariyan@gmail.com

doi: https://doi.org/10.37745/ejcsit.2013/vol14n53045 Published August 03, 2026

Citation: Salman M., Ali M.S., Purokayisto S. (2026) SDRL for Hourly Electricity Load Forecasting: A Prophet–Gradient-Boosting Hybrid with Statistically Validated Accuracy, *European Journal of Computer Science and Information Technology*, 14(5),30-45

Abstract: *Accurate short-term load forecasting underpins unit commitment, economic dispatch, and electricity trading, and forecast error carries direct economic cost. This study proposes Sequential Decomposition–Residual Learning (SDRL), a three-stage hybrid in which a Prophet model extracts trend, multi-scale seasonality, and holiday effects; gradient-boosted trees model the remaining non-linear residual from 55 engineered features; and base learners are combined by an unweighted average. Evaluated on 145,392 hourly observations from the PJME region (2002–2018) under a strict temporal split reserving 40,201 hours for testing, the principal ensemble attains a mean absolute error of 182.92 MW (MAPE 0.563%), the lowest among thirteen benchmarked models, including two contemporary deep architectures. Its 15.52 MW advantage over the strongest non-decomposition baseline is significant under the Diebold–Mariano test ($p < 0.001$) and a moving-block bootstrap. A CPU-only variant attains 190.03 MW, and SHAP attribution elucidates the underlying mechanism, yielding an accurate, interpretable, and reproducible forecaster.*

Keywords: Electricity Load Forecasting, Time-Series Decomposition, Prophet, Gradient Boosting, Ensemble Forecasting, Interpretable Machine Learning

INTRODUCTION

Electricity occupies a singular position among commodities: it cannot be stored economically at scale and must therefore be generated at the moment of consumption. System operation is consequently governed by a continuous balance constraint, whereby total generation must equal total demand plus losses at every instant; a sufficiently large imbalance displaces the system frequency, activates protective schemes, and may ultimately precipitate a blackout (Hong and Fan, 2016). Operators must therefore act in anticipation of demand. Unit commitment determines which generating units to start over the coming hours, economic dispatch allocates output among committed units, reserve sizing establishes the spare capacity held against contingencies, and market operators clear day-ahead and intraday auctions against expected demand. Each of these decisions is conditioned on a short-term load forecast, typically issued one hour to one week ahead at hourly resolution, and forecasting error at this stage propagates into every downstream decision.

The economic consequences of forecast error are substantial and asymmetric. Under-forecasting leaves insufficient capacity committed, compelling costly short-notice procurement or exposing the system to shortfall risk, whereas over-forecasting inflates start-up costs and the fuel consumed by idle plant. Unit-commitment simulation indicates that, within the common operating range, a one-percentage-point reduction in mean absolute percentage error lowers variable generation cost by approximately 0.1 to 0.3 percent, a proportion equivalent to millions of dollars annually for a large system (Hobbs et al., 1999). The forecasting task is, moreover, becoming more demanding: variable renewable generation, electric vehicles, and heat pumps render the net load that operators must track more volatile and less regular than historical demand. Hourly demand nevertheless retains pronounced layered structure, comprising daily, weekly, and annual cycles, calendar and holiday effects, and a markedly non-linear response to temperature. A series whose behaviour is partly determined by such structure is one that can be decomposed, modelled, and forecast.

The principal method families each satisfy only part of the task's requirements. Statistical and decomposition models are transparent and represent seasonality directly, yet their additive, essentially linear formulations cannot express interactions such as the manner in which temperature reshapes the load profile differently across the hours of the day. Gradient-boosted tree ensembles capture precisely such interactions on tabular features, but they possess no intrinsic representation of time and must recover known cyclical structure indirectly from engineered inputs. Deep neural architectures aspire to learn the complete mapping, but they are data- and compute-intensive, largely opaque, and predominantly designed for long multivariate horizons rather than one-step-ahead load; indeed, a single linear layer has been shown to match several Transformer forecasters on standard long-horizon benchmarks (Zeng et al., 2023). Hybrid designs that couple a structural model with a learned component enjoy strong competition evidence (Smyl, 2020), yet a considerable portion of the hybrid literature is evaluated against weak classical baselines, on a single dataset and split, and without tests of statistical significance, leaving the reader unable to judge whether reported gains are systematic or incidental.

This study addresses both the methodological and the evidential deficiency. It proposes Sequential Decomposition–Residual Learning (SDRL), a framework that (i) employs Prophet to remove trend, multi-scale seasonality, and holiday effects from the load series; (ii) trains gradient-boosted residual learners on 55 engineered features, including lags of the decomposition residual itself; and (iii) combines the base learners through a simple, unweighted average that contains no parameters capable of overfitting validation data. The framework is benchmarked against thirteen models, ranging from naïve persistence and ARIMA through standalone gradient boosting to a Temporal Fusion Transformer and an iTransformer, on 16.5 years of PJME data under a strict temporal split, with every claimed difference examined by the Diebold–Mariano procedure and moving-block bootstrap confidence intervals. The contributions are fourfold: a decomposition–residual–ensemble hybrid whose accuracy gain is statistically established rather than asserted; a broad and rigorous benchmark incorporating contemporary deep baselines; an interpretability analysis, comprising SHAP attribution and component ablation, that explains the mechanism of the improvement; and a fully CPU-reproducible variant that achieves near-principal accuracy on commodity hardware.

LITERATURE/THEORETICAL UNDERPINNING

Statistical and Decomposition Methods

The statistical family represents load through explicit equations with few parameters, and its defining strength is transparency. Regression on calendar and meteorological covariates constitutes the simplest formulation, and the seminal analysis of Engle et al. (1986) established that the relationship between electricity sales and temperature is non-linear and weather-driven. The autoregressive integrated moving average model captures short-range autocorrelation within a mature inferential theory (Box and Jenkins, 1970), but presupposes linearity and approximate stationarity and accommodates multiple overlapping seasonal cycles only with difficulty. Exponential smoothing tracks level, trend, and seasonality through recursively weighted averages (Winters, 1960), was subsequently embedded within a general state-space framework (Hyndman et al., 2002), and has been extended with double-seasonal formulations designed expressly for electricity demand (Taylor, 2003). Decomposition methods address structure directly: seasonal-trend decomposition based on Loess separates a series into trend, seasonal, and remainder components (Cleveland et al., 1990), and Prophet operationalises this principle as a practical additive framework comprising a piecewise trend, Fourier-based seasonalities, and holiday terms (Taylor and Letham, 2018). The family's shared limitation is an additive, essentially linear core that cannot represent interaction effects; its shared strength is a clean, interpretable account of structure. This complementarity motivates the use of a decomposition model as a first stage whose residual is delegated to a more flexible learner.

Machine Learning Methods

Machine-learning approaches learn a mapping from engineered features to load and thereby accommodate the non-linearities that statistical specifications omit. Support vector regression was an early and influential contribution (Chen et al., 2004), and shallow neural networks first demonstrated the value of learned non-linearity for load forecasting (Park et al., 1991). Tree-based ensembles have since become the dominant methodology for tabular prediction: random forests reduce variance by averaging decorrelated trees (Breiman, 2001), whereas gradient boosting constructs trees sequentially such that each corrects the errors of the current ensemble (Friedman, 2001). Two implementations define contemporary practice, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017); both model rich feature interactions with limited tuning and train economically on large datasets. Their structural deficiency is the absence of any intrinsic concept of time, trend, or seasonality: cyclical behaviour must be recovered from calendar features and lagged inputs, and a tree-based model cannot extrapolate a trend beyond the range of its training targets, because prediction is formed by averaging within leaves. A decomposition stage remedies precisely this deficiency, supplying the temporal structure that trees lack so that boosting capacity is devoted to the non-linear remainder.

Deep Learning Methods

Deep learning seeks to learn the forecasting function in its entirety. Long short-term memory networks enabled information to be carried across extended spans (Hochreiter and Schmidhuber, 1997) and were applied widely to load forecasting (Kong et al., 2019), with probabilistic formulations such as DeepAR extending the approach to full predictive distributions (Salinas et al., 2020). The Transformer supplanted recurrence with self-attention (Vaswani et al., 2017), prompting a succession of variants directed at long-horizon forecasting: Informer with sparsified attention (Zhou et al., 2021), Autoformer with decomposition embedded in the architecture (Wu et al., 2021), FEDformer operating in the

frequency domain (Zhou et al., 2022), PatchTST with channel-independent patching (Nie et al., 2023), TimesNet with two-dimensional temporal reshaping (Wu et al., 2023), and iTransformer with attention applied across variables rather than time steps (Liu et al., 2024). Efficient designs based on multilayer perceptrons include N-BEATS (Oreshkin et al., 2020) and N-HiTS (Challu et al., 2023), while the Temporal Fusion Transformer couples accuracy with variable-selection interpretability (Lim et al., 2021). A pointed critique demonstrated that a single linear layer matches several of these architectures on standard benchmarks, suggesting that some reported gains derive from experimental configuration rather than architectural substance (Zeng et al., 2023). For short-horizon load forecasting with strong engineered features, three reservations follow: these models are optimised for long multivariate horizons; they require GPU resources and extended training; and, with partial exceptions, they resist interpretation.

Hybrid Methods, Forecast Combination, and Weather Inputs

Two strands of hybrid research underpin the present design. In residual, or error-correction, hybrids, a base model captures one component of the signal and a second model learns what remains; the canonical ARIMA–neural-network pairing outperformed either constituent in isolation (Zhang, 2003). Competition evidence lends strong support: the M4 competition was won by a hybrid of exponential smoothing and a recurrent network (Smyl, 2020), and the competition further established that combinations of methods delivered the most reliable accuracy overall (Makridakis et al., 2020). Combination theory explains the phenomenon: distinct models commit distinct errors, and an average cancels a portion of them (Bates and Granger, 1969). Extensive empirical evidence indicates that a simple equal-weighted average is remarkably difficult to surpass, the so-called forecast-combination puzzle, which motivates an unweighted final stage possessing no estimable weights and hence no capacity to overfit validation data. With respect to exogenous inputs, temperature is the dominant meteorological driver of demand, and its effect follows a flattened V, conventionally encoded through heating and cooling degree-days (Engle et al., 1986); the thermal inertia of buildings additionally induces a lagged response. Reanalysis archives such as ERA5 supply consistent, gap-free hourly temperature fields suitable for this purpose (Hersbach et al., 2020). The research gap is accordingly twofold: no prior study combines a Prophet structural stage, gradient-boosted residual learners, and a simple-average ensemble for hourly load; and much of the hybrid literature lacks contemporary deep baselines, multiple evaluation metrics, and formal significance testing. The present study addresses both deficiencies.

METHODOLOGY

Problem Formulation and Framework Overview

Let y_t denote the electricity load in megawatts at hour t and x_t the corresponding temperature, the sole exogenous covariate employed. The task is one-step-ahead point forecasting: at hour t, the load y_{t+1} is to be predicted using only the information set I_t, which comprises the loads and temperatures observed up to and including t together with the calendar attributes and temperature of the target hour, both of which are available at forecast time, the latter as a meteorological forecast in operational deployment. Conditioning on I_t enforces causality and precludes data leakage by construction.

SDRL constructs the forecaster in three sequential stages (Figure 1). In the first stage, decomposition, Prophet fits the structural component of demand and produces a smooth forecast $\hat{y}_p$ embodying the trend, the daily, weekly, and yearly cycles, and holiday effects. In the second stage, residual learning,

the residual $r_t = y_t - \hat{y}_{p,t}$ retains the sharp, non-linear, weather-driven remainder; a gradient-boosted learner g maps a feature vector φ_t to a residual prediction $\hat{r}_{t+1} = g(\varphi_t)$, and the hybrid forecast is reconstructed as $\hat{y}_{p+1} = \hat{y}_{p,t+1} + \hat{r}_{t+1}$. In the third stage, combination, the base learners are averaged with equal weights, $\hat{y}^* = (1/K) \Sigma \hat{y}_i$. The residual learners are trained under a mean-absolute-error objective, which is robust to occasional demand spikes and closely aligned with operational cost. Two principles govern the entire pipeline: causality, whereby every computation employs only past or known-in-advance information, and reproducibility, whereby all random seeds are fixed and the core pipeline executes on a laptop-class CPU.

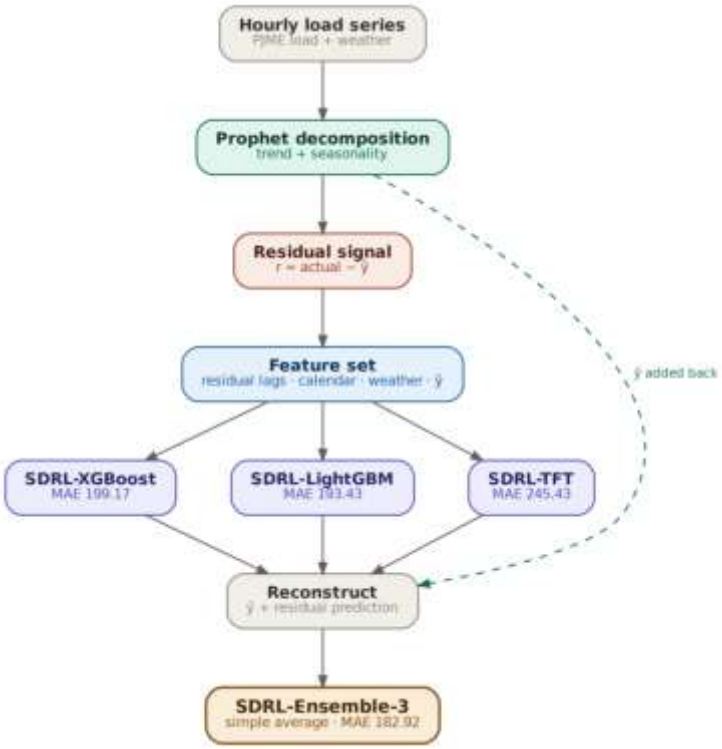

Figure 1. Architecture of the SDRL framework, comprising Prophet decomposition, residual learning by gradient-boosted trees, and hybrid reconstruction; the ensemble stage averages the base learners.

Stage 1: Structural Decomposition with Prophet

Prophet represents the series as $y(t) = g(t)[1 + s(t) + h(t)] + \varepsilon(t)$, where g denotes a piecewise-linear trend, s a sum of Fourier seasonalities, and h a holiday term (Taylor and Letham, 2018). Multiplicative seasonality is adopted so that seasonal and holiday amplitudes scale with the level of the trend. The trend employs 25 changepoints with a changepoint prior scale of 0.05, a configuration sufficiently flexible to accommodate genuine structural shifts, such as the post-2008 decline in demand, without pursuing noise. Three cycles are modelled with Fourier orders of 4 (daily), 3 (weekly), and 10 (yearly); the highest order is assigned to the yearly cycle because its bimodal profile is the most complex, and United States national holidays are each afforded an individual offset. These settings are grounded in the structure of the series documented in Section 3.5, namely the multi-seasonal autocorrelation, the bounded long-run trend confirmed by an STL diagnostic, and stationarity confirmed by Augmented Dickey–Fuller tests. Standalone Prophet is a weak forecaster on this task (MAE of 3,040.29 MW); this

is anticipated rather than problematic, since the model's function within SDRL is not accurate forecasting but the clean removal of structure, such that the residual retains only the difficult component of the problem, and the residual learner absorbs modest variation in the structural fit.

Stage 2: Feature Engineering and Residual Learners

The SDRL residual learner receives 55 features organised in five groups; the standalone gradient-boosting controls employ the 44-feature subset that carries no dependence on Prophet (Table 1). Calendar features (27) encode the position of each hour within its cycles; hour, day of week, month, and day of year are additionally mapped onto the unit circle through sine–cosine pairs, thereby removing the artificial discontinuity between hour 23 and hour 0. Load lags and rolling statistics (10) are selected in accordance with the autocorrelation structure of the series, with lags at 1, 2, 3, 24, 48, 168, and 336 hours motivated by autocorrelations of approximately 0.975, 0.891, and 0.782 at the one-hour, daily, and weekly offsets respectively. Residual lags and rolling statistics (10) constitute the defining elements of SDRL: lags of the Prophet residual endow the learner with a memory of its own recent target, such that a structural forecast running persistently high or low is detected and corrected. Weather features (7) comprise the target-hour temperature; heating and cooling degree-days computed against a 65 °F balance point, encoding the V-shaped response of demand to temperature; a centred quadratic term supplying additional curvature; and one-hour and one-day temperature lags together with a rolling daily mean, capturing the thermal inertia of the building stock. Finally, the structural forecast $\hat{y}_p$ is itself supplied as a feature, affording the learner direct access to the structural level. All lag and rolling features employ values strictly preceding the target hour.

Table 1. Feature groups for the SDRL residual learner (55 features) and the standalone control (44 features).

Feature group	Count	Representative members	In standalone (44)?
Calendar and cyclical	27	hour, dayofweek, month, is_weekend; hour_sin/cos, dow_sin/cos, month_sin/cos, doy_sin/cos	Yes
Load lags and rolling	10	lag_1h, lag_2h, lag_24h, lag_168h; roll_mean_24h, roll_std_24h	Yes
Residual lags and rolling	10	res_lag_1h, res_lag_2h, res_lag_24h; res_roll_mean_24h, res_roll_std_24h	No
Weather	7	temp, HDD, CDD, $(T-65)^2$; temp_lag_1h, temp_lag_24h, temp_roll_mean_24h	Yes
Structural forecast	1	prophet_yhat	No
Total	**55**		

Two gradient-boosting implementations serve as residual learners, ensuring that any observed gain is attributable to the SDRL design rather than to a particular library: XGBoost (Chen and Guestrin, 2016), the primary learner of the eponymous Prophet–XGBoost hybrid, and LightGBM (Ke et al., 2017). Both employ a learning rate of 0.03, a maximum of 3,000 boosting rounds with early-stopping patience of 50, row and column subsampling ratios of 0.8, and L1 and L2 regularisation of 0.1 and 1.0

respectively, with a maximum depth of 8 for XGBoost and 63 leaves for LightGBM, under a mean-absolute-error objective. Training observes a strict temporal protocol: each learner is fitted on data preceding 2013, validated on the calendar year 2013 (8,760 hours) for early stopping, and refitted on all data preceding 2014 with the selected number of rounds; no observation from 2014 onward is consulted at any point. A third residual learner, the Temporal Fusion Transformer (Lim et al., 2021), is trained on GPU hardware under the identical temporal split; although the weakest of the three in isolation, it contributes decorrelated errors to the ensemble.

Stage 3: Ensemble Construction

The base forecasts are combined by a simple average with equal weights. This choice is deliberate: an unweighted average possesses no estimable parameters, cannot overfit a validation set, and exhibits robustness across conditions, in accordance with the forecast-combination literature (Bates and Granger, 1969; Makridakis et al., 2020). Three ensembles are reported. SDRL-Ensemble-GBM averages the two gradient-boosted SDRL learners and executes entirely on CPU hardware. SDRL-Ensemble-3 additionally incorporates the Temporal Fusion Transformer and constitutes the most accurate configuration, at the cost of one GPU-trained member. The Standalone-Ensemble averages the two standalone learners and serves as a control isolating the contribution of decomposition from that of ensembling itself.

Data, Preprocessing, and Evaluation Protocol

The empirical evaluation employs the PJME hourly consumption series published by PJM Interconnection, corresponding to the eastern zone of a system serving approximately sixty-five million people. Following minimal cleaning, the dataset comprises 145,392 hourly observations from January 2002 to August 2018: four duplicate timestamps were removed and thirty missing hours interpolated onto a regular grid; no non-physical values were detected; and no statistical outlier filtering was applied, since the extreme observations are genuine summer cooling peaks and precisely the hours of greatest operational consequence. The series exhibits a mean of 32,079 MW, a standard deviation of 6,464 MW (coefficient of variation 20.2%), a range of 14,544 to 62,009 MW, and positive skewness of 0.739. Hourly temperature is obtained from the Open-Meteo archive, constructed upon the ERA5 reanalysis (Hersbach et al., 2020), for a central Philadelphia location (39.95° N, 75.16° W); the series merges without gaps and spans −7.4 to 103.8 °F. The correlation between load and raw temperature of +0.243 rises to +0.317 when temperature is expressed as the absolute deviation from 65 °F, confirming the V-shaped response that the degree-day features encode (Figure 2).

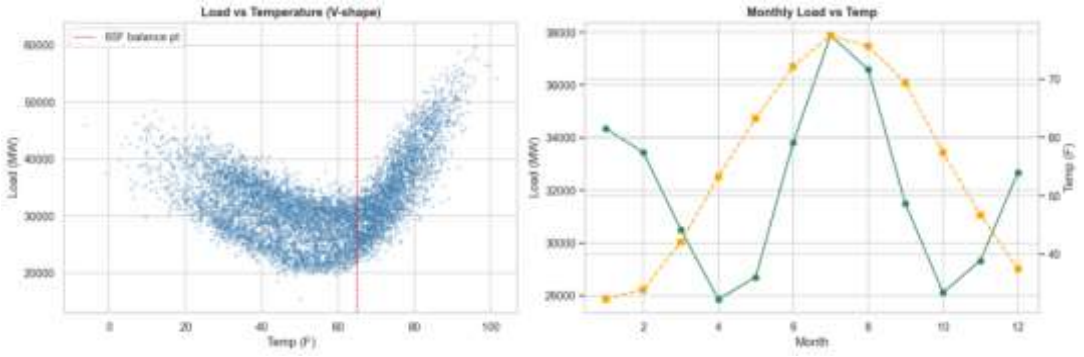

Figure 2. Load as a function of temperature (left), exhibiting the V-shaped response with its minimum near 65 °F, and monthly average load and temperature (right).

The data are partitioned chronologically rather than at random, since random partitioning disperses future observations into the training set and yields optimistic accuracy estimates (Bergmeir and Benítez, 2012). Training spans 2002–2013 (105,191 hours, 72.3%), within which 2013 constitutes the validation block; the test set spans January 2014 to August 2018 (40,201 hours, 27.7%) and encompasses multiple complete seasonal cycles. Because feature construction employs lags of up to 336 hours, the initial fortnight is discarded, reducing the modelling set to 104,855 hours without affecting the test set. Four metrics are reported: MAE, adopted as the primary criterion, RMSE, MAPE, and R^2. Statistical significance is assessed through two independent procedures: the Diebold–Mariano test applied to per-hour loss differentials, conducted separately under absolute-error and squared-error loss (Diebold and Mariano, 1995), and a moving-block bootstrap employing one-week blocks of 168 hours with 3,000 resamples, thereby preserving the daily and weekly autocorrelation of the errors (Künsch, 1989). Significance is adjudicated on paired per-hour differences and never on the overlap of marginal intervals. The gradient-boosting models and Prophet execute on a quad-core 2.0 GHz Intel Core i5 laptop with 16 GB of memory using the CPU alone; the deep models are trained on a Google Colab T4 GPU. All seeds are fixed, model predictions are persisted to disk, and all metrics are recomputed from the persisted files.

RESULTS/FINDINGS

Main Performance Comparison

Table 2 presents the seven strongest of the thirteen evaluated models on the 40,201-hour test set; the reference ladder beneath them comprises naïve persistence (MAE 1,060.64 MW), the seasonal-naïve method (3,310.96 MW), ARIMA (407.13 MW), and standalone Prophet (3,040.29 MW). The leading model is SDRL-Ensemble-3, the simple average of the SDRL-XGBoost, SDRL-LightGBM, and SDRL-TFT hybrids, which attains an MAE of 182.92 MW, a MAPE of 0.563%, and an R^2 of 0.9980, approximately 10% below standalone gradient boosting and 25% below the strongest deep model.

Table 2. Principal performance comparison on the 40,201-hour test set (the seven strongest of thirteen models, ranked by MAE).

Model	MAE (MW)	RMSE (MW)	MAPE (%)	R^2
SDRL-Ensemble-3 (XGB+LGB+TFT)	**182.92**	282.91	**0.563**	0.9980
SDRL-Ensemble-GBM (XGB+LGB)	190.03	308.24	0.578	0.9977
SDRL-LightGBM	193.43	307.13	0.591	0.9977
Standalone-Ensemble (XGB+LGB)	198.44	279.83	0.619	0.9981
SDRL-XGBoost	199.17	323.22	0.606	0.9974
Standalone-XGBoost	203.04	286.20	0.634	0.9980
Standalone-LightGBM	207.31	289.27	0.647	0.9979

Three findings emerge, and each is stated without overstatement. First, decomposition reduces absolute error for every learner: SDRL-LightGBM improves upon its standalone counterpart by approximately 14 MW, SDRL-XGBoost by approximately 4 MW, and the SDRL gradient-boosting ensemble by approximately 8 MW, indicating that the decomposition stage furnishes the residual learner with a more tractable target. Second, the advantage is specific to MAE and MAPE and does not extend to an outright RMSE improvement: the Standalone-Ensemble records the lowest RMSE (279.83 MW), because the residual stage sharpens the fit at typical hours while marginally enlarging a small number of large errors, which squared-error metrics weight disproportionately; the principal ensemble nevertheless recovers most of this margin, recording 282.91 MW, within three megawatts of the minimum. The framework is accordingly claimed to be superior on MAE and MAPE and competitive, though not superior, on RMSE. Third, combination is beneficial, and the strongest combination incorporates a deep member: appending the Temporal Fusion Transformer to the two-tree ensemble reduces MAE from 190.03 to 182.92 MW despite the TFT being the weakest member in isolation, consistent with the expected benefit of averaging learners with decorrelated errors. Figure 3 depicts the principal model over the most demanding interval of the test period, the week containing the test-set peak on 2 July 2018, across which the forecast tracks a swing from approximately 26 to 57 GW; the weekly MAE of 340.82 MW exceeds the test-set average precisely because the interval is the most difficult, and the model is thereby exhibited under stress rather than under favourable conditions.

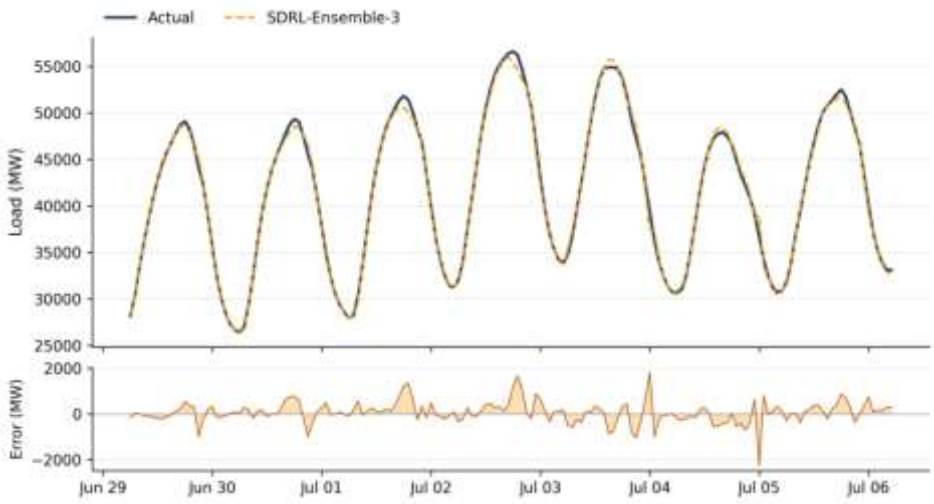

Figure 3. SDRL-Ensemble-3 forecast against actual load over the peak-demand week of the test period (late June to early July 2018); the lower panel presents the hourly errors, which concentrate at the steep afternoon ramps.

Statistical Significance

Since a lower error may arise by chance on a single test set, every comparative claim is subjected to formal testing. Table 3 reports Diebold–Mariano statistics under both loss functions, against the Standalone-XGBoost baseline (Panel A) and against the Standalone-Ensemble control (Panel B). Under absolute-error loss, every SDRL model surpasses its reference at the 0.05 level, and the ensembles do so decisively: SDRL-Ensemble-3 records DM statistics of +10.31 against Standalone-XGBoost and +7.92 against the Standalone-Ensemble, both with $p < 0.0001$. Under squared-error loss the conclusion is deliberately narrower: no Panel A comparison attains significance; the gradient-

boosting-only ensemble is significantly inferior to the Standalone-Ensemble (DM −2.15, p = 0.032); and the principal ensemble is statistically indistinguishable from it (DM −0.48, p = 0.628), indicating that the inclusion of the TFT closes the RMSE deficit that the tree-only ensemble leaves open.

Table 3. Diebold–Mariano test results under absolute-error (MAE) and squared-error (RMSE) loss; positive statistics favour the SDRL model.

Comparison	MAE-loss DM (p)	RMSE-loss DM (p)	Verdict
Panel A — reference: Standalone-XGBoost			
SDRL-LightGBM	+4.02 (0.0001)	−1.72 (0.086)	MAE significantly lower; RMSE tied
SDRL-Ensemble-GBM	+5.11 (<0.0001)	−1.71 (0.088)	MAE significantly lower; RMSE tied
SDRL-Ensemble-3	+10.31 (<0.0001)	+0.53 (0.596)	MAE significantly lower; RMSE tied
Panel B — reference: Standalone-Ensemble			
SDRL-Ensemble-GBM	+3.26 (0.0011)	−2.15 (0.032)	MAE significantly lower; RMSE significantly worse
SDRL-Ensemble-3	+7.92 (<0.0001)	−0.48 (0.628)	MAE significantly lower; RMSE tied

The moving-block bootstrap corroborates this verdict through an independent mechanism. The principal MAE of 182.92 MW is accompanied by a 95% interval of [173.06, 192.28] MW. Because competing models forecast identical hours with correlated errors, significance is adjudicated on the paired per-hour difference in absolute error, which cancels the shared variance: relative to the Standalone-Ensemble, SDRL-Ensemble-3 is more accurate by a mean of 15.52 MW with a 95% interval of [10.26, 20.83], and relative to its nearest competitor, SDRL-Ensemble-GBM, by 7.11 MW with an interval of [3.25, 11.78] (Figure 4). Both intervals lie entirely above zero. It should be emphasised that overlap among the marginal intervals of the three ensembles must not be interpreted as equivalence; inference from the overlap of marginal intervals understates significance when errors are correlated, and the practice is avoided here.

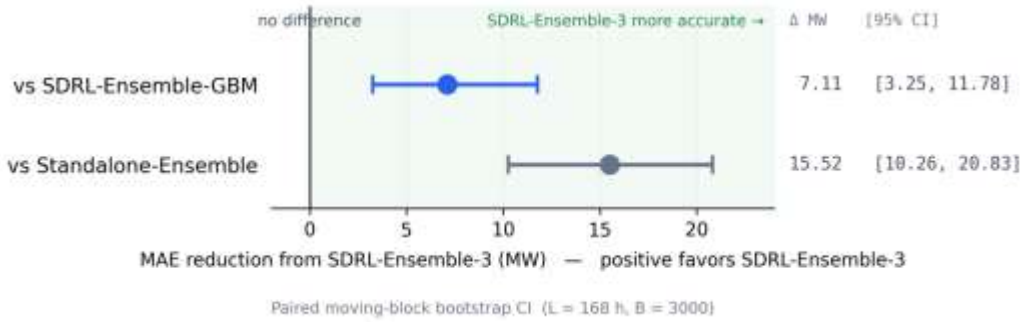

Figure 4. Paired per-hour MAE differences relative to SDRL-Ensemble-3, with 95% moving-block bootstrap confidence intervals (block length 168 hours; 3,000 resamples).

Deep-Learning Residual Backbones and Computational Cost

The residual-learner slot within SDRL is modular, permitting deep architectures to be substituted while all other elements remain fixed. The Temporal Fusion Transformer trained without difficulty to an MAE of 245.43 MW (RMSE 393.09 MW), a creditable result that nonetheless trails both gradient-boosted residual learners. The iTransformer proved markedly weaker at 711.90 MW, inferior even to classical ARIMA, its cross-variable attention failing to exploit the engineered tabular features as effectively as the tree-based models. An LSTM backbone failed to converge under the shared protocol; this outcome is reported as a failure rather than suppressed, since it is itself informative regarding robustness. The computational dimension sharpens the comparison: Prophet is fitted in approximately 1.3 minutes and the four gradient-boosting learners are trained in a measured 360.7 seconds on a laptop CPU, whereas the TFT requires approximately 7.5 minutes on a T4 GPU. The CPU-only SDRL-Ensemble-GBM (190.03 MW) consequently surpasses the GPU-trained transformer employed as a standalone model (245.43 MW) at a fraction of the training cost, and the transformer justifies its inclusion solely as one decorrelated member of the full ensemble, in which capacity it contributes a further improvement of approximately seven megawatts.

Ablation, Feature Attribution, and Error Structure

An ablation study removes each feature group in turn from SDRL-XGBoost (full-model MAE of 199.17 MW) and retrains under the identical protocol (Table 4). Lag information dominates all other contributions: removing the load and residual lags jointly raises the MAE by 742 MW, reducing the model to little more than a calendar climatology, as anticipated for a series whose lag-one autocorrelation is 0.975. The two lag-related rows warrant joint interpretation. Removing the residual lags alone is severely damaging (+105.63 MW) even though removing the decomposition entirely costs merely +3.87 MW; the former preserves the residual-learning architecture while depriving it of its most informative inputs, whereas the latter reverts to direct load prediction, in which the raw load lags convey the autoregressive signal themselves. Residual lags are therefore critical conditional upon the framework, whereas the decomposition step in isolation yields only a modest single-model gain; the larger improvements in Table 2 derive from the combination of diverse learners. Removal of the calendar features costs 80.98 MW, and removal of the weather features 6.22 MW; the latter contribution is modest because recent load already encodes the greater part of the thermal response, such that weather information refines the forecast principally during heat waves and cold spells. Removing the decomposition additionally improves the RMSE from 323.22 to 286.20 MW, exposing at the level of a single learner the same MAE–RMSE trade-off observed among the ensembles.

Table 4. Feature-group ablation of SDRL-XGBoost on the test set.

Configuration	MAE (MW)	Change vs. full (MW)
SDRL-XGBoost (full, 55 features)	199.17	—
− all lag features (load + residual)	≈ 941	+742
− residual lag/rolling features	304.80	+105.63
− calendar features	280.15	+80.98
− weather features	205.39	+6.22
− decomposition (= Standalone-XGBoost)	203.04	+3.87

SHAP attribution renders the mechanism explicit. Because the principal ensemble comprises heterogeneous tree-based and neural members, no single exact attribution exists for it; the analysis therefore addresses the SDRL-XGBoost residual learner, for which TreeSHAP is exact (Lundberg et al., 2020), over a fixed random sample of 8,000 test hours. The one-hour residual lag dominates, with a mean absolute SHAP value of 2,464 MW, approximately eight times that of any other feature: the residual learner operates, in essence, as an autoregressive corrector that propagates the most recent deviation from the structural forecast (Figure 5). Beneath it, the two-hour residual lag (321 MW) extends the same logic; the cyclical hour encoding outranks the raw integer hour (89 against 85 MW), validating the encoding design; the quadratic temperature term (84 MW) and cooling degree-days (81 MW) furnish the weather correction; and the Prophet forecast itself contributes approximately 70 MW, confirming the dual role of the decomposition. The error structure completes the account: errors concentrate where demand is elevated and rapidly changing, at the evening peak (301.40 MW at 19:00, the most difficult hour, with an RMSE of 423.85 MW) and the morning ramp (270.43 MW at 09:00), against 92.78 MW in the pre-dawn trough; the temperature-extreme months of January (256.81 MW) and July (210.37 MW) are the most demanding, whereas October (122.98 MW) is the least; peak hours average 209.38 MW against 151.65 MW off-peak; and error rises gradually across the horizon, from 164.45 MW in 2014 to 229.58 MW in the partial year 2018, the anticipated signature of forecasting at increasing distance from the training period.

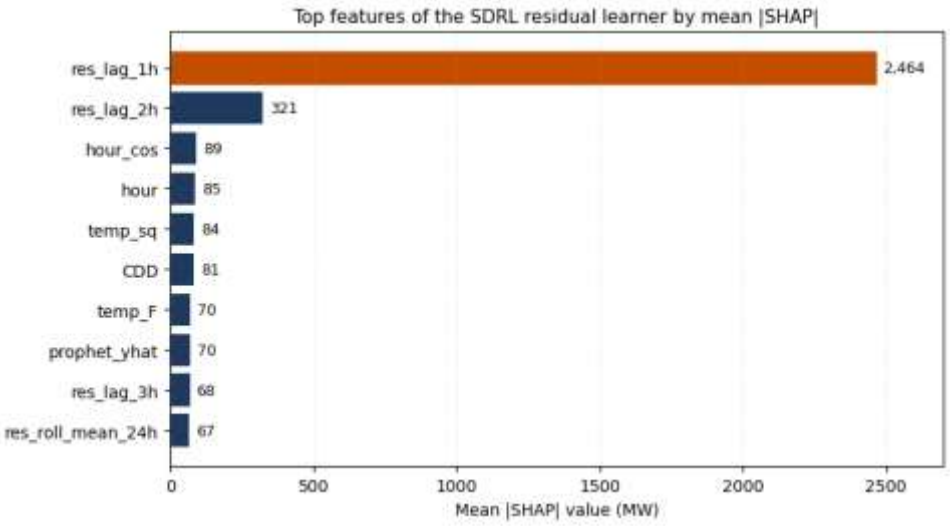

Figure 5. Mean absolute SHAP values (MW) for the SDRL-XGBoost residual learner, computed with exact TreeSHAP over a fixed 8,000-hour test sample.

DISCUSSION

The results admit a precise account of why the design succeeds. Decomposition contributes no additional information; rather, it reshapes the learning target. The Prophet stage supplies the learner with a residual that is strongly autocorrelated at short lags and purged of trend and seasonality, and the SHAP analysis demonstrates that the learner exploits exactly this property: recent residual memory constitutes the principal driver of the forecast, refined by hour-of-day and temperature terms. The same account explains the metric-specific character of the improvement. Optimising a more tractable target under an absolute-error objective tightens the fit at typical hours, which MAE and MAPE reward, whereas the small number of hours at which the structural fit is misaligned receive marginally larger

erroneous corrections, which squared-error metrics penalise disproportionately. The deep ensemble member, whose errors are differently distributed, offsets precisely those hours, which explains why SDRL-Ensemble-3 incurs no significant RMSE penalty while the tree-only ensemble does.

The comparison with deep learning warrants careful formulation. On this one-step, feature-rich, single-series task, gradient-boosted trees proved the more dependable residual learner: faster, more stable, and more accurate than either transformer, with one deep backbone failing outright under the shared protocol. This observation accords with the linear-baseline critique of long-horizon transformer benchmarks (Zeng et al., 2023) and with the established strength of tree ensembles on tabular data, yet it is asserted narrowly: the deep models were trained under a common, reasonable protocol rather than an exhaustive architecture-specific search, and the verdict is therefore specific to the computational budget employed. The productive role identified for the deep model is that of a source of decorrelated errors within a combination rather than a replacement for the trees; that a weaker but differently erring model improves the aggregate is itself a practically valuable observation, consistent with classical combination theory (Bates and Granger, 1969).

The evaluation design likewise carries a methodological implication for the hybrid-forecasting literature. Two independent significance procedures, a Diebold–Mariano test on paired loss differentials and a moving-block bootstrap on paired differences, concur in confining the claim to absolute error, and the boundary they delineate, namely that average accuracy is improved while worst-case accuracy is statistically unchanged, is more informative than an undifferentiated assertion of superiority. Employing the strongest non-decomposition ensemble as the control, rather than a weak classical baseline, is what renders the attribution of the observed gain to the decomposition stage credible.

IMPLICATION TO RESEARCH AND PRACTICE

For practice, the cost profile of the framework constitutes a deployable advantage. The complete CPU pipeline, comprising Prophet and both gradient-boosted learners, trains within minutes on an ordinary laptop and attains an MAE of 190.03 MW; a utility operating on commodity hardware may therefore retrain with sufficient frequency to track evolving demand patterns without GPU infrastructure, reserving the GPU-trained ensemble member for settings in which a further seven megawatts of average accuracy justifies the additional hardware. The model is furthermore auditable in a form that regulators and control-room engineers can employ: its accuracy rests upon a compact set of transparent inputs whose contributions are quantified exactly, and the error segmentation identifies the conditions under which the model is least reliable, namely the morning and evening ramps and the temperature-extreme months, which are precisely the circumstances in which operator scrutiny of any forecast should be concentrated.

For research, the study contributes two elements of general applicability beyond the model itself. The first is a modular residual-learning slot: any architecture mapping features to the decomposition residual may be evaluated within an identical protocol, as demonstrated here through the benchmarking of two transformers and an LSTM in that position. The second is an evidential standard for hybrid studies, comprising a strict temporal split, a control condition that isolates the claimed mechanism, loss-specific Diebold–Mariano tests, and paired block-bootstrap intervals. Broader adoption of this standard would render reported gains in the load-forecasting literature substantially more comparable and more credible.

CONCLUSION

This study has proposed Sequential Decomposition–Residual Learning, a hybrid hourly load forecaster in which Prophet removes trend, multi-scale seasonality, and holiday structure; gradient-boosted trees model the non-linear residual from 55 engineered features; and the base learners are combined by an unweighted average. On 16.5 years of PJME data under a strict temporal split, the principal configuration, SDRL-Ensemble-3, achieved the lowest absolute and percentage error among thirteen benchmarked models (MAE 182.92 MW, MAPE 0.563%, R^2 0.9980), improving upon the strongest non-decomposition control by 15.52 MW, an advantage confirmed by the Diebold–Mariano test ($p < 0.001$) and by a moving-block bootstrap interval of [10.26, 20.83] that excludes zero, while remaining statistically indistinguishable on RMSE. A fully CPU-reproducible variant attained 190.03 MW. Ablation and exact SHAP attribution elucidated the underlying mechanism, identifying recent residual memory as the dominant driver, such that the design is explained rather than merely asserted. The claims are deliberately bounded: the demonstrated gain concerns average rather than worst-case accuracy; the evidence derives from a single grid region and a single chronological split; the model consumes measured rather than forecast temperature; and the trees-versus-deep verdict is specific to the computational budget employed. Within this stated scope, SDRL constitutes an accurate, interpretable, statistically validated, and inexpensive approach to short-term load forecasting, whose contribution resides as much in the discipline of its evaluation as in the accuracy of its forecasts.

FUTURE RESEARCH

Five avenues follow directly from the stated boundaries. First, cross-market validation: evaluating SDRL on systems with contrasting climates, including a cooling-dominated system such as ERCOT, a heating-dominated northern system, and national-scale series such as the Spanish grid or ENTSO-E data, together with public benchmarks such as the GEFCom competition datasets under rolling-origin cross-validation, would distinguish a genuine property of the method from an artefact of a single series. Second, extended horizons: adapting the framework to day-ahead forecasting requires residual features that do not depend upon the most recent actual observations, through recursive or direct multi-step formulations, and the advantage should be expected to diminish as the dominant one-hour residual lag becomes unavailable. Third, operational deployment: consuming forecast rather than measured weather, and extending the accompanying results dashboard into a live service with scheduled retraining, would close the gap between offline study and operational tool. Fourth, more exacting deep comparisons: placing recent architectures such as PatchTST and TimesNet in the modular residual position under an enlarged tuning budget would revisit the trees-versus-deep question on stronger terms. Fifth, tail-aware objectives: the design of a peak-weighted loss function could extend the demonstrated advantage from average accuracy to the worst-case accuracy of greatest consequence for grid reliability.

REFERENCES

Bates, J. M. and Granger, C. W. J. (1969) The combination of forecasts, Operational Research Quarterly, 20 (4), 451-468.

Bergmeir, C. and Benítez, J. M. (2012) On the use of cross-validation for time series predictor evaluation, Information Sciences, 191, 192-213.

Box, G. E. P. and Jenkins, G. M. (1970) Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco.

Breiman, L. (2001) Random forests, Machine Learning, 45 (1), 5-32.

Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler-Canseco, M. and Dubrawski, A. (2023) N-HiTS: Neural hierarchical interpolation for time series forecasting, In Proceedings of the AAAI Conference on Artificial Intelligence, 37 (6), 6989-6997.

Chen, B.-J., Chang, M.-W. and Lin, C.-J. (2004) Load forecasting using support vector machines: a study on EUNITE competition 2001, IEEE Transactions on Power Systems, 19 (4), 1821-1830.

Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system, In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, USA, pp. 785-794.

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990) STL: A seasonal-trend decomposition procedure based on Loess, Journal of Official Statistics, 6 (1), 3-73.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy, Journal of Business and Economic Statistics, 13 (3), 253-263.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986) Semiparametric estimates of the relation between weather and electricity sales, Journal of the American Statistical Association, 81 (394), 310-320.

Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine, Annals of Statistics, 29 (5), 1189-1232.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146 (730), 1999-2049.

Hobbs, B. F., Jitprapaikulsarn, S., Konda, S., Chankong, V., Loparo, K. A. and Maratukulam, D. J. (1999) Analysis of the value for unit commitment of improved load forecasts, IEEE Transactions on Power Systems, 14 (4), 1342-1348.

Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory, Neural Computation, 9 (8), 1735-1780.

Hong, T. and Fan, S. (2016) Probabilistic electric load forecasting: A tutorial review, International Journal of Forecasting, 32 (3), 914-938.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. and Grose, S. (2002) A state space framework for automatic forecasting using exponential smoothing methods, International Journal of Forecasting, 18 (3), 439-454.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) LightGBM: A highly efficient gradient boosting decision tree, In Advances in Neural Information Processing Systems 30, Long Beach, USA, pp. 3146-3154.

Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y. and Zhang, Y. (2019) Short-term residential load forecasting based on LSTM recurrent neural network, IEEE Transactions on Smart Grid, 10 (1), 841-851.

Künsch, H. R. (1989) The jackknife and the bootstrap for general stationary observations, Annals of Statistics, 17 (3), 1217-1241.

Lim, B., Arık, S. Ö., Loeff, N. and Pfister, T. (2021) Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting, 37 (4), 1748-1764.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L. and Long, M. (2024) iTransformer: Inverted transformers are effective for time series forecasting, In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020) From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2 (1), 56-67.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020) The M4 Competition: 100,000 time series and 61 forecasting methods, International Journal of Forecasting, 36 (1), 54-74.

Nie, Y., Nguyen, N. H., Sinthong, P. and Kalagnanam, J. (2023) A time series is worth 64 words: Long-term forecasting with transformers, In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda.

Oreshkin, B. N., Carpov, D., Chapados, N. and Bengio, Y. (2020) N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia.

Park, D. C., El-Sharkawi, M. A., Marks, R. J., Atlas, L. E. and Damborg, M. J. (1991) Electric load forecasting using an artificial neural network, IEEE Transactions on Power Systems, 6 (2), 442-449.

Salinas, D., Flunkert, V., Gasthaus, J. and Januschowski, T. (2020) DeepAR: Probabilistic forecasting with autoregressive recurrent networks, International Journal of Forecasting, 36 (3), 1181-1191.

Smyl, S. (2020) A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting, International Journal of Forecasting, 36 (1), 75-85.

Taylor, J. W. (2003) Short-term electricity demand forecasting using double seasonal exponential smoothing, Journal of the Operational Research Society, 54 (8), 799-805.

Taylor, S. J. and Letham, B. (2018) Forecasting at scale, The American Statistician, 72 (1), 37-45.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017) Attention is all you need, In Advances in Neural Information Processing Systems 30, Long Beach, USA, pp. 5998-6008.

Winters, P. R. (1960) Forecasting sales by exponentially weighted moving averages, Management Science, 6 (3), 324-342.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J. and Long, M. (2023) TimesNet: Temporal 2D-variation modeling for general time series analysis, In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda.

Wu, H., Xu, J., Wang, J. and Long, M. (2021) Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, In Advances in Neural Information Processing Systems 34, pp. 22419-22430.

Zeng, A., Chen, M., Zhang, L. and Xu, Q. (2023) Are transformers effective for time series forecasting?, In Proceedings of the AAAI Conference on Artificial Intelligence, 37 (9), 11121-11128.

Zhang, G. P. (2003) Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, 50, 159-175.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. and Zhang, W. (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting, In Proceedings of the AAAI Conference on Artificial Intelligence, 35 (12), 11106-11115.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L. and Jin, R. (2022) FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting, In Proceedings of the 39th International Conference on Machine Learning, PMLR, Baltimore, USA, pp. 27268-27286.