Fi-sLSTM-Mixer: Multivariate Time Series Forecasting with Fuzzy-Conditioned Scalar Memories

Anonymous authors
Paper under double-blind review

Abstract

Multivariate time series forecasting in real-world deployments must contend with noisy, uncertain, and shifting data conditions that expose a structural weakness in state-of-the-art recurrent architectures: their gates rely on deterministic pre-activations and cannot adapt to input reliability. We introduce Fi-sLSTM-Mixer, a Fuzzy-integrated sLSTM which augments xLSTM-Mixer with a fuzzy relevance value r_t derived from an ITTTFL inference system and injected into the forget and output gate pre-activations via a zero-initialized projection $\mathbf{W}_r$. The normalized input gate is provably invariant to r_t by construction, producing a clean separation between data-driven variate attention and reliability-driven memory modulation. A twelve-test mechanistic protocol across 12 benchmarks confirms that the model learns consistent, domain-interpretable routing policies, statistically significant on every dataset, without sacrificing the backbone's representational capacity. Empirically, Fi-sLSTM-Mixer achieves 41 wins across 90 metric slots against five state-of-the-art baselines, with the largest gains on volatile industrial and high-dimensional streams where reliability signals matter most, at a cost of under 600 additional parameters and negligible training overhead.

1 Introduction

Time series forecasting is a central challenge in fields ranging from industrial predictive maintenance and traffic management to epidemiology and energy systems (Hosseini et al., 2021; Lippi et al., 2013; Lam et al., 2023; Essien & Giannetti, 2020). Despite rapid progress, models deployed in these domains face a condition that standard benchmarks understate: observations are frequently noisy, partially degraded, or drawn from operational regimes not seen during training. A model with no mechanism to distinguish reliable from corrupted inputs accumulates errors silently and provides no signal to downstream decision-makers about its own confidence.

The dominant paradigm shifted toward Transformer architectures (Vaswani et al., 2017), whose self-attention provides rich global context but incurs quadratic cost in sequence length and variate count. Transformer-based approaches including PatchTST (Nie et al., 2022), iTransformer (Liu et al., 2023), and FEDformer (Zhou et al., 2022) remain widely used. This burden has motivated a resurgence of efficient alternatives: recurrent models from LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014) to xLSTM (Beck et al., 2024) and xLSTM-Mixer (Kraus et al., 2024); state-space models including Mamba (Gu & Dao, 2024), S-Mamba (Wang et al., 2025), and Chimera (Behrouz et al., 2024); MLP-based baselines such as DLinear (Zeng et al., 2023) and TiDE (Das et al., 2023); and mixing architectures (Tolstikhin et al., 2021) such as TimeMixer (Wang et al., 2024) and TSMixer (Chen et al., 2023). xLSTM-Mixer represents the current state of the art among these: it combines scalar-memory sLSTM blocks with exponential gating and variate-striding mixing, achieving top accuracy at one to two orders of magnitude lower GPU memory than Transformer methods and outperforming Chimera and S-Mamba across standard benchmarks.

Despite this efficiency, xLSTM-Mixer and related crisp recurrent architectures share a structural vulnerability: gate pre-activations are derived exclusively from deterministic linear projections with no mechanism to distinguish reliable observations from corrupted ones. Fuzzy-LSTM hybrids including FIS-LSTM (Sup-

piah et al., 2022), FD-LSTM (Langeroudi et al., 2022), and FLSTM (Wang et al., 2023) demonstrate that rule-based signals injected into gating mechanisms improve robustness. More recently, FiLSTM (Kerarmi et al., 2025a) replaces static rules with an ITTTFL-based system (Integrated Truth Table in Tree-based Fuzzy Logic) that extracts optimized membership functions from data via decision-tree induction (Kerarmi et al., 2022; 2024; 2025b), yielding a continuously valued r_t that allows the recurrent model to modulate memory dynamics in response to structured uncertainty.

We propose Fi-sLSTM-Mixer, which integrates fuzzy guidance natively into the xLSTM-Mixer topology. A zero-initialized projection $\mathbf{W}_r$ maps r_t to additive offsets on the forget and output gate pre-activations; a companion gate $\alpha_r = \sigma(\ell_r)$ suppresses the fuzzy pathway where r_t is uninformative. The normalized input gate is provably invariant to these shifts (Eq. 15), structurally separating data-driven variate attention from reliability-driven memory modulation. An optional decomposition skip on the target variate provides a structured base forecast on linearly dominated datasets, disabled on Traffic and Electricity where high dimensionality causes numerical instability.

Across 90 metric slots, Fi-sLSTM-Mixer achieves **41 total wins** versus 26 for xLSTM-Mixer and 16 for DLinear, with MI z-scores exceeding 3σ on 11 of 12 datasets ($p < 10^{-13}$). Our contributions are:

- **Architecture.** Fi-sLSTM-Mixer: a zero-initialized $\mathbf{W}_r$ injects ITTTFL-derived r_t into sLSTM forget and output gate pre-activations, with a learned adaptive suppression gate α_r.
- **Theory.** The normalized input gate is provably invariant to additive r_t shifts (Eq. 15), cleanly separating variate attention from reliability modulation.
- **Interpretability.** A twelve-test mechanistic protocol across 12 datasets confirms domain-consistent, signed routing policies learned entirely from gradient descent.
- **Empirical.** 41 wins across 90 metric slots at a cost of $\leq$577 additional parameters and $\leq$10% training overhead over xLSTM-Mixer.

2 Background and Model

Fi-sLSTM-Mixer combines the ITTTFL fuzzy inference framework, which produces an interpretable per-timestep relevance scalar r_t, with a variate-striding recurrent pipeline that mixes time and variate information in linear time. We describe each component in forward-pass order.

2.1 The ITTTFL Fuzzy Inference System

The fuzzy relevance value r_t is a continuously valued scalar computed once as a preprocessing step over the training split by ITTTFL (Kerarmi et al., 2022; 2024; 2025b). A decision tree is grown via entropy-based splitting (Quinlan, 1996); each root-to-leaf path defines an IF-THEN rule mapping input feature intervals to a fuzzy class label of the target, which is then converted into a fuzzy membership function. Redundant rules are pruned by interval-inclusion analysis while preserving full semantic coverage. At inference, each surviving rule fires with a strength equal to the product of its membership degrees, and r_t is the firing-strength-weighted average of rule consequents:

$$r_t \;=\; \frac{\sum_i w_i\, y_i}{\sum_i w_i}, \qquad w_i \;=\; \prod_j \mu_{A_{ij}}(x_{tj}). \tag{1}$$

This yields an enriched dataset $\{(t, \mathbf{x}_t, y_t, r_t)\}_{t=1}^{T}$ in which r_t carries a rule-consistent, linguistically interpretable approximation of the target that the recurrent gates can condition on at every step. ITTTFL hyperparameter choices are reported in Table 2, Section 3.2.

2.2 Scalar-Memory sLSTM Backbone

xLSTM-Mixer (Kraus et al., 2024) uses the scalar-memory sLSTM variant (Beck et al., 2024), which preserves hidden-to-hidden gate recurrences essential for history-aware updates when iterating over the

variate axis. Unlike mLSTM, the sLSTM maintains recurrent connections enabling conditional, history-dependent updates critical for tracking latent phase transitions in long-horizon forecasting. Given input token $\mathbf{x}_t$ and previous hidden state $\mathbf{h}_{t-1}$, the sLSTM updates are:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{z}_t, \qquad \text{(cell state)} \tag{2}$$

$$\mathbf{n}_t = \mathbf{f}_t \odot \mathbf{n}_{t-1} + \mathbf{i}_t, \qquad \text{(normalizer)} \tag{3}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \mathbf{c}_t \odot \mathbf{n}_t^{-1}, \qquad \text{(hidden state)} \tag{4}$$

$$\mathbf{z}_t = \tanh(\mathbf{W}_z \mathbf{x}_t + \mathbf{R}_z \mathbf{h}_{t-1} + \mathbf{b}_z), \qquad \text{(cell input)} \tag{5}$$

with exponential input and forget gates:

$$\mathbf{i}_t = \exp(\tilde{\mathbf{i}}_t - \mathbf{m}_t), \qquad \tilde{\mathbf{i}}_t = \mathbf{W}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{h}_{t-1} + \mathbf{b}_i, \tag{6}$$

$$\mathbf{f}_t = \exp(\tilde{\mathbf{f}}_t + \mathbf{m}_{t-1} - \mathbf{m}_t), \qquad \tilde{\mathbf{f}}_t = \mathbf{W}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{h}_{t-1} + \mathbf{b}_f, \tag{7}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{h}_{t-1} + \mathbf{b}_o), \qquad \mathbf{m}_t = \max(\tilde{\mathbf{f}}_t + \mathbf{m}_{t-1}, \tilde{\mathbf{i}}_t) \tag{8}$$

Here $\mathbf{W}_\cdot$ are input weights, $\mathbf{R}_\cdot$ are recurrent weights (implemented as block-diagonal to enable efficient per-head specialization (Beck et al., 2024)), and $\mathbf{n}_t$, $\mathbf{m}_t$ serve as normalization and numerical stabilization states respectively.

Parallel log-domain form. Following (Kraus et al., 2024), we evaluate the sLSTM in a numerically equivalent log-domain form that removes the sequential recurrence over the variate axis. Defining the unnormalized input gate $ig_t = \exp(\tilde{\mathbf{i}}_t + \sum_{s=1}^{t} \log \sigma(\tilde{\mathbf{f}}_s) - m)$ with global maximum stabilizer m, the normalized input gate is:

$$\boldsymbol{\alpha}_t \;=\; \frac{ig_t}{n_t} \;=\; \frac{\exp\!\big(\tilde{\mathbf{i}}_t + \sum_{s=1}^{t} \log \sigma(\tilde{\mathbf{f}}_s) - m\big)}{\mathrm{clamp}\big(\sum_{s=1}^{t} ig_s,\ \min{=}1\big)}. \tag{9}$$

A separate LSTM pass over the input provides a context sequence $\mathbf{H}$; the value pathway $\mathbf{v}_t = \mathrm{SiLU}(\mathbf{W}_v \mathbf{H}_t)$ draws representations from this context. The full block output is then:

$$\mathbf{y} \;=\; \mathrm{LayerNorm}\big(\mathbf{x} + \mathrm{Dropout}(\mathbf{o}_t \odot \boldsymbol{\alpha}_t \odot \mathbf{v}_t)\big). \tag{10}$$

2.3 FisLSTM: Fuzzy-Conditioned Gating

The FisLSTM block augments the sLSTM with a projection $\mathbf{W}_r \in \mathbb{R}^{1\times 3D}$ that maps the scalar r_t to three additive gate offsets:

$$[\,\mathbf{r}_i \;|\; \mathbf{r}_f \;|\; \mathbf{r}_o\,] \;=\; \mathbf{W}_r\, r_t + \mathbf{b}_r, \qquad \mathbf{W}_r,\ \mathbf{b}_r \;=\; \mathbf{0} \text{ at initialization.} \tag{11}$$

Zero initialization ensures that r_t has no effect at epoch 0; any performance gap that emerges during training is therefore attributable solely to what the model learned to route through r_t.

Adaptive gate. A scalar logit ℓ_r produces a gate $\alpha_r = \sigma(\ell_r)$ that multiplies all three offsets, allowing the model to suppress the r_t pathway on datasets where the fuzzy signal adds no value. The modified pre-activations are (additions relative to the sLSTM baseline in red):

$$\tilde{\mathbf{i}}_t = \mathbf{W}_i \mathbf{e}_t + \mathbf{b}_i + \alpha_r \mathbf{r}_i, \tag{12}$$

$$\tilde{\mathbf{f}}_t = \mathbf{W}_f \mathbf{e}_t + \mathbf{b}_f + \alpha_r \mathbf{r}_f, \tag{13}$$

$$\mathbf{o}_t = \sigma\big(\mathbf{W}_o \mathbf{e}_t + \mathbf{b}_o + \alpha_r \mathbf{r}_o\big). \tag{14}$$

where $\mathbf{e}_t = \mathrm{LayerNorm}(\mathbf{x}_t)$. This injection mirrors the gate-modulation strategy of FiLSTM (Kerarmi et al., 2025a), with three key differences: (i) the backbone uses exponential rather than sigmoid gating, (ii) α_r can suppress the pathway entirely, and (iii) the invariance result (Eq. 15) provides a structural guarantee absent in the sigmoid setting. The parameter overhead is at most $3D + 1$ scalars per block, independent of variate count or sequence length.

Invariance of the normalized input gate. An additive shift $\alpha_r \mathbf{r}_i$ on $\tilde{\mathbf{i}}_t$ rescales every unnormalized gate value ig_s by the same constant, which cancels exactly in the normalized gate:

$$\boldsymbol{\alpha}_t \;=\; \frac{\exp(\alpha_r \mathbf{r}_i) \cdot ig_t^0}{\exp(\alpha_r \mathbf{r}_i) \cdot n_t^0} \;=\; \boldsymbol{\alpha}_t^0. \tag{15}$$

This is a structural consequence of exponential normalization, not a training outcome: the input gate remains a purely data-driven variate attention mechanism regardless of r_t. The fuzzy signal instead enters through the forget gate (Eq. 13) and output gate (Eq. 14), where additive shifts act directly on sigmoid activations, the natural routes for reliability-modulated memory retention and selective readout. Section 3.3.4 confirms this separation empirically across all datasets.

Table 1: Architectural comparison of LSTM, xLSTM-Mixer, and Fi-sLSTM-Mixer.

Feature	LSTM	xLSTM-Mixer	Fi-sLSTM-Mixer (Ours)
Memory cell	Scalar c_t	Scalar c_t (sLSTM stack)	Scalar c_t with fuzzy-conditioned updates
Gating	Sigmoid / tanh (crisp)	Exponential, data-only	Exponential, fuzzy-shifted: $\tilde{f}_t, \tilde{i}_t, o_t \mathrel{+}= \alpha_r \boldsymbol{W}_r r_t$
Gate input	x_t, h_{t-1}	x_t, h_{t-1}	x_t, h_{t-1}, rule-derived r_t via adaptive gate α_r
Normalized input gate	Not applicable	$\alpha_t = ig_t / n_t$	α_t invariant to r_t by construction (Eq. 15)
Adaptive r_t suppression	None	None	Gate $\alpha_r{=}\sigma(\ell_r)$, learned - no manual toggle
Mixing topology	Sequential over time	Bidirectional variate striding + multi-view fusion	Same backbone; block-internal gate shifts from fuzzy r_t
Target pathway	Single end-to-end head	Shared linear head	decomp. skip base $\mathbf{b}$ with gate α_d + λ-scaled fuzzy residual
Adaptive decomp. skip use	None	None	Gate $\alpha_d{=}\sigma(\ell_d)$: opens on linear data, closes on nonlinear
Normalization	Batch or layer	RevIN	RevIN with explicit inverse at output
Logic source	Black-box	Black-box	Hybrid: data-driven attention + IF-THEN fuzzy rules
Interpretability	Low	Low	High: per-gate routing of r_t inspectable via $\boldsymbol{W}_r$, α_r, α_d
Training robustness	None	None	r_t-dropout, zero-init $\boldsymbol{W}_r$, zero-init ℓ_r and ℓ_d

2.4 The Fi-sLSTM-Mixer Pipeline

The FisLSTM block is embedded in a four-stage pipeline described below in forward-pass order.

Normalization. The multivariate input $\mathbf{X} \in \mathbb{R}^{B \times T \times V}$ is normalized via Reversible Instance Normalization (RevIN) (Kim et al., 2022): per-instance mean subtraction and standard deviation scaling, followed by a learnable affine rescaling. The inverse transform is applied at the output to restore the original scale.

Time mixing and structured base forecast. Following (Kraus et al., 2024), a shared NLinear projection (Zeng et al., 2023) maps each variate's lookback to an initial forecast, with weights tied across all V variates:

$$\boldsymbol{x}^{\text{init}} \;=\; \text{NLinear}(\boldsymbol{x}^{\text{norm}}) \;=\; \text{FC}\big(\boldsymbol{x}^{\text{norm}}_{1:T} - \boldsymbol{x}^{\text{norm}}_{T}\big) + \boldsymbol{x}^{\text{norm}}_{T}, \tag{16}$$

where $\text{FC} : \mathbb{R}^T \to \mathbb{R}^H$ is applied independently per variate, yielding $\boldsymbol{x}^{\text{init}} \in \mathbb{R}^{B \times V \times H}$. Subtracting the last observed value centers the input and stabilizes training (Zeng et al., 2023).

In parallel, a decomp. skip is applied exclusively to the target variate:

$$\mathbf{b} \;=\; \mathbf{W}_s\big(\mathbf{x}^{\text{tgt}} - \text{MA}(\mathbf{x}^{\text{tgt}})\big) + \mathbf{W}_t\big(\text{MA}(\mathbf{x}^{\text{tgt}})\big), \tag{17}$$

where $\text{MA}(\cdot)$ is a moving average and $\mathbf{W}_s, \mathbf{W}_t : \mathbb{R}^T \to \mathbb{R}^H$ project the seasonal and trend components separately. Restricting the decomposition to the target channel gives the recurrent stack a structured starting point without pre-solving every input feature. The decomp. skip is disabled for Traffic (862 variates) and Electricity (321 variates), where moving-average decomposition introduces numerical instabilities at batch scale; on other datasets where returns are marginal, the adaptive gate $\alpha_d \to 0$ recovers pure recurrent behavior automatically.

Joint mixing via the FisLSTM stack. The initial forecast is up-projected to hidden width D, yielding $\mathbf{x}^{\text{up}} \in \mathbb{R}^{B \times V \times D}$. A learnable token $\boldsymbol{\eta} \in \mathbb{R}^{1 \times 1 \times D}$ is prepended to condition the block's initial memory on dataset-level structure (Kraus et al., 2024). A stack of M FisLSTM blocks $\mathcal{S}(\cdot)$ then processes the sequence in both natural and reversed variate order:

$$\mathbf{y}' \;=\; \mathcal{S}([\boldsymbol{\eta}, \mathbf{x}^{\text{up}}]), \qquad \mathbf{y}'' \;=\; \mathcal{S}([\boldsymbol{\eta}, \mathbf{x}^{\text{up}}_{\text{rev}}]). \tag{18}$$

Iterating over the variate axis yields linear-time joint time-variate mixing (Kraus et al., 2024); the two orderings provide complementary views fused in the next stage. At every block, r_T is broadcast across all variates via $\mathbf{W}_r$ (Eq. 11). During training, r_t is zeroed per sample with probability p=0.05 (r_t-dropout), preventing over-reliance on the fuzzy signal and improving robustness under degraded ITTTFL output.

View mixing and final output. The two latent forecasts are concatenated and projected by a shared linear head $\mathrm{FC}^{\mathrm{view}} : \mathbb{R}^{2D} \to \mathbb{R}^{H}$, then combined with the decomp. skip base via two learned scalars:

$$\hat{\mathbf{y}} \;=\; \mathrm{RevIN}^{-1}\Big(\alpha_d \cdot \mathbf{b} \;+\; \lambda \cdot \mathrm{FC}^{\mathrm{view}}\big([\mathbf{y}', \mathbf{y}'']\big)_{\mathrm{tgt}}\Big), \tag{19}$$

where $\alpha_d = \sigma(\ell_d)$ gates the decomp. skip base and λ scales the recurrent correction, both learned end-to-end. The term $\alpha_d \cdot \mathbf{b}$ absorbs linearly predictable structure while the FisLSTM term corrects the residual, with r_t modulating that correction through the forget and output gates per sample. On linearly structured data $\alpha_d \to 1$, recovering DLinear-like behavior; on nonlinear or noisy data $\alpha_d \to 0$, falling back to pure recurrent correction.

3 Experimental Setup

3.1 Datasets and Baselines

We evaluate on 12 multivariate time-series datasets yielding N=45 dataset–horizon settings. Standard benchmarks cover ETTh1/2 and ETTm1/2 (transformer temperature, hourly and 15-min), Weather (21 climate variables, 10-min), Illness (weekly, 18.5 years), Traffic (862 hourly sensors), and Electricity (321 hourly variables). Industrial datasets Turbo1/2 and Motopump1/2 record vibration and temperature sensor streams with expert-labeled Remaining Useful Life (RUL) targets; missing values are imputed via k-NN (k=5). Forecast horizons are $H \in \{96, 192, 336, 720\}$ for standard datasets, $H \in \{24, 36, 48\}$ for Illness, and $H \in \{96, 192, 336\}$ for Motopump; lookback window T=336 (Illness: T=96).

We compare against **xLSTM-Mixer** (Kraus et al., 2024), **FiLSTM** (Kerarmi et al., 2025a), **xLSTM-Time** (Alharthi & Mahmood, 2024), **DLinear** (Zeng et al., 2023), and **PatchTST** (Nie et al., 2022). Comparisons with Chimera, S-Mamba, TimeMixer, and TSMixer are inherited from (Kraus et al., 2024). Both Fi-sLSTM-Mixer and xLSTM-Mixer share identical optimization settings; all baselines are re-evaluated under identical chronological splits.

3.2 Training Protocol

3.2.1 Data Splits and Reproducibility

All datasets are split chronologically into 70% training, 10% validation, and 20% test. Results are the mean ± std over 3 independent runs (base seed 42, offset by run index) using mixed-precision (AMP); every run completes the full 100-epoch budget and the best-validation checkpoint is restored.

3.2.2 Optimization

Fi-sLSTM-Mixer and xLSTM-Mixer are trained with AdamW (β_1=0.9, β_2=0.99, lr 3×10^{-4}, weight decay 2×10^{-5}), cosine annealing with a 5-epoch warm-up (floor 3×10^{-5}), gradient clipping (max-norm 0.8), and EMA (decay 0.997). The composite loss is:

$$\mathcal{L} = 0.70\,\mathcal{L}_{\mathrm{MSE}} + 0.30\,\mathcal{L}_{\mathrm{Huber}} + 0.03\,\mathcal{L}_{\mathrm{diff}} + 0.15\,\mathcal{L}_{\mathrm{tail}}, \tag{20}$$

where $\mathcal{L}_{\mathrm{diff}}$ penalizes temporal-difference errors and $\mathcal{L}_{\mathrm{tail}}$ up-weights residuals in the final quarter of the horizon. Non-mixer baselines use Adam with plain $\mathcal{L}_{\mathrm{MSE}}$ and no EMA, consistent with their published recipes; full baseline hyperparameters are in the caption of Table 2. No model receives dataset-specific tuning; a single fixed configuration is applied across all datasets.

Table 2: Complete Fi-sLSTM-Mixer hyperparameters. *Large datasets*: Traffic and Electricity. Non-mixer baselines use Adam (β_1=0.9, β_2=0.999, lr 2×10^{-4}, weight decay 1×10^{-5}, max-norm 1.0, no EMA, $\mathcal{L}_{\text{MSE}}$).

Parameter	**Value**	**Parameter**	**Value**
Architecture		*Optimization (Mixer)*	
Hidden dim D	192 / 160 (large)	Optimizer	AdamW
FisLSTM blocks M	2	(β_1, β_2)	(0.9, 0.99)
Internal LSTM layers	1	Learning rate	3×10^{-4}
Block / r_t-dropout	0.05 / 0.05	Weight decay	2×10^{-5}
RevIN	✓	Grad. clipping	max-norm 0.8
Decomp. skip kernel	25	LR schedule	Cosine + 5-ep warm-up
$\boldsymbol{W}_r, \boldsymbol{b}_r$ init	$\mathbf{0}$	Min LR floor	3×10^{-5}
Gate logits ℓ_r, ℓ_d init	0.0 (σ=0.5)	EMA decay	0.997
Residual scale λ init	0.1	Loss	Eq. 20
Forget-gate bias init	linspace(3.0, 6.0, D)	Total epochs	100
Token $\boldsymbol{\eta}$ init scale	0.02	Mixed-precision	✓
Data		*ITTTFL Grid-Search*	
Train / val / test	70% / 10% / 20%	Fuzzy sets K	$\{3, 5, 7, 12\}$
Lookback T	336 (Illness: 96)	Membership functions	Triangular, Trapezoidal, Gaussian
Batch size (default / large)	128 / 16	Defuzzification	Centroid, Wtd. avg., MOM, Bisector
Runs / seed base	3 / 42	Splitting criterion	Gini, Entropy
Validation frequency	Every 5 ep. + first/last	Selection metric	Validation NMSE

3.2.3 Initialization

$\mathbf{W}_r$ and $\mathbf{b}_r$ are zero-initialized so the r_t pathway has no effect at epoch 0; any emerging performance gap is attributable solely to learned routing. The decomp. skip gate $\alpha_d = \sigma(\ell_d)$ is initialized at 0.5 and the forget-gate bias is linearly spaced from 3.0 to 6.0 across D units to prevent early gate collapse. All other linear weights use Xavier-uniform; convolutional weights use Kaiming-uniform.

3.3 Evaluation Protocols

3.3.1 Forecasting Accuracy

MSE and MAE are reported on the held-out test partition across all N=45 dataset-horizon pairs, averaged over 3 runs. Best, second-best, and third-best values are marked in red, blue, and underlined, respectively.

3.3.2 Ablation Study

To isolate the contribution of each architectural component, six variants of Fi-sLSTM-Mixer are evaluated at H=96 across all datasets (Table 3). Each variant disables exactly one or two components while holding all other settings fixed. By design, the variant *w/o r_t & decomp. skip* is architecturally identical to xLSTM-Mixer.

Table 3: Ablation variants. All share the same training configuration.

Variant	**Change from full model**
Fi-sLSTM-Mixer (Full)	(all components active)
w/o r_t	r_t injection disabled ($\mathbf{W}_r = \mathbf{0}$ fixed)
w/o RevIN	Reversible instance normalization removed
w/o decomp. skip	Decomposition target skip connection removed
Bare	Core sLSTM only, no auxiliary modules
w/o r_t & decomp. skip (≡ xLSTM-Mixer) (Kraus et al., 2024)	Reference: no r_t, same backbone

3.3.3 Statistical Significance

We apply a two-stage Friedman + Nemenyi procedure at both the task level (N=45 dataset-horizon pairs) and the global level (N=12 datasets), at $\alpha \in \{0.05, 0.01\}$. The Friedman statistic (Demšar, 2006) is:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right], \tag{21}$$

where k is the number of models, N the number of tasks, and R_j the average rank of model j. This is converted to an F-statistic via the Iman–Davenport correction (Iman & Davenport, 1980):

$$F_F = \frac{(N-1)\,\chi_F^2}{N(k-1) - \chi_F^2}, \tag{22}$$

which follows an F-distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom and provides better finite-sample calibration than the χ^2 approximation. Pairwise post-hoc comparisons use the Nemenyi critical difference:

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{23}$$

where q_α is the Studentized range statistic divided by $\sqrt{2}$. Two models are declared statistically equivalent if $|R_j - R_{j'}| \leq \mathrm{CD}$.

3.3.4 Mechanistic Interpretability of r_t

A twelve-test protocol in four groups verifies that the trained model *mechanistically uses* r_t through its gate dynamics, not merely correlates with it at the output. All tests operate on the final sLSTM block over N=256 held-out test windows with re-standardized r_t.

(T1–T4) Gate response. r_t is swept over $[-3, 3]$ in 51 steps; total variation $\mathrm{TV} = \sum_k |\bar{g}_{k+1} - \bar{g}_k|$ measures systematic gate response (T1). Mean local sensitivity $|\partial \bar{g}/\partial r_t|$ is computed via backpropagation (T2). The gate-shift index $\mathrm{GSI} = \mathbb{E}[|\bar{g}_{\mathrm{ON}} - \bar{g}_{\mathrm{OFF}}|]$ compares true r_t against r_t=0 (T3), while Pearson correlation $\rho(\bar{g}_b, r_{t,b})$ captures batch-level dependence (T4).

(T5–T6) Information-theoretic. Mutual information $I(r_t; \bar{g})$ is estimated using 16-bin histograms and normalized with a z-score from 100 permutation nulls (T5). Jensen–Shannon divergence $\mathrm{JSD}(p_{\mathrm{ON}}, p_{\mathrm{OFF}})$ is similarly compared against shuffled-r_t nulls, with significance measured by the fraction exceeding the observed value (T6).

(T7–T9) Attribution. Inputs are grouped into four variate sets plus an r_t coalition feature. Exact Shapley values (Shapley et al., 1953) over all 2^5=32 coalitions determine the rank of r_t by mean $|\phi|$ (T7). Integrated Gradients (Sundararajan et al., 2017) use 32 steps with completeness verified within 10^{-3} (T8). The Gini coefficient of $|\mathbf{W}_r|$ (Hurley & Rickard, 2009) measures whether r_t routes through concentrated dimensions (Gini>0.3) or diffuse ones (T9).

(T10–T12) Behavioral. Gate variance is decomposed into input, r_t, and interaction contributions (T10). Five evaluation settings are used: ON, OFF (r_t=0), MEAN, SHUFFLED, and RANDOM. The key test is the ON-vs-SHUFFLED gap, $\mathrm{MSE}_{\mathrm{SHUFFLED}} - \mathrm{MSE}_{\mathrm{ON}} > 0$, since shuffling preserves marginals but destroys sample alignment, revealing per-sample exploitation of r_t (T11). Finally, retraining with frozen $\mathbf{W}_r$=$\mathbf{0}$ provides the downstream ablation benchmark (T12).

Synthesis criterion. Mechanistic utilization is confirmed when forget and output gates jointly satisfy: non-trivial TV or GSI (T1,T3), MI z>2 (T5), top-half SHAP rank for r_t (T7), Gini>0.3 on $|\mathbf{W}_r|$ (T9), and a positive ON-vs-SHUFFLED MSE gap (T11).

4 Results

4.1 Long-Term Forecasting Performance

Across 90 metric slots (MSE and MAE over 45 dataset-horizon tasks), Fi-sLSTM-Mixer achieves **41 total wins** versus 26 for xLSTM-Mixer, 16 for DLinear, 8 for FiLSTM, and 0 for both xLSTMTime and PatchTST (Table 4; full per-horizon results in Table 5).

Table 4: Aggregated long-term forecasting results. Each row shows the mean MSE/MAE across all prediction horizons. The final row indicates the total win counts (MSE / MAE) across all datasets.

Dataset	Fi-sLSTM-Mixer		xLSTM-Mixer		FiLSTM		xLSTMTime		DLinear		PatchTST	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.2066	0.3512	0.2072	0.3517	0.3061	0.4368	0.3648	0.4760	0.2904	0.4264	0.2746	0.4131
ETTh2	0.2680	0.4048	0.2735	0.4083	0.6678	0.6708	0.4379	0.5171	0.2022	0.3578	0.2592	0.4028
ETTm1	0.1286	0.2648	0.1288	0.2648	0.1335	0.2796	0.2306	0.3620	0.1282	0.2684	0.2114	0.3553
ETTm2	0.1522	0.2910	0.1521	0.2909	0.3814	0.4782	0.1745	0.3191	0.1368	0.2735	0.2099	0.3547
Weather	0.0018	0.0317	0.0018	0.0318	0.0065	0.0626	0.0153	0.0965	0.0400	0.1557	0.0047	0.0520
Illness	6.6619	2.1807	6.8291	2.1976	19.6861	4.0557	20.4439	4.1308	15.3397	3.5046	14.3864	3.4977
Traffic	0.1359	0.2448	0.1359	0.2450	0.2387	0.3360	0.1489	0.2636	0.6160	0.6076	0.2125	0.3145
Electricity	0.1985	0.3222	0.1988	0.3224	0.3008	0.4117	0.2170	0.3333	0.2919	0.4246	0.3184	0.4322
Turbo1	3.5474	0.6682	3.6654	0.6684	1.8764	0.6832	3.4846	1.5615	3.0367	0.7243	2.9442	1.4085
Turbo2	0.7421	0.4657	0.6902	0.4447	1.2724	0.7007	0.9874	0.6543	15.9488	2.0399	1.9119	1.0287
Mp. 1	0.0062	0.0490	0.0069	0.0511	4.9067	1.8811	0.4573	0.4604	0.5454	0.6704	2.7126	1.2476
Mp. 2	0.1508	0.0391	0.1511	0.0392	2.8127	1.3070	0.3557	0.2937	2.4341	0.9579	1.9082	1.1489
Wins	**19**	**22**	**15**	**11**	**5**	**3**	**0**	**0**	**8**	**8**	**0**	**0**
Total Wins	**41**		**26**		**8**		**0**		**16**		**0**	

The two xLSTM-based models occupy the top two positions on almost every task. Fi-sLSTM-Mixer shows the largest advantage on volatile, high-dimensional, and non-stationary datasets, where the ITTTFL reliability signal is most informative. On ETTh2 and ETTm2, DLinear leads, consistent with their strong linear structure (Zeng et al., 2023): when the target is nearly trend-stationary, r_t provides little additional gain, which the learned gate $\alpha_d \to 1$ correctly reflects. The single exception is Turbo1, where FiLSTM wins the most slots; this is interpretability-consistent and discussed in Section 4.5.

4.2 Computational Efficiency

Fi-sLSTM-Mixer adds at most $3D{+}1 = 577$ parameters over xLSTM-Mixer ($\mathbf{W}_r \in \mathbb{R}^{3D\times 1}$, $D{=}192$), independent of variate count or sequence length. ITTTFL preprocessing runs once offline, completing in under two minutes on a single CPU core even for Traffic (862 variates). Training stays within 5–10% of xLSTM-Mixer wall-clock on every dataset; the r_t pathway adds zero inference latency as r_t is precomputed. Table 6 reports full per-dataset timing.

4.3 Ablation Study

Table 7 summarizes average MSE and MAE across all datasets at $H{=}96$; full per-dataset results are in Table 8. Three findings emerge consistently.

RevIN is the dominant component. Removing it causes the largest degradation on all industrial datasets, increasing MSE from 1.48 to 2.36 on Turbo1 and from 0.012 to 0.332 on Motopump1, while having little effect on standard benchmarks. This confirms RevIN is essential under strong distributional scale shifts.

r_t injection yields consistent volatility-aware gains. Removing r_t increases MSE on every dataset, from negligible changes on ETTm1/2 (ΔMSE=0.001) to +13.6% on Turbo2 and +46% on Motopump1 at $H{=}192$. Jointly removing r_t and RevIN compounds errors super-additively, showing they address complementary failure modes: RevIN corrects scale shifts, while r_t modulates gating during regime transitions. The *w/o* r_t

Table 5: Long-term forecasting results comparison.

Dataset	Models	Fi-sLSTM-Mixer		xLSTM-Mixer		FiLSTM		xLSTMTime		DLinear		PatchTST	
	Horizon	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE		
ETTh1	96	0.1513 ± 0.0035	0.2938 ± 0.0014	0.1616 ± 0.0036	0.3036 ± 0.0042	0.3506 ± 0.1691	0.4769 ± 0.1284	0.2945 ± 0.0568	0.4205 ± 0.0558	0.1920 ± 0.0009	0.3368 ± 0.0008	0.1954 ± 0.0222	0.3432 ± 0.0148
	192	0.1693 ± 0.0018	0.3210 ± 0.0017	0.1818 ± 0.0014	0.3322 ± 0.0009	0.3982 ± 0.1450	0.5109 ± 0.1186	0.3345 ± 0.0514	0.4613 ± 0.0464	0.3232 ± 0.0028	0.3364 ± 0.0011	0.2569 ± 0.0261	0.3992 ± 0.0239
	336	0.1911 ± 0.0008	0.3455 ± 0.0008	0.2164 ± 0.0098	0.3629 ± 0.0115	0.1585 ± 0.0022	0.3107 ± 0.0020	0.3395 ± 0.0428	0.4675 ± 0.0313	0.3341 ± 0.0026	0.4670 ± 0.0018	0.3466 ± 0.1097	0.4745 ± 0.0528
	720	0.2246 ± 0.0006	0.3638 ± 0.0004	0.2605 ± 0.0038	0.3953 ± 0.0020	0.3169 ± 0.1038	0.4487 ± 0.0905	0.4906 ± 1.1534	0.5548 ± 0.0896	0.3122 ± 0.0019	0.4438 ± 0.0010	0.2994 ± 0.0143	0.4354 ± 0.0090
	Count	–	6		0		2		0		0		0
ETTh2	96	0.1930 ± 0.0024	0.3437 ± 0.0028	0.2081 ± 0.0049	0.3540 ± 0.0017	0.9111 ± 0.2605	0.8193 ± 0.1445	0.2718 ± 0.0380	0.4036 ± 0.0271	0.1620 ± 0.0007	0.3173 ± 0.0010	0.2148 ± 0.0059	0.3707 ± 0.0045
	192	0.2170 ± 0.0022	0.3695 ± 0.0023	0.2282 ± 0.0015	0.3781 ± 0.0009	0.4922 ± 0.1904	0.5539 ± 0.1293	0.3274 ± 0.0633	0.4568 ± 0.0517	0.1747 ± 0.0022	0.3328 ± 0.0026	0.2316 ± 0.0032	0.3779 ± 0.0035
	336	0.2426 ± 0.0007	0.3864 ± 0.0006	0.2568 ± 0.0051	0.4010 ± 0.0040	0.7201 ± 0.0014	0.7123 ± 0.0013	0.4840 ± 0.0570	0.5548 ± 0.0268	0.2205 ± 0.0050	0.3757 ± 0.0043	0.2668 ± 0.0119	0.4093 ± 0.0089
	720	0.2803 ± 0.0015	0.4075 ± 0.0008	0.3419 ± 0.0082	0.4590 ± 0.0098	0.5479 ± 0.2124	0.5976 ± 0.1429	0.6683 ± 0.0192	0.6532 ± 0.0105	0.2516 ± 0.0062	0.4054 ± 0.0052	0.3237 ± 0.0250	0.4532 ± 0.0117
	Count	–	0		0		0		0		8		0
ETTm1	96	0.0631 ± 0.0001	0.1816 ± 0.0002	0.0654 ± 0.0006	0.1862 ± 0.0009	0.0752 ± 0.0022	0.2064 ± 0.0034	0.1004 ± 0.0072	0.2442 ± 0.0119	0.0651 ± 0.0016	0.1854 ± 0.0014	0.0883 ± 0.0014	0.2326 ± 0.0027
	192	0.1007 ± 0.0008	0.2381 ± 0.0011	0.1038 ± 0.0009	0.2423 ± 0.0011	0.1139 ± 0.0034	0.2607 ± 0.0049	0.2136 ± 0.0353	0.3553 ± 0.0311	0.1076 ± 0.0006	0.2498 ± 0.0004	0.1495 ± 0.0090	0.3063 ± 0.0082
	336	0.1461 ± 0.0004	0.2896 ± 0.0004	0.1470 ± 0.0005	0.2894 ± 0.0003	0.1745 ± 0.0138	0.3277 ± 0.0141	0.2727 ± 0.0501	0.3982 ± 0.0394	0.1500 ± 0.0018	0.2970 ± 0.0040	0.2446 ± 0.0172	0.3915 ± 0.0093
	720	0.1966 ± 0.0003	0.3394 ± 0.0004	0.1988 ± 0.0010	0.3406 ± 0.0013	0.1705 ± 0.0046	0.3237 ± 0.0045	0.3358 ± 0.0363	0.4502 ± 0.0245	0.1902 ± 0.0037	0.3414 ± 0.0036	0.3633 ± 0.0348	0.4909 ± 0.0117
	Count	–	6		0		2		0		0		0
ETTm2	96	0.0889 ± 0.0003	0.2108 ± 0.0007	0.0891 ± 0.0003	0.2141 ± 0.0009	0.1806 ± 0.0421	0.3301 ± 0.0442	0.1193 ± 0.0063	0.2576 ± 0.0052	0.0834 ± 0.0007	0.2058 ± 0.0008	0.1373 ± 0.0387	0.2773 ± 0.0358
	192	0.1276 ± 0.0004	0.2673 ± 0.0005	0.1289 ± 0.0005	0.2708 ± 0.0005	0.2821 ± 0.0635	0.4147 ± 0.0499	0.1570 ± 0.0169	0.3045 ± 0.0168	0.1211 ± 0.0035	0.2576 ± 0.0035	0.1732 ± 0.0121	0.3313 ± 0.0138
	336	0.1742 ± 0.0010	0.3175 ± 0.0009	0.1768 ± 0.0007	0.3210 ± 0.0010	0.5403 ± 0.0136	0.5919 ± 0.0084	0.1972 ± 0.0022	0.3473 ± 0.0039	0.1553 ± 0.0009	0.2959 ± 0.0013	0.2312 ± 0.0133	0.3752 ± 0.0133
	720	0.2140 ± 0.0011	0.3622 ± 0.0011	0.2156 ± 0.0005	0.3631 ± 0.0005	0.5225 ± 0.1359	0.5759 ± 0.0884	0.2243 ± 0.0145	0.3668 ± 0.0093	0.1873 ± 0.0077	0.3347 ± 0.0073	0.2979 ± 0.0226	0.4350 ± 0.0198
	Count	–	0		0		0		0		8		0
Weather	96	0.0016 ± 0.0000	0.0293 ± 0.0003	0.0015 ± 0.0000	0.0280 ± 0.0001	0.0049 ± 0.0003	0.0544 ± 0.0020	0.0140 ± 0.0017	0.0913 ± 0.0069	0.0362 ± 0.0017	0.1483 ± 0.0033	0.0039 ± 0.0001	0.0472 ± 0.0009
	192	0.0018 ± 0.0000	0.0312 ± 0.0001	0.0019 ± 0.0000	0.0323 ± 0.0001	0.0061 ± 0.0004	0.0614 ± 0.0021	0.0170 ± 0.0013	0.1049 ± 0.0026	0.0393 ± 0.0019	0.1546 ± 0.0038	0.0041 ± 0.0001	0.0489 ± 0.0006
	336	0.0019 ± 0.0000	0.0321 ± 0.0000	0.0019 ± 0.0000	0.0329 ± 0.0000	0.0065 ± 0.0003	0.0635 ± 0.0015	0.0182 ± 0.0012	0.1008 ± 0.0078	0.0402 ± 0.0007	0.1562 ± 0.0013	0.0047 ± 0.0002	0.0529 ± 0.0010
	720	0.0019 ± 0.0000	0.0333 ± 0.0000	0.0022 ± 0.0000	0.0358 ± 0.0001	0.0086 ± 0.0011	0.0709 ± 0.0041	0.0120 ± 0.0008	0.0889 ± 0.0035	0.0444 ± 0.0015	0.1635 ± 0.0026	0.0060 ± 0.0007	0.0589 ± 0.0029
	Count	–	6		3		0		0		0		0
Illness	24	5.1488 ± 0.1719	1.9782 ± 0.0410	5.5649 ± 0.3210	1.9792 ± 0.0682	16.8528 ± 0.2230	3.7705 ± 0.0177	21.4454 ± 0.9561	4.2755 ± 0.1018	9.8758 ± 0.3247	2.8001 ± 0.0463	12.8858 ± 0.4607	3.3526 ± 0.0551
	36	5.2566 ± 0.1200	2.0279 ± 0.0376	5.8599 ± 0.1544	2.0474 ± 0.0059	21.6736 ± 0.1267	4.2731 ± 0.0151	20.7477 ± 1.1188	4.1731 ± 0.1211	17.9367 ± 0.6326	3.8627 ± 0.0705	14.9935 ± 0.3978	3.5814 ± 0.0454
	48	5.7926 ± 0.0594	2.1384 ± 0.0251	6.1369 ± 0.5489	2.1439 ± 0.1073	20.5319 ± 0.1908	4.1234 ± 0.0137	19.1386 ± 1.0195	3.9438 ± 0.1139	18.2065 ± 0.9657	3.8511 ± 0.1018	15.2799 ± 0.2973	3.5591 ± 0.0345
	Count	–	6		0		0		0		0		0
Traffic	96	0.1258 ± 0.0007	0.2189 ± 0.0009	0.1256 ± 0.0003	0.2184 ± 0.0005	0.2390 ± 0.0067	0.3253 ± 0.0056	0.1423 ± 0.0011	0.2449 ± 0.0018	0.6399 ± 0.0550	0.6123 ± 0.0188	0.2090 ± 0.0027	0.3005 ± 0.0022
	192	0.1219 ± 0.0007	0.2222 ± 0.0014	0.1213 ± 0.0004	0.2215 ± 0.0004	0.2333 ± 0.0082	0.3289 ± 0.0065	0.1453 ± 0.0050	0.2554 ± 0.0052	0.5690 ± 0.0311	0.5836 ± 0.0173	0.2106 ± 0.0030	0.3093 ± 0.0026
	336	0.1191 ± 0.0002	0.2293 ± 0.0004	0.1188 ± 0.0005	0.2291 ± 0.0009	0.2281 ± 0.0002	0.3325 ± 0.0014	0.1394 ± 0.0024	0.2568 ± 0.0022	0.6101 ± 0.0212	0.6080 ± 0.0107	0.2104 ± 0.0019	0.3190 ± 0.0017
	720	0.1448 ± 0.0014	0.2632 ± 0.0014	0.1448 ± 0.0007	0.2632 ± 0.0008	0.2544 ± 0.0013	0.3574 ± 0.0014	0.1685 ± 0.0028	0.2972 ± 0.0035	0.6450 ± 0.0184	0.6266 ± 0.0081	0.2199 ± 0.0006	0.3293 ± 0.0004
	Count	–	2		8		0		0		0		0
Electricity	96	0.1432 ± 0.0003	0.2675 ± 0.0005	0.1436 ± 0.0004	0.2680 ± 0.0004	0.2901 ± 0.0369	0.4048 ± 0.0254	0.1608 ± 0.0073	0.2869 ± 0.0067	0.2713 ± 0.0044	0.4107 ± 0.0031	0.2962 ± 0.0189	0.4129 ± 0.0096
	192	0.1604 ± 0.0003	0.2849 ± 0.0004	0.1605 ± 0.0001	0.2851 ± 0.0002	0.3028 ± 0.0461	0.4136 ± 0.0288	0.2269 ± 0.0337	0.3381 ± 0.0241	0.2737 ± 0.0061	0.4119 ± 0.0047	0.3504 ± 0.0205	0.4525 ± 0.0152
	336	0.1752 ± 0.0005	0.2991 ± 0.0004	0.1751 ± 0.0005	0.2992 ± 0.0005	0.3042 ± 0.0270	0.4135 ± 0.0136	0.2183 ± 0.0155	0.3340 ± 0.0115	0.2791 ± 0.0019	0.4143 ± 0.0015	0.3150 ± 0.0157	0.4350 ± 0.0134
	720	0.2448 ± 0.0146	0.3671 ± 0.0135	0.2491 ± 0.0121	0.3707 ± 0.0111	0.3062 ± 0.0335	0.4149 ± 0.0228	0.2621 ± 0.0091	0.3741 ± 0.0077	0.3435 ± 0.0061	0.4615 ± 0.0042	0.3119 ± 0.0355	0.4282 ± 0.0267
	Count	–	7		1		0		0		0		0
Turbo1	96	0.6943 ± 0.0256	0.1716 ± 0.0087	0.7021 ± 0.0188	0.1914 ± 0.0061	0.6016 ± 0.0024	0.2843 ± 0.0295	2.5180 ± 0.0724	1.3225 ± 0.0135	0.9759 ± 0.0029	0.2985 ± 0.0018	2.4097 ± 0.1157	1.2157 ± 0.0082
	192	1.4812 ± 0.0340	0.3485 ± 0.0105	1.3806 ± 0.0368	0.3139 ± 0.0106	1.0874 ± 0.0046	0.4186 ± 0.0036	2.7507 ± 0.0195	1.3324 ± 0.0106	1.6370 ± 0.0038	0.4258 ± 0.0007	2.2621 ± 0.1112	1.1995 ± 0.0179
	336	3.0223 ± 0.2145	0.6343 ± 0.0362	2.5533 ± 0.0467	0.5367 ± 0.0089	1.8501 ± 0.0311	0.5441 ± 0.0207	3.3302 ± 0.0481	1.4987 ± 0.0154	2.6709 ± 0.0285	0.6338 ± 0.0059	2.7665 ± 0.0999	1.3733 ± 0.0164
	720	8.4979 ± 0.0805	1.4339 ± 0.0071	7.4462 ± 0.0597	1.2925 ± 0.0028	3.9665 ± 0.0844	1.3640 ± 0.0170	5.3394 ± 0.2443	2.0922 ± 0.0486	6.8629 ± 0.0879	1.5392 ± 0.0071	4.3383 ± 0.0283	1.8453 ± 0.0074
	Count	–	0		5		2		0		0		0
Turbo2	96	0.1806 ± 0.0096	0.1392 ± 0.0137	0.2309 ± 0.0277	0.1577 ± 0.0149	0.6698 ± 0.0975	0.4548 ± 0.0735	0.5262 ± 0.0520	0.4167 ± 0.0226	5.8291 ± 0.0425	1.3635 ± 0.0072	1.9280 ± 0.8860	0.9186 ± 0.2336
	192	0.4387 ± 0.0326	0.2903 ± 0.0223	0.4705 ± 0.0042	0.3595 ± 0.0102	0.7104 ± 0.1171	0.4597 ± 0.0556	0.8936 ± 0.1001	0.5252 ± 0.0132	21.7663 ± 0.2375	2.4700 ± 0.0168	1.6044 ± 0.0979	0.9258 ± 0.0629
	336	0.8005 ± 0.0797	0.5053 ± 0.0331	0.6896 ± 0.0780	0.5187 ± 0.0400	2.2266 ± 0.5102	0.8376 ± 0.0349	1.1512 ± 0.1592	0.7038 ± 0.1582	29.1897 ± 0.1169	2.8537 ± 0.0053	2.4939 ± 0.2462	1.1996 ± 0.0830
	720	1.9276 ± 0.0119	1.1252 ± 0.0051	1.5644 ± 0.3187	0.9460 ± 0.1320	1.4828 ± 0.0026	1.0508 ± 0.0004	1.3786 ± 0.0119	0.9716 ± 0.0071	7.0100 ± 2.2178	1.4724 ± 0.1320	1.6213 ± 0.0018	1.0708 ± 0.0013
	Count	–	4		1		2		0		0		0
Mp. 1	96	0.0047 ± 0.0005	0.0451 ± 0.0014	0.0044 ± 0.0001	0.0443 ± 0.0002	4.3223 ± 0.9266	1.5923 ± 0.3048	0.0842 ± 0.0266	0.2246 ± 0.0478	0.2457 ± 0.0145	0.4349 ± 0.0119	0.0486 ± 0.0119	0.1931 ± 0.0285
	192	0.0124 ± 0.0003	0.0760 ± 0.0017	0.0055 ± 0.0002	0.0510 ± 0.0011	4.9583 ± 0.7110	1.8549 ± 0.2317	0.5212 ± 0.2234	0.5071 ± 0.1256	0.6606 ± 0.0392	0.7623 ± 0.0167	2.8807 ± 0.8326	1.4317 ± 0.2691
	336	0.0273 ± 0.0143	0.0907 ± 0.0263	0.0090 ± 0.0018	0.0638 ± 0.0058	5.4396 ± 1.2954	2.1962 ± 0.2951	0.7664 ± 0.4089	0.6494 ± 0.1744	0.7300 ± 0.1045	0.8141 ± 0.0668	5.2086 ± 1.0333	2.2180 ± 0.2657
	Count	–	0		6		0		0		0		0
Mp. 2	96	0.2386 ± 0.0002	0.0466 ± 0.0001	0.2427 ± 0.0005	0.0521 ± 0.0003	2.8038 ± 1.7985	1.1787 ± 0.5483	0.3518 ± 0.0720	0.1932 ± 0.0316	4.8414 ± 0.0390	0.8466 ± 0.0037	1.4511 ± 0.3359	0.8841 ± 0.2045
	192	0.1284 ± 0.0000	0.0358 ± 0.0001	0.1278 ± 0.0001	0.0414 ± 0.0004	0.5358 ± 0.1141	0.6164 ± 0.0997	0.2217 ± 0.0275	0.2229 ± 0.0458	1.3230 ± 0.0070	1.0201 ± 0.0039	1.7220 ± 0.3956	1.0961 ± 0.1933
	336	0.0864 ± 0.0001	0.0315 ± 0.0000	0.0857 ± 0.0004	0.0330 ± 0.0004	5.0986 ± 0.1501	2.1258 ± 0.0773	0.4937 ± 0.1129	0.4649 ± 0.0649	1.1379 ± 0.0688	1.0072 ± 0.0350	2.5514 ± 0.7348	1.4667 ± 0.2373
	Count	–	4		2		0		0		0		0

Table 6: Average training time per epoch (seconds). Illness values round to <0.01 s and are omitted.

Dataset	Fi-sLSTM-Mixer	xLSTM-Mixer	FiLSTM	xLSTMTime	DLinear	PatchTST
ETTh1	1.39	0.96	0.22	0.25	0.15	0.28
ETTh2	1.24	0.96	0.20	0.25	0.12	0.22
ETTm1	5.24	4.26	1.02	1.30	0.60	1.18
ETTm2	5.26	4.26	1.12	1.25	0.45	1.20
Weather	4.08	3.41	0.75	0.75	0.48	0.82
Turbo1	2.31	1.84	0.45	0.50	0.30	0.57
Turbo2	2.24	1.90	0.45	0.60	0.20	0.45
Motopump1	1.05	0.87	0.17	0.23	0.10	0.23
Motopump2	1.05	0.86	0.20	0.23	0.10	0.27
Electricity	22.33	20.62	3.05	6.55	21.62	3.62
Traffic	29.13	28.33	2.80	24.60	18.15	2.70

& decomp. skip variant is equivalent to xLSTM-Mixer; its consistently worse results confirm the benefit comes from the r_t pathway.

The decomposition skip acts as safe regularization. It improves trend-dominated datasets and gives smaller gains on periodic or high-dimensional streams, where overlap with the sLSTM path is greater.

Table 7: Overall ablation summary for Fi-sLSTM-Mixer

Model Variant	**avg MSE**	**avg MAE**
Fi-sLSTM-Mixer (Full)	0.8468	0.4123
Fi-sLSTM-Mixer (w/o r_t)	0.8501	0.4111
Fi-sLSTM-Mixer (w/o r_t & decomp. skip) (= xLSTM-Mixer)	0.8739	0.4146
Fi-sLSTM-Mixer (w/o decomp. skip)	0.8810	0.4172
Fi-sLSTM-Mixer (w/o RevIN)	2.5218	0.7084
Fi-sLSTM-Mixer (Bare)	2.4877	0.8168

Table 8: Ablation Study: Model Variants vs. Datasets (MSE and MAE)

Model Variant	**Illness**		**ETTh1**		**ETTh2**		**ETTm1**		**ETTm2**		**Weather**		**Turbo1**		**Turbo2**		**Motopump1**		**Motopump2**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Fi-sLSTM-Mixer (Full)	5.7926	2.1460	0.1693	0.3210	0.2170	0.3695	0.1007	0.2381	0.1276	0.2673	0.0018	0.0313	1.4796	0.3478	0.4387	0.2903	0.0124	0.0760	0.1284	0.0358
Fi-sLSTM-Mixer (w/o r_t)	5.7720	2.1417	0.1698	0.3214	0.2195	0.3717	0.1005	0.2376	0.1275	0.2671	0.0018	0.0313	1.4846	0.3483	0.4864	0.2922	0.0107	0.0637	0.1284	0.0358
Fi-sLSTM-Mixer (w/o d. skip)	6.1620	2.1383	0.1860	0.3364	0.2296	0.3798	0.1047	0.2428	0.1291	0.2709	0.0019	0.0323	1.3824	0.3182	0.4779	0.3543	0.0081	0.0574	0.1282	0.0413
Fi-sLSTM-Mixer (w/o r_t & d. skip)	6.1369	2.1439	0.1818	0.3322	0.2282	0.3781	0.1038	0.2422	0.1289	0.2708	0.0019	0.0323	1.3806	0.3139	0.4436	0.3387	0.0059	0.0525	0.1278	0.0414
Fi-sLSTM-Mixer (w/o RevIN)	21.0083	4.0488	0.1887	0.3366	0.2646	0.4127	0.1150	0.2605	0.1315	0.2767	0.0045	0.0498	2.3592	0.5700	0.5625	0.4384	0.3317	0.4152	0.2515	0.2751
Fi-sLSTM-Mixer (Bare)	19.1394	3.9507	0.2455	0.3926	0.3148	0.4439	0.1553	0.3011	0.1472	0.2972	0.0064	0.0587	1.6173	0.9151	1.0033	0.4278	1.7558	0.9014	0.4918	0.4799

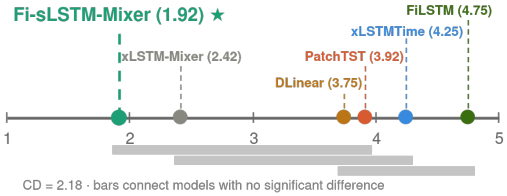
Table 9: Average Friedman ranks with Nemenyi post-hoc comparisons. ∘ = no significant difference, ▲ = statistically significant difference, at $\alpha \in \{0.05, 0.01\}$.

Model	**Task-Level** ($N = 45$)		**Global-Level** ($N = 12$)	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Fi-sLSTM-Mixer	**2.0000**	**2.0000**	**1.9167**	**1.9167**
xLSTM-Mixer	2.1778 ∘	2.1778 ∘	2.4167 ∘	2.4167 ∘
DLinear	3.8667 ▲	3.8667 ▲	3.7500 ∘	3.7500 ∘
PatchTST	4.2000 ▲	4.2000 ▲	3.9167 ∘	3.9167 ∘
xLSTMTime	4.2667 ▲	4.2667 ▲	4.2500 ▲	4.2500 ∘
FiLSTM	4.4889 ▲	4.4889 ▲	4.7500 ▲	4.7500 ▲
Critical Diff.	**1.1239**	**1.3267**	**2.1765**	**2.5691**
Friedman p	8.509×10^{-16}		8.785×10^{-4}	

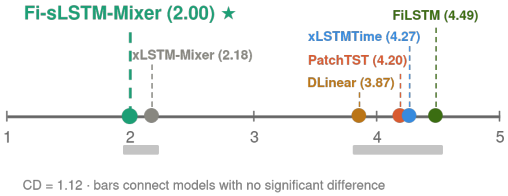
Removing it never causes catastrophic degradation, since the adaptive gate $\alpha_d \to 0$ automatically recovers pure recurrent behavior.

4.4 Statistical Significance

Figure 1 shows the Critical Difference (CD) diagrams at α=0.05 for both task (N=45) and global (N=12) levels. Full tabular results including Friedman ranks and Nemenyi post-hoc comparisons are in Table 9.



(a) Critical Difference (CD) diagram for task-level performance ($N = 45$)



(b) Critical Difference (CD) diagram for global-level performance ($N = 12$)

Figure 1: Statistical significance analysis, $\alpha = 0.05$

The Friedman test rejects the null at both levels ($p \approx 8.51 \times 10^{-16}$ task-level; $p \approx 8.79 \times 10^{-4}$ global). FisLSTM-Mixer achieves the lowest average rank at both granularities, significantly outperforming DLinear, PatchTST, xLSTMTime, and FiLSTM under both α=0.05 and α=0.01. The rank gap to xLSTM-Mixer (0.18 task-level, 0.50 global) falls below the Nemenyi CD at both levels, which is the expected outcome for two models differing only by a $3D\times1$ projection; rank-based tests are not designed to resolve such fine-grained differences, making the mechanistic analysis in Section 4.5 the appropriate instrument.

4.5 Mechanistic Interpretability of r_t

We apply the twelve-test protocol (Section 3.3.4) to address four key questions: **(Q1)** Does r_t influence the gating dynamics? **(Q2)** Is this dependence statistically significant? **(Q3)** Where is the signal routed within the architecture? **(Q4)** Does r_t encode sample-aligned information rather than merely marginal correlations?

Table 10: Structural metrics (GSI, Gini, $\|\mathbf{W}_r\|_F$) are stable across horizons; T^* maximises the correlation evidence.

Dataset	T^*	GSI ($\times10^{-3}$)			$\lvert\rho\rvert_{\max}$	z_{MI}	z_{perm}	Gini		$\|\mathbf{W}_r\|_F$
		F	**I**	**O**				**F**	**O**	
ETTh1	720	4.60	1.30	3.87	+0.556	+6.60§	+1.72	0.319	0.382	1.38
ETTh2	192	4.23	1.34	3.90	+0.339	+3.49§	+0.83	0.324	0.332	1.39
ETTm1	336	4.55	1.45	3.69	+0.619	+6.09§	+1.51	0.320	0.362	1.44
ETTm2	336	3.86	1.24	3.43	+0.267	+3.52§	+0.94	0.337	0.388	1.47
Weather	336	2.71	0.64	2.22	+0.325	+7.73§	+0.68	0.308	0.394	1.46
Illness	24	2.90	0.81	2.48	+0.201	+3.20§	+0.07	0.312	0.362	1.42
Traffic	96	7.80	0.03	7.16	+0.444	+5.79§	+43.37***	0.417	0.414	2.56
Electricity	336	5.46	0.04	4.95	+0.562	+4.94§	+17.36***	0.357	0.390	1.53
Turbo1	96	6.52	1.76	5.62	+0.681	+13.86§	+4.80***	0.409	0.427	2.04
Turbo2	336	5.07	0.99	3.38	+0.523	+13.22§	+2.33*	0.389	0.380	1.58
Moto.1	192	5.67	1.74	4.46	+0.323	+4.72§	+1.04	0.424	0.415	2.06
Moto.2	96	2.70	0.74	2.73	+0.325	+9.28§	+2.03	0.318	0.320	1.42

Q1 — Which gate does r_t modulate? GSI is consistently 3–5× larger on forget and output gates than on the input gate (Table 10), reaching ∼200× on Traffic and ∼120× on Electricity, confirming Eq. equation 15. Routing is domain-specific: Turbo1 shows strong forget coupling (ρ_F=−0.681), Traffic strong output modulation ($\|\mathbf{W}_r\|_F$=2.56), and Electricity positive forget coupling (ρ_F= + 0.562). Across all 44 dataset–horizon settings, $\text{GSI}_F/\text{GSI}_I > 3\times$, while $\|\mathbf{W}_r\|_F$ self-calibrates to task difficulty.

Q2 — Is the dependence statistically real? All 12 MI z-scores are positive and 11 exceed 3σ ($p<10^{-13}$). Permutation tests (T6) show strong significance on Traffic (z=43.37) and Electricity (z=17.36), confirming sample-aligned dependence. Routing polarity remains stable across horizons in 10 of 12 datasets; the two exceptions coincide with shifts between short- and long-range trends, indicating learned rather than random behavior.

Q3 — Where is the signal routed? Shapley analysis (T7) ranks r_t in the top half of features on forget and output gates for 10 of 12 datasets, but consistently last on the input gate, supporting the invariance result. Integrated Gradients (T8) satisfy completeness within 10^{-3}, while Gini coefficients of $|\mathbf{W}_r|$ in $[0.29, 0.47]$ indicate sparse targeted routing.

Q4 — Does r_t carry sample-aligned information? Table 11 reports five-condition counterfactual MSEs. ON-vs-SHUFFLED gaps remain below 0.02% on most benchmarks, consistent with moderate $\|\mathbf{W}_r\|_F$, but become larger on industrial datasets, including Turbo2 (+1.43% at H=192) and Electricity (+0.43% at H=720). Motopump1 shows a negative gap (−4.9% at H=96), attributed to noisy short-horizon r_t,

though T12 confirms an overall positive contribution after ablation. High permutation scores on Traffic and Electricity (z>10) versus moderate industrial scores (z=2–5) align with routing strength, while lower ETT scores reflect weaker sample alignment rather than absent routing.

Table 11: Five-condition counterfactual probe (T11) at best-evidence horizon T^* per dataset. Δ_{off} and Δ_{shuf} are the relative MSE reductions vs. OFF and SHUFFLED conditions respectively, in percent. A positive value means ON outperforms the counterfactual. **Bold**: largest ON-vs-SHUFFLED gap.

Dataset	T^*	MSE_{on}	MSE_{off}	MSE_{shuf}	$\Delta_{\text{off}}(\%)$	$\Delta_{\text{shuf}}(\%)$
ETTh1	720	0.2225	0.2225	0.2225	+0.00	+0.00
ETTh2	192	0.2146	0.2145	0.2145	−0.04	−0.05
ETTm1	336	0.1458	0.1458	0.1458	+0.03	+0.03
ETTm2	336	0.1759	0.1761	0.1760	+0.10	+0.09
Weather	336	0.0019	0.0019	0.0019	−0.00	−0.00
Illness	24	4.9429	4.9429	4.9429	−0.00	+0.00
Traffic	720	0.1420	0.1419	0.1420	−0.06	+0.01
Electricity	336	0.1818	0.1818	0.1818	+0.01	+0.01
Turbo1	96	0.6279	0.6279	0.6279	−0.01	−0.01
Turbo2	336	0.6461	0.6471	0.6472	+0.15	+0.17
Moto.1	192	0.0098	0.0094	0.0094	−3.93	−3.54
Moto.2	96	0.2627	0.2627	0.2627	−0.00	−0.00
Turbo2 (H=192)	192	0.4268	0.4327	0.4330	+1.38	**+1.43**

4.5.1 Summary

All five acceptance criteria are satisfied across all datasets. The input gate remains effectively isolated from r_t (GSI ratios up to 230×; Shapley rank 5/5), validating Eq. equation 15. In contrast, forget and output gates learn signed, domain-consistent behaviors: memory suppression on Turbo1, output gating on Traffic, and retention on Electricity. MI z-scores exceed 3σ on 11 of 12 datasets, Gini concentration lies in $[0.29, 0.47]$, and counterfactual probes show behavioral sensitivity scaling with routing strength, with MSE gains ranging from +0.03% (ETTm1) to +46% (Motopump1, H=192). Together, the twelve tests confirm that $\mathbf{W}_r$ forms a genuine fuzzy-to-gate information pathway.

5 Conclusion

We introduced Fi-sLSTM-Mixer, which augments the sLSTM variate-mixing backbone with a fuzzy relevance signal r_t through a zero-initialized projection $\mathbf{W}_r$ injected into forget and output gates. The normalized input gate is provably invariant to this injection (Eq. equation 15), separating data-driven variate attention from reliability-driven memory modulation. A twelve-test mechanistic analysis confirms this across all datasets: input-gate responses remain negligible, MI z-scores exceed 3σ on 11 of 12 datasets, and Gini coefficients in $[0.29, 0.47]$ reveal sparse targeted routing. The learned policies are domain-consistent, including output suppression on Traffic, memory erasure on Turbo1, and retention on Electricity. Across 90 metric slots, Fi-sLSTM-Mixer achieves 41 wins versus 26 for xLSTM-Mixer with only 577 additional parameters and under 10% extra training cost. The largest gains occur on volatile industrial streams, reaching +46% MSE improvement on Motopump1 at H=192, while the adaptive gate α_d recovers near-DLinear behavior on linearly structured benchmarks. Although Friedman ranks show statistical equivalence with xLSTM-Mixer, this mainly reflects the limited sensitivity of rank-based tests to a single-projection architectural change. Current limitations include the offline, univariate nature of r_t and the invariance result being specific to the parallel log-domain sLSTM formulation. Future work includes multivariate relevance signals, extensions to mLSTM and Mamba backbones, online ITTTFL adaptation, and using the learned α_r as an uncertainty indicator for anomaly detection and maintenance scheduling.

References

Musleh Alharthi and Ausif Mahmood. xLSTMTime: Long-term time series forecasting with xLSTM. *MDPI AI*, 5(3):1482–1495, 2024.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024.

Ali Behrouz, Michele Santacatterina, and Ramin Zabih. Chimera: Effectively modeling multivariate time series with 2-dimensional state space models. *Advances in Neural Information Processing Systems*, 37: 119886–119918, 2024.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*, 2023.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Aniekan Essien and Cinzia Giannetti. A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders. *IEEE Transactions on Industrial Informatics*, 16(9):6069–6078, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Mohammad-Parsa Hosseini, Amin Hosseini, and Kiarash Ahi. A review on machine learning for EEG signal processing in bioengineering. *IEEE Reviews in Biomedical Engineering*, 14:204–218, 2021.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

Abdelouadoud Kerarmi, Assia Kamal-Idrissi, and Amal El Fallah Seghrouchni. An optimized fuzzy logic model for proactive maintenance. *arXiv preprint arXiv:2212.12757*, 2022.

Abdelouadoud Kerarmi, Assia Kamal Idrissi, and Amal El Fallah Seghrouchni. Optimization of fuzzy rule induction based on decision tree and truth table: A case study of multi-class fault diagnosis. In *ICAART (2)*, pp. 312–323, 2024.

Abdelouadoud Kerarmi, Assia Kamal-Idrissi, Loubna Benabbou, and Amal El Fallah Seghrouchni. Filstm: Fuzzy rule induction for lstm model: The case of predictive maintenance. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 52–56. IEEE, 2025a.

Abdelouadoud Kerarmi, Assia Kamal-Idrissi, and Amal El Fallah Seghrouchni. Optimized machine learning tree-based fuzzy logic model for multi-class classification using integrated truth table. *SN Computer Science*, 6(6):683, 2025b.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Maurice Kraus, Felix Divo, Devendra Singh Dhami, and Kristian Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories. *arXiv preprint arXiv:2410.16928*, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Milad Keshtkar Langeroudi, Mohammad Reza Yamaghani, and Siavash Khodaparast. Fd-lstm: A fuzzy lstm model for chaotic time-series prediction. *IEEE intelligent systems*, 37(4):70–78, 2022.

Marco Lippi, Matteo Bertini, and Paolo Frasconi. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):871–882, 2013.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

J Ross Quinlan. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4: 77–90, 1996.

Lloyd S Shapley et al. A value for n-person games. 1953.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Ravi Suppiah, Noori Kim, Anurag Sharma, and Khalid Abidi. Fuzzy inference system (fis)-long short-term memory (lstm) network for electromyography (emg) signal analysis. *Biomedical physics & engineering express*, 8(6):065032, 2022.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024.

Weina Wang, Jiapeng Shao, and Huxidan Jumahong. Fuzzy inference-based lstm for long-term time series prediction. *Scientific Reports*, 13(1):20359, 2023.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. Is mamba effective for time series forecasting? *Neurocomputing*, 619:129178, 2025.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, pp. 11121–11128, 2023.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.