Review

Mamba for Time Series Analysis: A Contemporary Survey

Thanh Tam Nguyen *, Ming Jin, Trinh Pham, Shirui Pan, Quoc Viet Hung Nguyen

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0995.v1

Keywords: time series analysis; selective state-space models; Mamba; deep learning; sequence modeling

 doi:10.20944/preprints202605.0995.v1

Review

Mamba for Time Series Analysis: A Contemporary Survey

Thanh Tam Nguyen *, Ming Jin, Trinh Pham, Shirui Pan and Quoc Viet Hung Nguyen

Griffith University, 170 Kessels Rd, Nathan QLD 4111, Australia
* Correspondence: t.nguyen19@griffith.edu.au

Abstract

Time series analysis (TSA) – forecasting, anomaly detection, imputation, classification, and unified multi-task analytics – has become a battleground for sequence-modeling backbones, where Transformers, MLPs, convolutions, and state-space models compete on long-context benchmarks. Since the 2023 release of MAMBA, its linear-time recurrence and input-dependent selectivity have triggered a surge of MAMBA-based time-series models that report strong numbers yet resist like-for-like comparison. This survey provides the first focused treatment of MAMBA for time series analysis, organized around four orthogonal perspectives: Model, Task, Data, and Application. The Model perspective formalizes a five-axis design space – tokenization, channel strategy, directional scan, hybridization, and decomposition – together with a three-pattern architectural toolbox of pure, bidirectional, and hybrid designs. The Task perspective re-indexes the corpus across the five canonical TSA tasks. The Data perspective re-tags it by data shape, spanning univariate, multivariate, spatio-temporal graph, and irregular series. The Application perspective then surveys deployments in healthcare, energy, traffic, climate, finance, activity recognition, and foundation-scale settings. Building on these views, we distill practical guidelines for variant selection, training recipes, and MAMBA-specific pitfalls. We also catalog public implementations, datasets, and metrics, and we map out open frontiers in gain attribution, modeling regimes, and cross-task unification. An online repository is maintained at https://github.com/tamlhp/awesome-mamba-ts.

Keywords: time series analysis; selective state-space models; MAMBA ; deep learning; sequence modeling

1. Introduction

Time series analysis spans forecasting, anomaly detection, imputation, classification, and increasingly unified multi-task analytics. These five tasks underpin the operational core of electricity markets [97], traffic control systems [15], clinical early-warning streams [9], and industrial monitoring at scale. Across all of them, the sequence backbone is asked to do the same job: capture long-range dependencies, model multi-scale seasonalities, react to regime changes, and exploit cross-channel interactions.

Three architectural families dominate the post-Transformer era, each with shortcomings. *Transformers* (Informer, Autoformer, PatchTST, iTransformer [58,66,105,128]) pay $\mathcal{O}(L^2)$ on long lookbacks ($L \geq 512$) that improve accuracy. *MLP and linear* models such as DLinear [120] and N-BEATS [68] match those baselines with closed-form decompositions and shallow projections, but lack a real temporal-mixing primitive of their own. *Convolutional and 2D* models like TimesNet [104] and TimeMixer++ [99] reshape the time axis into period-aligned or multi-scale 2D tensors to expose multi-period structure and have proven especially strong across tasks under one backbone, at the cost of a hard locality bias. The picture is fragmented: no backbone uniformly wins, and reported gains often hinge on small implementation details.

 doi:10.20944/preprints202605.0995.v1

MAMBA [30], a selective SSM with input-dependent transitions and a hardware-aware parallel scan, reopened the question. It inherits the $\mathcal{O}(L)$ inference cost and the unbounded receptive field of structured SSMs (S4 [32], S5 [90]), but adds the data-dependent gating that prior linear-time recurrences lacked, matching or exceeding same-size Transformers up to 3B parameters on language modeling. For time series the appeal is sharper still. Long lookback windows that strain self-attention become tractable, the recurrent form maps cleanly onto streaming-inference deployment, and the same selectivity machinery can serve both predictive and representational objectives – a single backbone for many tasks.

Since the release of MAMBA, peer-reviewed MAMBA-based time-series models have spread across all five canonical TSA tasks. Forecasting attracted the largest wave, organized along three architectural patterns (Figure 1): some papers use MAMBA as a drop-in encoder [2,101]; others introduce bidirectional or multi-directional scans to compensate for the unidirectional bias [52]; a third strand combines MAMBA with attention, MLPs, or signal-processing front ends [71,127]. A second wave extended the same architectural toolbox to anomaly detection [80,87], imputation [27, 91], and classification [3,33,50]. A third wave, building on foundational work like TimesNet [104] and TimeMixer++ [99], develops *unified analytics* models targeting multiple tasks under a single backbone [6,10,71]. Cutting across these waves, a separate strand targets specific domains – finance, spatio-temporal grids, and foundation models [13,116].

1.1. Why a MAMBA *Survey for TSA?*

Despite this surge, *no existing survey* treats selective SSMs as a primary object at full TSA scope: prior reviews either target image backbones [57], cover theoretical foundations [92], or focus on Transformers [103], channel strategies [82], or diffusion [117] – almost all narrow to forecasting alone. Three gaps motivate this survey. *First, empirical evidence is mixed*: audits question whether MAMBA-based forecasters consistently beat linear or Transformer baselines [83,101,120], and variants for other TSA tasks lack a shared evaluation regime. *Second, the design space* – tokenization, channel strategy, directional scan, hybrid block, decomposition – *lacks a unified vocabulary*. *Third, cross-task coupling is unmapped*: the five tasks share representation machinery, yet no survey shows how design axes specialize by task.

1.2. Challenges Targeted by This Survey

Long-Context Efficiency. Pre-Mamba evidence shows longer look-back monotonically improves accuracy when the backbone can use it [66], yet self-attention's $\mathcal{O}(L^2)$ cost forces truncation. MAMBA trains in $\mathcal{O}(L \log L)$ and runs inference in $\mathcal{O}(L)$, making $L \geq 1024$ practical [2,71]. The same benefit applies to anomaly detection (rare events demand long evidence) and imputation (irregular EHRs are long and sparse).

Channel Modeling at Scale. Multivariate TSA must balance *channel independence* (robust under shift) against *channel mixing* (higher capacity) [58,66]. MAMBA variants proliferate in both directions – channel-aware (CMamba [121], MambaMixer [11]) and channel-independent (DTMamba [109]) – compete on the same benchmarks, and the optimal choice depends on task as well as data.

Directional Scan Bias. A left-to-right scan suits causal language modeling but is asymmetric for non-causal TSA. Bidirectional and multi-directional designs (Bi-Mamba [52], S-Mamba [101], ms-Mamba [41]) address this in forecasting but introduce concatenation vs. summation, shared vs. separate parameters without a settled best practice. The question is *settled* for non-forecasting tasks (reconstruction, contrastive, discriminative objectives are all non-causal) yet *unsettled* for forecasting itself.

Reproducibility. Small protocol changes in lookback, normalization, or tuning budget can flip method orderings [83,101]. Forecasters report on non-identical subsets of ETT, Electricity, Traffic, Weather, ILI, Solar, and PEMS [34]; anomaly detectors disagree on threshold calibration; imputers on masking

 doi:10.20944/preprints202605.0995.v1

ratios; classifiers on UCR/UEA splits. Headline numbers across MAMBA variants are not directly comparable.

Cross-Task Generalization of the Selective Scan. Forecasting, anomaly detection, imputation, and classification expose different facets of temporal modeling – future prediction, density estimation, conditional reconstruction, and discriminative summarization [99,104]. Whether a single MAMBA backbone transfers across them or each task demands its own design-axis configuration is unsettled; unified analytics models aspiring to one-backbone-fits-all coverage are still rare and the empirical evidence preliminary.

Table 1. Coverage of this survey versus related reviews on seven dimensions (defined in subsection 1.4). ● full, ◐ partial, ○ none.

Survey	Year	Mamba	TSA	Persp.	Axes	Taxon.	Resource	Guide.
DL [12]	2022	○	◐	◐	◐	○	◐	◐
Transformers [103]	2023	○	◐	◐	◐	○	◐	○
Foundation TS [53]	2024	○	◐	◐	○	◐	◐	○
Mamba-360 [72]	2024	●	◐	◐	◐	◐	◐	○
Mamba Survey [84]	2024	●	◐	◐	◐	◐	◐	○
Vision Mamba [57]	2025	●	○	○	◐	○	◐	○
S4 to Mamba [92]	2025	●	◐	◐	◐	◐	○	○
Channel [82]	2025	○	◐	◐	◐	○	◐	◐
DiffusionTS [117]	2026	○	◐	●	○	◐	◐	○
This Survey	**2026**	●	●	●	●	●	●	●

1.3. Contributions

This survey provides the first focused treatment of MAMBA for TSA, organized around four orthogonal *perspectives* – Model, Task, Data, Application – each a valid entry point to the corpus (Figure 4), with the following contributions:

- *Four-perspective body with one master catalog.* Model (section 3), Task (section 4), Data (section 5), and Application (section 6), unified by a one-row-per-method catalog (Table A11) cross-indexed across all perspectives.
- *Unified design vocabulary.* Five orthogonal axes (tokenization, channel strategy, directional scan, hybridization, decomposition) and three architectural patterns, each with a mechanism diagram and design-comparison table.
- *Per-task coverage at full TSA scope.* For each of the five tasks we survey principal designs and provide a comparison table; cross-task synthesis appears in subsection 7.1.
- *Data and application perspectives.* section 5 re-tags the corpus by input regime; section 6 by domain (healthcare, energy, traffic, climate, finance, activity, foundation), exposing patterns invisible at the task level.
- *Practical guidelines* (section 7): a model-selection decision tree, training recipes, and MAMBA-specific pitfalls in numerics and evaluation.
- *Open research frontiers* (section 8): twelve frontiers spanning gain attribution, modeling regimes, evaluation, deployment, and cross-task unification.
- *Reproducibility resources* (Appendix A): benchmarks, datasets, metrics, and a public-implementation registry.

1.4. Differences from Existing Surveys

We compare against every prior survey that overlaps with our scope along either axis: MAMBA-broad reviews that include time series as one of several application domains, and TSA-broad reviews that cover selective SSMs or close neighbors. Table 1 scores these on seven dimensions: **Mamba** (selective SSMs as primary object), **TSA** (full task stack beyond forecasting), **Persp.** (all four orthogonal perspectives), **Axes** (unified design vocabulary over the five axes), **Taxon.** (structural taxonomy), **Resource** (datasets, metrics, code registry), and **Guide.** (selection recipes and pitfalls). The two closest works cover adjacent territory. Liang et al. [53] survey foundation models for time series but absorb

 doi:10.20944/preprints202605.0995.v1

SSMs into a non-Transformer bucket without MAMBA-specific axes. Yang et al. [117] survey diffusion models for time series; their coverage is orthogonal (they focus on output-distribution machinery; we cover the sequence backbone), with overlap handled in our hybridization and imputation sections. Qu et al. [84] and Patro et al. [72] survey MAMBA across multiple domains (vision, NLP, medical imaging, graph learning, and time series), but treat time-series applications as one subsection rather than the primary lens; neither provides the five design axes, the four-perspective taxonomy, nor the practitioner guide developed here. The remaining rows of Table 1 each touch only part of our scope: Benidis et al. [12] survey deep forecasters predating selective SSMs; Wen et al. [103] survey Transformers for time series and treat SSMs as a non-attention alternative; Lin et al. [82] compare channel-modeling strategies orthogonal to backbone choice; Xu et al. [57] survey vision-domain MAMBA variants with 1-D time series only in passing; and Wang et al. [92] trace the S4-to-MAMBA lineage theoretically with limited TSA-specific coverage. No prior survey simultaneously provides MAMBA-specific coverage at full TSA scope, all four perspectives, a unified design vocabulary, and practical guidelines.

1.5. Paper Organization

The survey is organized in three parts. *Part I* (section 2–subsection 2.5) lays the foundation: SSM lineage, the architecture landscape, selectivity, and the four-perspective master taxonomy. *Part II* (section 3–section 6) covers the four perspectives in turn – Model (design axes and patterns), Task (per-task analysis and synthesis), Data (input regimes), and Application (domain surveys). *Part III* (section 7–section 9) closes with a practitioner guide, open challenges, and conclusion.

Reading Guide. Readers new to selective SSMs should start with Part I (section 2–subsection 2.5). Practitioners can enter directly from any Part II perspective and use section 7 as a decision tree for model selection. Open-problem readers can skip ahead to section 8.

2. Background and Categorization

This section locates MAMBA in the sequence-modeling landscape, traces the lineage from S4 to MAMBA and MAMBA-2, and establishes the notation (Table 2) and design vocabulary (subsection 2.4) used throughout the survey.

Table 2. Notation used throughout the survey.

Symbol	Meaning
L	lookback (input) window length
H	forecast horizon
C	number of channels (variates)
P	patch length used by patch-tokenization
N	SSM latent state dimension
$X \in \mathbb{R}^{L\times C}$	multivariate input matrix
$\hat{Y} \in \mathbb{R}^{H\times C}$	forecast
$\mathbf{x}_t \in \mathbb{R}^{C}$	input vector at time step t
$\mathbf{h}_t \in \mathbb{R}^{N}$	SSM hidden state
A, B, C, D	SSM parameter matrices
$\bar{A}, \bar{B}$	discretized SSM matrices
Δ_t	input-dependent step size
f_θ	learnable forecaster
$\mathcal{O}(\cdot)$	asymptotic complexity

2.1. The Time Series Analysis Problem Family

A multivariate time series of length L with C channels is a matrix $X \in \mathbb{R}^{L\times C}$ with rows $\mathbf{x}_t \in \mathbb{R}^{C}$ for $t = 1, \ldots, L$. Univariate analysis is the special case $C = 1$. This survey covers five canonical TSA tasks, each of which the selective-SSM literature has begun to address.

Forecasting. Given a look-back window of L historical observations, the long-term forecasting task predicts the next H steps:

$$\hat{Y} = f_\theta(X_{1:L}), \qquad \hat{Y} \in \mathbb{R}^{H\times C}, \tag{1}$$

 doi:10.20944/preprints202605.0995.v1

where f_θ is the learnable model. This long-horizon setting is commonly abbreviated *long-term time series forecasting (LTSF)*. Standard LTSF horizons are $H \in \{96, 192, 336, 720\}$ with look-back $L \in \{96, 336, 512, 720, 1440\}$ [66,128].

Anomaly Detection. Given an unlabeled training stream of mostly normal behavior, an anomaly score $s_t \in \mathbb{R}$ is emitted for each test step (or window) and thresholded to flag deviations:

$$s_t = g\big(X_{t-L+1:t}\,;\,\theta\big), \qquad \hat{a}_t = \mathbf{1}\big[s_t > \tau\big]. \tag{2}$$

The score g is typically reconstruction error, prediction residual, or contrastive discrepancy [38,93]; τ is calibrated on a held-out normal distribution. Multivariate detection requires s_t to integrate evidence across channels.

Imputation. Given an observed mask $M \in \{0,1\}^{L\times C}$ and the masked input $X \odot M$, the model fills the unobserved positions:

$$\hat{X} = h_\theta(X \odot M,\, M), \qquad \text{loss} = \frac{1}{|\bar{M}|}\sum_{(t,c)\in\bar{M}} \big(\hat{X}_{t,c} - X_{t,c}\big)^2, \tag{3}$$

where $\bar{M}$ marks masked positions where loss is applied. Variants include random missing, structured missing (irregular EHRs, sensor dropouts), and probabilistic imputation (CSDI [94]).

Classification. A whole sequence (or fixed-length window) is mapped to a categorical label $y \in \{1, \ldots, K\}$:

$$\hat{y} = \arg\max_k \ \mathrm{softmax}\big(W\,\phi_\theta(X_{1:L})\big)_k, \tag{4}$$

where ϕ_θ is a backbone encoder (mean-pooled or [CLS]-tokenized) and W is a linear head. UCR/UEA-archive benchmarks [8,23] dominate evaluation; physiological-signal subfields (EEG, ECG, sleep, activity recognition) report on dataset-specific splits.

Unified Multi-Task Analytics. A single backbone ϕ_θ with shared parameters feeds task-specific heads $\{f^{\text{fcst}}, g^{\text{AD}}, h^{\text{imp}}, W^{\text{cls}}\}$, trained jointly or by sequential adaptation across tasks [6,10,71,99,104]. The desideratum is that ϕ_θ generalizes across tasks at substantially lower cost than four specialized models – an aspiration first concretized by TimesNet's 2D backbone, extended by TimeMixer++'s multi-scale pattern machine, and pursued under MAMBA in subsection 4.5.

2.2. Mamba vs. Other Sequence Backbones

Architectures split into two camps: *non-recurrent mixing* (MLP/Linear, CNN, Transformer) and the *recurrence family* (RNN, structured SSM, selective SSM). Figure 2 contrasts six representative mechanisms.

Non-recurrent Mixing (Panels a–c). Three families mix tokens without recurrence. *MLP / Linear* (DLinear [120], N-BEATS [68]) applies shallow per-token projections; the backbone itself does no temporal mixing. *CNN / 2D-reshaping* (TimesNet [104]) imposes a hard locality bias through 1D or 2D receptive windows. *Transformers* (Informer [128], PatchTST [66], iTransformer [58]) mix globally at $\mathcal{O}(L^2)$ cost and need external KV caching for recurrent inference.

Recurrence Family (Panels d–f). The recurrence family carries a state $\mathbf{h}_t \in \mathbb{R}^N$ forward in time, trading content-awareness against parallelizability. *Classical RNNs* (LSTM, GRU) use a *nonlinear* update $\mathbf{h}_t = \sigma(W_h\mathbf{h}_{t-1} + W_x\mathbf{x}_t)$ that is content-aware but $\mathcal{O}(L)$ sequential, surviving in recent practice mainly inside hybrids or xLSTM-style revivals. *Structured SSMs* (S4 [32], S5 [90]) drop the nonlinear gate for a *linear time-invariant* recurrence, enabling $\mathcal{O}(L \log L)$ parallel scans but losing input-dependent selectivity. *Selective SSMs* (MAMBA [30]) make (Δ_t, B_t, C_t) a function of $\mathbf{x}_t$, recovering content-awareness while keeping linearity in $\mathbf{h}_t$.

Training and Inference Paradigms. Orthogonal to the backbone is the *learning paradigm* that wraps it: *deterministic point prediction* (MSE/MAE), *probabilistic prediction* (Gaussian/quantile heads, normalizing

 doi:10.20944/preprints202605.0995.v1

flows, or diffusion – TimeGrad [85], CSDI [94]; see Yang et al. [117]), *reconstruction-based* learning (masked autoencoding for imputation and anomaly detection), and *self-supervised pretraining* (the foundation-model paradigm of Liang et al. [53]). Any backbone above can be inserted into any paradigm; e.g., the denoiser in a diffusion model is commonly a U-Net, Transformer, or – in the hybrids of subsection 4.3 and subsection 4.1 – a MAMBA.

Scope of This Survey. We cover *selective SSMs as sequence backbones* for all five TSA tasks – pure MAMBA, bidirectional scans, and hybrids with Transformer, MLP, or diffusion modules – excluding S4/S5 without selectivity. Probabilistic and foundation-model paradigms are included only when paired with a MAMBA backbone.

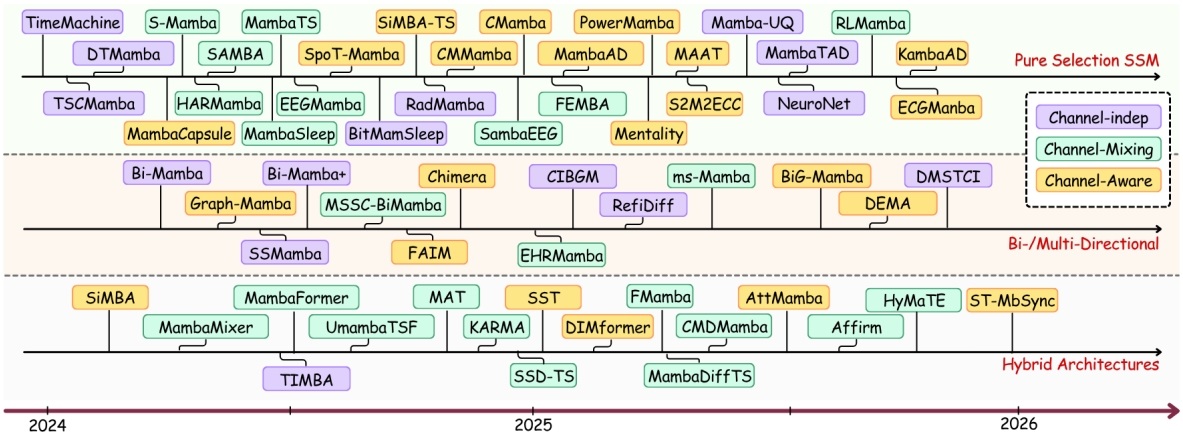

Figure 1. Roadmaps of representative MAMBA models for time series analysis.

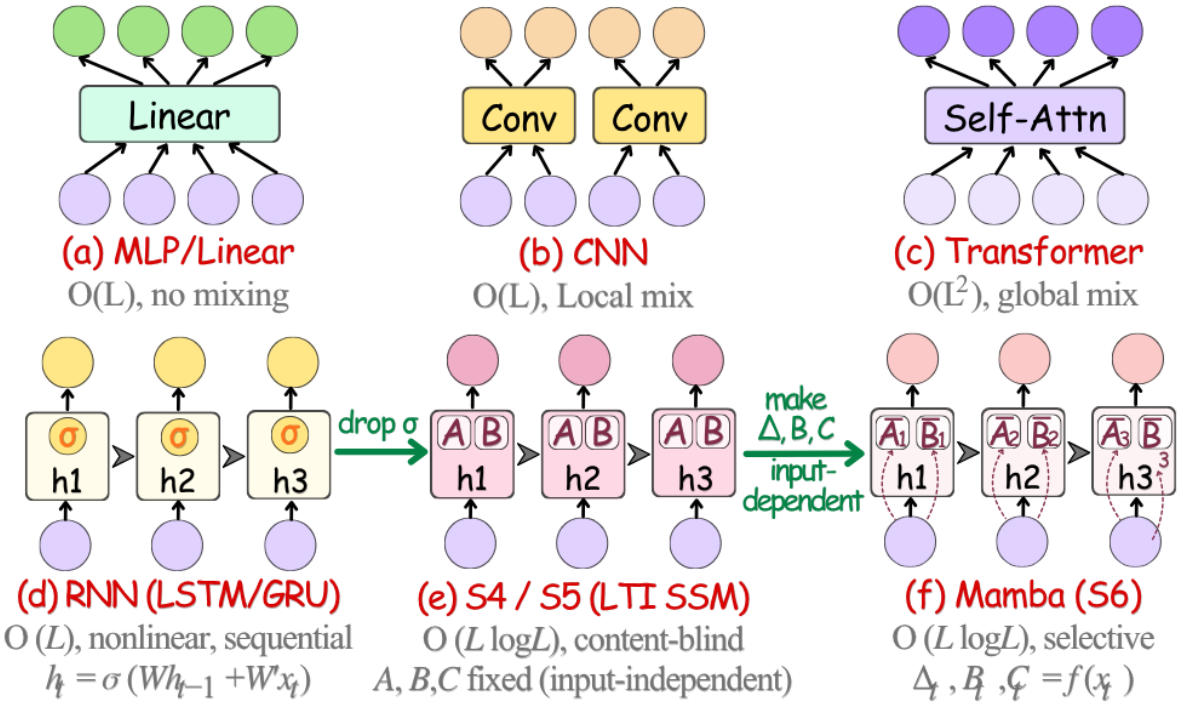

Figure 2. Six sequence backbones: *non-recurrent mixing* (a–c) and the *recurrence family* (d–f). In (d–f) each cell carries state h_t; top chips name the operation (RNN: σ; S4/S5: fixed A, B; MAMBA: red-dashed input-dependent $\bar{A}_t, \bar{B}_t$). Teal arrows trace the progression (d)→(e)→(f).

 doi:10.20944/preprints202605.0995.v1

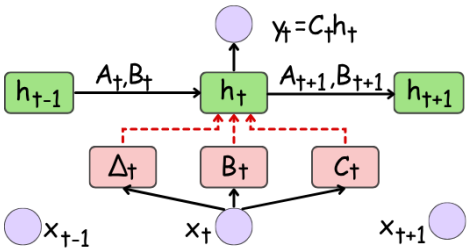

Figure 3. Mamba's selectivity. Input-dependent gates Δ_t, B_t, C_t (red, dashed) modulate state update $h_{t-1} \to h_t$ and read-out $y_t = C_t h_t$, letting it forget stale context after regime shifts.

Table 3. The five design axes for MAMBA-based time-series analysis methods.

Axis	Choices	Representative methods
Tokenization	pointwise / patch / channel-as-token	MambaTS [14] / SiMBA-TS [73] / S-Mamba [101]
Channel strategy	CI / CD / CC / dual-mixer	DTMamba [109] / iTransformer-style / CMamba [121] / MambaMixer [11]
Directional scan	forward / bidir / multi-scale / 2D / walk-based	MambaTS / Bi-Mamba+ [52] / ms-Mamba [41] / Chimera [10] / SpoT-Mamba [21]
Hybridization	none / +attn / +MLP / +FFT / +CNN / +diff / +decomp	MAT [127]/MambaMixer/SiMBA [71]/ CMDMamba [81] / MambaDiffTS [100]/KARMA [118]
Decomposition	none / trend-seasonal / multi-scale / Fourier	Bi-Mamba+ / KARMA / TimeMachine [2] / Affirm [108]

2.3. From HiPPO to S4 and Mamba

State space models originate in control and signal processing:

$$h'(t) = A\,h(t) + B\,x(t), \qquad y(t) = C\,h(t) + D\,x(t), \tag{5}$$

where $h(t) \in \mathbb{R}^N$ is the latent state, $x(t) \in \mathbb{R}$ is the input, and A, B, C, D are time-invariant matrices. Zero-order hold discretization with step Δ gives $\bar{A} = \exp(\Delta A)$ and $\bar{B} = (\Delta A)^{-1}\big(\exp(\Delta A) - I\big)\,\Delta B$, yielding the recurrence

$$\mathbf{h}_t = \bar{A}\,\mathbf{h}_{t-1} + \bar{B}\,x_t, \qquad y_t = C\,\mathbf{h}_t. \tag{6}$$

HiPPO: Principled Long-Range Memory. The High-order Polynomial Projection Operators (HiPPO) framework [31] derived A matrices optimally compressing input into a fixed-dimensional latent state under measure-theoretic objectives. The HiPPO-LegS variant gives $O(\log N)$-error reconstruction over arbitrary intervals and initializes nearly every modern SSM backbone.

S4: Structured State Spaces. S4 [32] made Equation 6 practical for long sequences via (i) a diagonal-plus-low-rank parameterization of A, (ii) the global convolution kernel

$$\bar{K} = (C\bar{B},\ C\bar{A}\bar{B},\ C\bar{A}^2\bar{B},\ \ldots,\ C\bar{A}^{L-1}\bar{B}) \tag{7}$$

computed in $\mathcal{O}(N \log N + L \log L)$ via FFT, and (iii) training in parallel while inferring recurrently. S4 established SSMs as a viable Transformer alternative on Long Range Arena.

S5: Diagonal SSMs. S5 [90] simplifies S4 to a fully diagonal SSM applied jointly across channels via a parallel associative scan, dropping the convolution kernel.

Mamba: Input-Dependent Selectivity. S4 and S5 are linear time-invariant: A, B, C do not depend on the input, so the network cannot *select* what to remember. MAMBA [30] (a.k.a. S6) makes B, C, and Δ functions of $\mathbf{x}_t$:

$$\begin{aligned} B_t &= \mathrm{Linear}_B(\mathbf{x}_t), \quad C_t = \mathrm{Linear}_C(\mathbf{x}_t), \\ \Delta_t &= \mathrm{softplus}\big(\mathrm{Linear}_\Delta(\mathbf{x}_t) + \mathbf{p}_\Delta\big), \end{aligned} \tag{8}$$

 doi:10.20944/preprints202605.0995.v1

yielding the time-varying recurrence

$$\mathbf{h}_t = \bar{A}_t\,\mathbf{h}_{t-1} + \bar{B}_t\,x_t, \qquad y_t = C_t\,\mathbf{h}_t. \tag{9}$$

Selectivity removes the convolution shortcut, requiring S5's parallel scan; MAMBA pairs it with a hardware-aware CUDA kernel that fuses scan and discretization, giving 3× training speedup over optimized attention at $L = 4096$ while keeping $\mathcal{O}(L)$ inference and $\mathcal{O}(L \log L)$ training.

Mamba-2 and the SSD Framework. Mamba-2 [22] introduces state-space duality (SSD), which unifies attention and SSMs and shows that a large class of attention variants are dual to an SSM family. Restricting A to scalar-times-identity per head enables matrix-multiplication-friendly implementations that run 2–8× faster than MAMBA at matched quality – the basis for several recent TSA hybrids [71, 101,127].

The Role of the Input-Dependent Triple. The architectural change that distinguishes MAMBA from S4/S5 is the input-dependent triple (Δ_t, B_t, C_t) in Equation 8. Figure 3 sketches the flow: Δ_t controls how much of the previous state to retain, B_t how new input enters the state, and C_t the read-out. Together they let the network *select* what to remember – something S4/S5 cannot.

Why This Matters for TSA. The original MAMBA ablations [30] show large drops on Selective-Copying and Induction-Heads when any of Δ_t, B_t, C_t is frozen back to its input-independent S5 form. For TSA the picture is mixed: matched-protocol audits find MAMBA variants do not consistently beat DLinear or PatchTST under a shared budget [83,101,120], and post-RevIN [43] benchmarks are nearly instance-stationary – a regime where an LTI SSM should already suffice. subsection 8.1 treats this as the corpus's first open frontier.

What MAMBA Buys Per Task. The five task subsections in section 4 test per-task hypotheses: long-context forecasting at $L \geq 1024$, anomaly detection across regime shifts, imputation with boundary-weighted reconstruction, classification on long physiological windows, and multi-task analytics sharing one selective recurrence across forecasting, reconstruction, and classification heads.

2.4. Design Axes

The surveyed corpus converges on five orthogonal *design choices*; any MAMBA-TSA model can be located by its choice on each (Table 3). Orthogonal to the axes, most models wrap the backbone in Reversible Instance Normalization (RevIN) [43], which subtracts and re-adds instance mean and variance to absorb the train-test mismatch caused by trends and changing variances.

Axis 1: Tokenization. Three strategies convert the series into tokens for the SSM. *Pointwise* treats each step as a token, preserving full resolution at the cost of long sequences. *Patch* tokenization [66] splits the series into windows of length P, cutting token count by P and adding local context. *Channel-as-token* [58] transposes the input so each channel becomes one length-L token, redirecting the scan to mix across channels.

Axis 2: Channel Strategy. Methods are channel-independent (CI [66]; shared weights across channels), channel-mixing (CD [58]; cross-channel attention or mixer projections), channel-correlated (CC [121]; group mixing on correlation-derived clusters), or *dual-mixer* interleaving time- and channel-axis scans in parallel [11]. The choice is *task-conditional*: forecasting tolerates any; anomaly detection and imputation typically benefit from CD/CC; classification depends on the modality.

Axis 3: Directional Scan. The native MAMBA scan is unidirectional. TSA variants explore *forward-only*, *bidirectional* (forward + reverse with concatenation or summation), *multi-scale* (parallel scans at sub-sampling rates), *2D* (joint scan over time and channel), and *walk-based* (multiple random walks over a spatio-temporal graph) configurations.

Axis 4: Hybridization. MAMBA blocks can stand alone or pair with attention, MLP-mixers, convolutional blocks, FFT/spectral mixers, diffusion denoisers, or decomposition modules. Hybrids trade the linear complexity of pure MAMBA for capacity in fine-grained channel mixing or multi-frequency modeling.

 doi:10.20944/preprints202605.0995.v1

Axis 5: Decomposition. Many models apply explicit signal decomposition (trend–seasonal, multi-scale, Fourier) before or in parallel with the sequence backbone, reflecting the Autoformer finding [105] that decomposition simplifies the modeling problem.

2.5. Survey Backbone and Master Taxonomy

The five axes of subsection 2.4 describe *a single model*; the four *perspectives* below organize the *corpus* of models. With the TSA problem family, sequence-backbone landscape, selective SSM lineage, and design axes in place, the MAMBA-TSA literature can be entered from four orthogonal perspectives. The **model perspective** (section 3) uses the five axes directly as its vocabulary, exploring selective-SSM variants and hybrid combinations. The **task perspective** (section 4) reviews state-of-the-art adaptations for forecasting, anomaly detection, imputation, classification, and multi-task analytics. The **data perspective** (section 5) discusses design defaults for various data shapes, while the **application perspective** (section 6) examines domain-specific constraints across fields like healthcare, energy, and finance.

Figure 4 consolidates the corpus into a unified taxonomy, cross-referencing these mutually re-entrant perspectives. A reader can seamlessly trace an application back to its model architecture, or follow a model's citations into its specific task chapter.

3. Model Perspective

The five design axes of subsection 2.4 define a vocabulary, but the corpus does not spread uniformly across that space: along Axis 3 (directional scan) and Axis 4 (hybridization) it clusters into three recurring *patterns*, while the other three axes (tokenization, channel strategy, decomposition) vary *within* each pattern. This section walks the three clusters: pure selective backbones (subsection 3.1; forward-only scan, no hybridization), bi- and multi-directional designs (subsection 3.2; generalized scan, light or no hybridization), and hybrid designs (subsection 3.3; MAMBA paired with attention, MLP, convolution, or spectral modules). Each pattern includes a mechanism figure and a design-summary table; Table 10 shows how the remaining axes specialize by task.

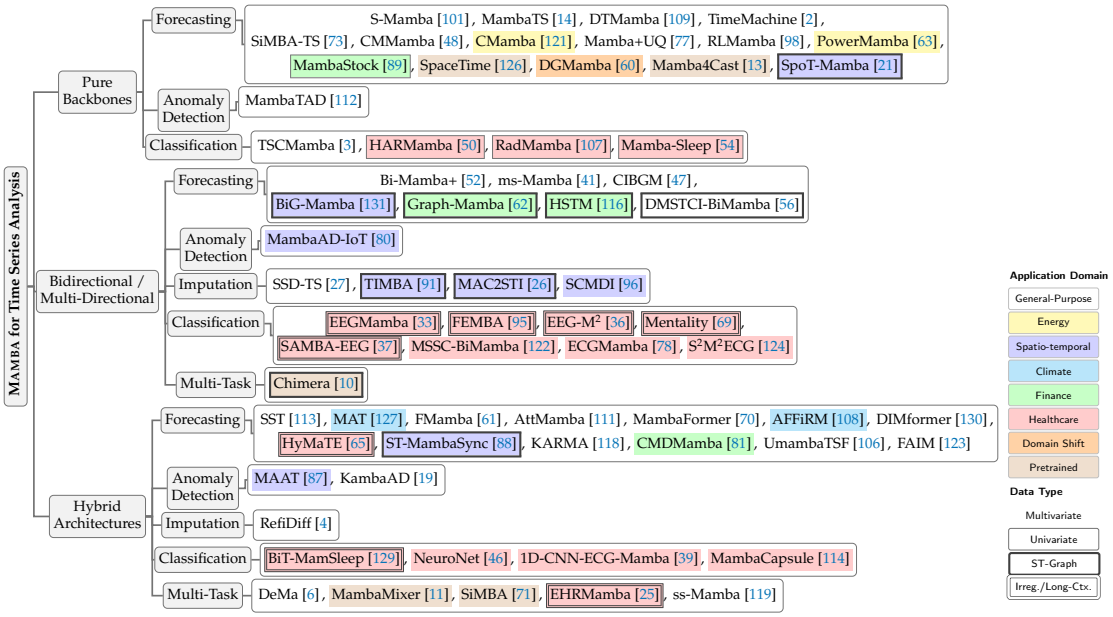

Figure 4. Master taxonomy of MAMBA-TSA. Level 1 branches: architectural pattern (Pure / Bidirectional / Hybrid); Level 2: task. Per-method, *background colour* marks application domain and *box border* marks data type (thin = univariate, thick = STG, double = irregular; multivariate is unbordered).

 doi:10.20944/preprints202605.0995.v1

3.1. *Pure* MAMBA *Backbones*

The *pure* pattern uses MAMBA blocks as the sole mixing layer, handling temporal and cross-channel mixing via the selective scan. Figure 5 shows canonical mechanisms; Table 4 summarizes variants by design axis.

Single-direction Backbones. *S-Mamba* [101] pairs a critical evaluation with a deliberately simplified design: matched-protocol audits show that several published MAMBA variants do not significantly outperform DLinear or PatchTST on ETT and Weather, motivating an iTransformer-style channel-as-token tokenization with a single MAMBA block per layer. On the seven standard ETT/Electricity/Traffic/Weather/Solar benchmarks, this design matches or beats iTransformer at lower training cost. *MambaTS* [14] introduces a Variable-Aware Scan along Time (VAST) that interleaves the C variables *across* time after patching each into $\lfloor L/s \rfloor$ tokens of stride s; the in-block convolution of standard MAMBA is dropped because VAST's ordering is already non-local, and the optimal scan order is decoded as an asymmetric traveling-salesman problem. *S-Mamba* [101] strips the gated MLP wrapper of the original MAMBA block, matching prior pure-MAMBA baselines with fewer parameters. *DTMamba* [109] runs two same-direction MAMBA blocks in parallel per layer with *independent* parameters for parameter-diversity (not bidirectionality), stacked at decreasing hidden size (n_1=256, n_2=128). *TimeMachine* [2] arranges four MAMBA blocks as two resolution stages – an outer pair on the n_1-token compressed representation and an inner pair on the n_2-token representation – with one block per pair scanning the length axis and the other the dimension axis, recovering the effect of a bidirectional scan without two full time-axis passes. *RLMamba* [98] integrates classical residual learning so each block produces a delta over the preceding layer.

Patch-based Pure SSM Backbones. *SiMBA-TS* [73] applies the SiMBA [71] block – an EinFFT channel mixer fused with a MAMBA time mixer – to long-term TSF benchmarks under patch tokenization. *Mamba+UQ* [77] is the first selective-SSM forecaster with calibrated probabilistic outputs: two parallel MAMBA backbones predict heteroscedastic mean and dispersion under joint NLL minimization, so each horizon step is modeled as $x_{L+\tau} \mid X_{1:L} \sim \mathcal{N}(\mu_\tau, \sigma_\tau^2)$ with single-pass inference. *CMamba* [121] targets the channel-correlation gap with an M-Mamba block (no x-branch convolution; feature-independent shared A; data-dependent skip D) and a global data-dependent (GDD) MLP that mixes channels via input-conditioned scale and shift; a Channel Mixup augmentation linearly mixes channels within the same sample during training. *CMMamba* [48] pairs a bidirectional MAMBA backbone with a similarity-weighted Top-K cross-channel aggregator, mirroring the token-mixer/channel-mixer alternation popularized by MLP-Mixer.

Domain-conditioned Pure Backbones. Two pure-pattern entries condition the design on domain structure. *PowerMamba* [63] is the first explicitly energy-specialized design, pairing a dual-MAMBA encoder with an external-predictor fusion head that consumes system-operator forecasts; it is paired with the GridSet 5-year ERCOT benchmark and discussed under healthcare/energy in subsection 6.2. *SpoT-Mamba* [21] generates multiple random walks over the sensor graph and feeds each walk as a separate sequence to a shared MAMBA encoder, fusing the per-walk outputs with a learned weighting; the construction targets spatio-temporal benchmarks (subsection 6.3).

3.2. *Bidirectional and Multi-Directional Scans*

Pure MAMBA inherits the causal recurrence of a left-to-right scan – natural for autoregressive language modeling, asymmetric for non-causal time series tasks. The *bidirectional* pattern runs separate scans forward and reverse and combines their hidden states; the *multi-directional* variant extends this to multi-scale parallel scans, 2D scans over time×channel grids, or walk-based scans on spatio-temporal graphs. Figure 6 shows the canonical mechanisms and the bi-/multi-directional rows of Table 4 summarize the variants. Bidirectional patterns dominate the non-forecasting tasks (anomaly detection, imputation, classification) because their training objectives – reconstruction, contrastive, or discriminative – are inherently non-causal; they remain a contested choice within forecasting itself.

 doi:10.20944/preprints202605.0995.v1

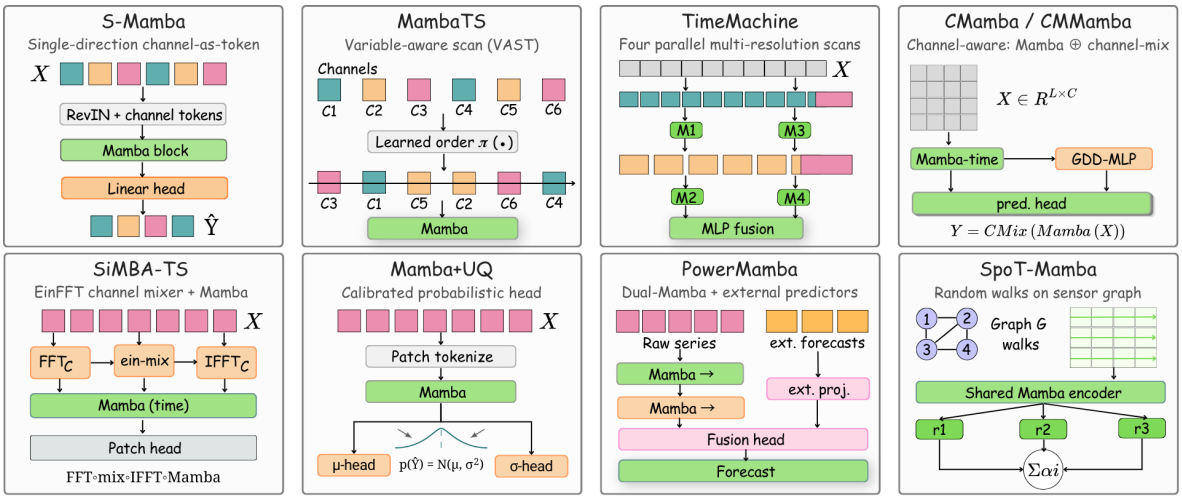

Figure 5. Pure MAMBA models. *S-Mamba*: single block. *MambaTS*: channel-token reordering. *TimeMachine*: multi-resolution MLP fusion. *CMamba/CMMamba*: channel-mixer after Mamba. *SiMBA-TS*: EinFFT + Mamba mixer. *Mamba+UQ*: probabilistic twin heads. *PowerMamba*: dual scans + external forecasts. *SpoT-Mamba*: shared encoder over graph walks.

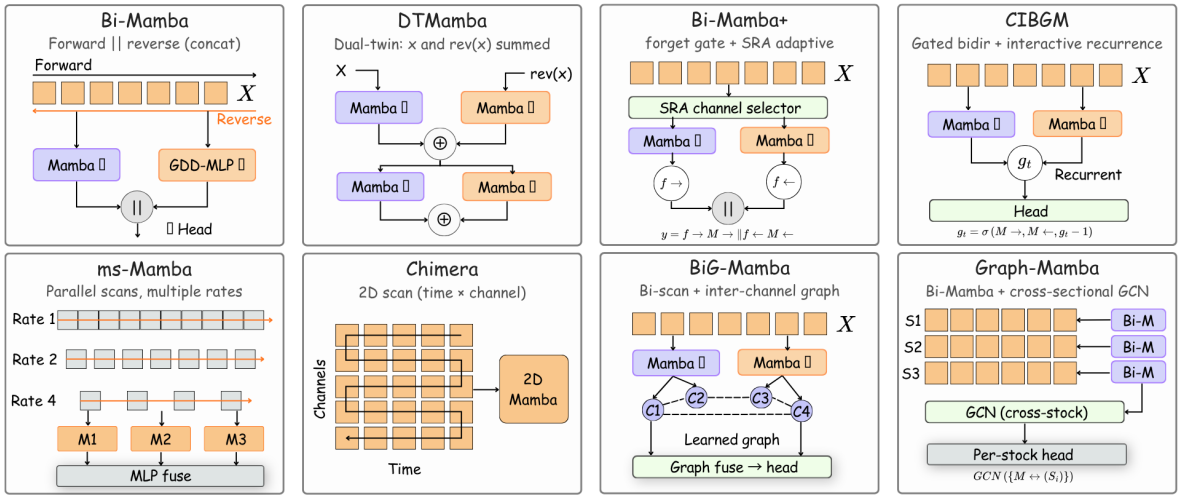

Figure 6. MAMBA models with bi- and multi-directional selective scans. *Bi-Mamba*: fwd+rev branches fused by concatenation. *Bi-Mamba+*: adds SRA and channel-strategy gate. *ms-Mamba*: multi-resolution parallel scans fused via MLP. *CIBGM*: gated bidirectional with interactive recurrence. *Chimera*: 2D scan over time×channel grid. *Graph-Mamba*: bidirectional Mamba with graph attention. *HSTM*: hierarchical spatio-temporal bidirectional Mamba.

Symmetric Bidirectional Designs. *Bi-Mamba+* [52] runs forward and reverse MAMBA scans in parallel and concatenates the outputs along the channel dimension; on top of this it adds a forget-gate variant of the MAMBA block and a series-relation-aware (SRA) decider that selects between channel-independent and channel-mixing tokenization per dataset. Applied along both intra-series (time) and inter-series (channel) axes, it consistently outperforms single-direction variants on ETT and Electricity. Bi-Mamba+ supersedes the earlier Bi-Mamba4TS draft (same arXiv ID, since collapsed on Google Scholar); the forget gate ties the new conv/SSM feature x' to the selective-scan output y through complementary sigmoids, preserving longer-range memory than vanilla MAMBA. *CIBGM* [47] adds an interactive recurrent mechanism that gates the bidirectional states with a learned recurrence over the gate values themselves.

Graph-augmented Bidirectional Designs. *BiG-Mamba* [131] couples a forward+reverse MAMBA scan with a learned inter-channel graph attention module, using the graph to mix bidirectional state across correlated variables for traffic-style benchmarks where the sensor adjacency matrix is informative but loosely grid-aligned. *Graph-Mamba* [62] pairs a bidirectional MAMBA scan over each stock's price

 doi:10.20944/preprints202605.0995.v1

history with an end-to-end learned graph convolution that mixes information across the cross-section of stocks; the inter-stock graph is built from a learnable node-embedding matrix via a Gaussian-kernel adjacency, followed by a K-order Chebyshev polynomial graph convolution on the bidirectional MAMBA output. The two designs differ in how the graph mixes with the scan: BiG-Mamba uses the graph as an attention layer between scans, while Graph-Mamba applies the graph convolution *after* the scan.

Multi-directional and Multi-Scale Scans. *ms-Mamba* [41] runs parallel MAMBA blocks with different discretization steps Δ_i, realizing multi-scale temporal coverage by varying Δ rather than by sub-sampling the input. *Chimera* [10] extends MAMBA to a genuine 2D recurrence on the time–channel grid: each grid cell (t, v) carries two hidden states, one accumulating along time under matrices $(\bar{A}_1, \bar{A}_2)$ and one along the variate axis under $(\bar{A}_3, \bar{A}_4)$, with output $y_{t,v} = C_1\mathbf{h}^{(1)}_{t,v} + C_2\mathbf{h}^{(2)}_{t,v}$ and data-dependent B, C, Δ on both axes. The 2D recurrence is executed as a parallel associative scan; bidirectionality applies only to the variate axis (the time axis remains causal). *DMSTCI-BiMamba* [56] generalizes ms-Mamba to bidirectional scans and adds an Autoformer-style decomposition front end that splits the input into trend and multi-scale seasonal components, each processed by a parallel bidirectional MAMBA branch with temporal–channel interaction.

3.3. Hybrid Architectures

The *hybrid* pattern pairs a MAMBA block with a different mixing primitive – Transformer (global attention), MLP/EinFFT mixer (channel/frequency mixing), convolution (local features), diffusion denoiser (probabilistic outputs), or signal-decomposition (trend/seasonality). Figure 7 and the hybrid rows of Table 4 summarize the variants. Hybrids dominate forecasting sub-branches and reappear in imputation (diffusion+MAMBA) and anomaly detection (attention+MAMBA).

Mamba + Attention. *MAT* [127] processes the look-back in two parallel branches – a MAMBA branch for slow long-range trend and a Transformer branch with multi-head self-attention for short-range fluctuations on the same patch-tokenized input – combined by a learned input-dependent gate so the routing varies per time step. *SST* [113] uses a two-expert mixture in which a MAMBA expert handles long-horizon global patterns and a Transformer expert handles short-horizon local patterns, with a soft (not sparse top-k) gating network producing the routing weight. *FMamba* [61] runs a fast linear-attention branch $\phi(q)\phi(k)^\top v$ in parallel with the selective scan, with the two branches axis-orthogonal: the MAMBA scan operates on time and the linear-attention kernel on the channel axis. *MambaFormer* [70] interleaves MAMBA and Transformer blocks in alternating layers. *AttMamba* [111] wraps the MAMBA block in an Adaptive Average Pooling front-end (downsamples $L \to L'$ before scan and self-attention) and a Receptive-Field-Enhancement module (parallel dilated 1D convolutions $\{d{=}1, 2, 4, \dots\}$). *DIMformer* [130] embeds a MAMBA block inside an iTransformer-style channel-as-token backbone, replacing self-attention with MAMBA-driven cross-variable linear attention. *HyMaTE* [65] encodes each patient's event sequence with MAMBA blocks and adds a Transformer block to mix across the medical-code dimension; it is the cleanest example in the corpus of a domain-driven MAMBA +Transformer hybrid for healthcare (subsection 6.1). *ST-MambaSync* [88] fuses a bidirectional MAMBA branch for linear-time temporal modeling with a lightweight ST-Transformer branch for spatial attention over the sensor graph, cutting the compute cost of full ST-Transformer baselines while preserving accuracy on METR-LA / PEMS-BAY / PEMS03–08.

Mamba + MLP / EinFFT Mixers. *SiMBA* [71] stacks a MAMBA time mixer and an EinFFT channel mixer sequentially: $X \to$ Mamba $\to$ EinFFT $\to +$res. The EinFFT block applies a real 1D FFT along the channel dimension, mixes in the spectral domain via a learned complex einsum, and inverts via IFFT, replacing the MLP/Monarch channel mixer used in earlier vision-MAMBA variants. *MambaMixer* [11] realizes both token and channel mixing via two selective S6 blocks per layer: the token-axis scan sweeps the L-dimensional sequence and the channel-axis scan sweeps the C-dimensional variate axis, treating the channel dimension itself as a sequence to be selectively scanned. Both scans are bidirectional and carry their own input-dependent $(\bar{A}_t, \bar{B}_t, C_t, \Delta_t)$, with a learnable weighted-averaging module fusing

 doi:10.20944/preprints202605.0995.v1

features from preceding blocks. *KARMA* [118] pairs a classical decomposition front-end with one MAMBA block per frequency band: STL splits the input into trend, seasonal, and residual; each is routed to its own MAMBA branch and the outputs are summed and linearly projected.

Mamba + CNN / Decomposition. *CMDMamba* [81] replaces MAMBA's gated MLP with a dual-convolution FFN containing a temporal depth-wise branch and a cross-variable point-wise branch; the backbone itself is a dual-layer MAMBA stack with a shallow short-receptive-field layer and a deep long-receptive-field layer. The arrangement explicitly targets noise suppression in financial series. *UmambaTSF* [106] adopts a U-shaped encoder–decoder with MAMBA blocks at each level for multi-scale processing of long lookback windows. *Affirm* [108] places an adaptive Fourier filter (AFF) branch alongside a bidirectional MAMBA inside each block: the AFF takes a real FFT along the time axis, multiplies by a data-dependent complex filter that modulates both amplitude and phase, then inverts via IFFT, so frequency-domain filtering is input-conditioned rather than realized by static global spectral weights.

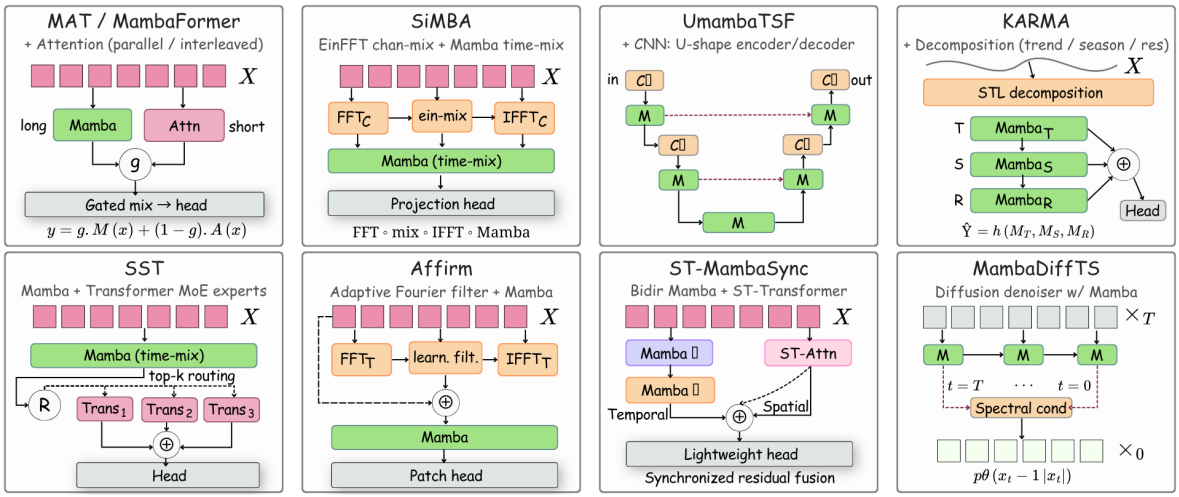

Figure 7. Hybrid MAMBA models. *MAT*: parallel Mamba+Transformer branches for weather. *SST*: Mamba–Transformer MoE. *SiMBA*: Mamba with EinFFT channel mixer. *MambaMixer*: dual token+channel selective mixers. *Affirm*: bidirectional Mamba with adaptive Fourier filters. *KARMA*: decomposition+Mamba with hierarchical fusion. *ST-MambaSync*: Mamba–Transformer for traffic. *MambaDiffTS*: Mamba encoder + diffusion decoder.

Mamba + Diffusion. A distinct hybrid family uses MAMBA as the *denoiser backbone* inside a diffusion model, combining the linear-time scan with probabilistic forecasting or imputation. *TIMBA* [91] replaces the time-oriented Transformer inside CSDI [94] and PriSTI with bidirectional MAMBA blocks, keeping the conditional score-based diffusion formulation intact. *SSD-TS* [27] (originally DiffImp; KDD'25) attaches a Bidirectional Attention Mamba block for forward-and-reverse temporal scanning and a Channel Mamba Block for inter-channel mixing as the denoiser inside a DDPM. *MambaDiffTS* [100] targets long-horizon forecasting (rather than imputation): a MAMBA encoder captures long-range dependencies in linear time, a frequency-aware spectral decomposition isolates trend and seasonal components, and a spectral-energy-guided noise schedule adapts the diffusion forward process to the signal's frequency content. These hybrids are the right choice when the application demands a full predictive distribution over long horizons; the linear-time scan alleviates the main scaling bottleneck of Transformer-based diffusion denoisers.

4. Task Perspective

Where section 3 grouped models by their architectural pattern, this section re-tags the *same* corpus by task: each task induces task-conditional *defaults* along each of the five design axes (subsection 2.4). For each of the five TSA tasks we state the problem, identify which axis choices the surveyed MAMBA methods converge on, and report a comparison table with open issues. Forecasting (subsection 4.1) holds two-thirds of the corpus and is the only task with a contested directional choice; anomaly

 doi:10.20944/preprints202605.0995.v1

detection (subsection 4.2), imputation (subsection 4.3), classification (subsection 4.4), and multi-task analytics (subsection 4.5) are smaller and default to bidirectional scans. subsection 7.1 consolidates the pattern.

4.1. Forecasting

Forecasting is the largest and most mature segment of the MAMBA-TSA corpus, holding two-thirds of the surveyed methods. MAMBA's advantage is sharpest here: *linear-time recurrence with input-dependent gating delivers long-context benefits that quadratic attention can only afford truncated.* All three architectural patterns of section 3 appear; Table 4 summarizes the corpus along the five design axes (subsection 2.4).

Table 4. MAMBA forecasting methods grouped by architectural pattern: pure selective backbones (subsection 3.1), bi-/multi-directional scans (subsection 3.2), and hybrid designs (subsection 3.3). Columns are the five design axes of subsection 2.4. Cells use the axis vocabulary from Table 3; method-specific flavors appear in parentheses. – marks the axis default (no decomposition; no hybrid partner; forward scan).

Method	Tokenization	Channel	Direction	Hybridization	Decomposition
Pure selective backbones					
S-Mamba [101]	ch.-token	CD	forward	–	–
MambaTS [14]	pointwise	CC	forward (VAST)	–	–
DTMamba [109]	patch	CI	forward (dual)	–	–
TimeMachine [2]	patch	CI	multi-scale	–	multi-scale
SiMBA-TS [73]	patch	CD	forward	–	–
CMamba [121]	patch	CC	forward	–	–
CMMamba [48]	patch	CC	bidir	–	–
Mamba+UQ [77]	patch	CI	forward	–	–
PowerMamba [63]	ch.-token	CC	forward (dual)	–	–
SpoT-Mamba [21]	pointwise	CC	walk-based	–	–
RLMamba [98]	patch	CI	forward	–	–
Bidirectional and multi-directional scans					
Bi-Mamba+ [52]	patch	dual-mixer	bidir	–	–
CIBGM [47]	patch	CI	bidir (gated)	–	–
ms-Mamba [41]	patch	CD	multi-scale	+MLP (fusion)	–
Chimera [10]	patch (2D)	CD	2D	–	–
BiG-Mamba [131]	ch.-token	CC	bidir	+attn (graph)	–
Graph-Mamba [62]	ch.-token	CC	bidir	+CNN (graph)	–
DMSTCI-BiMamba [56]	patch	CD	bidir + multi-scale	–	multi-scale
Hybrid architectures					
MAT [127]	patch	CD	–	+attn (Transformer)	–
SST [113]	patch	CD	–	+attn (MoE)	–
FMamba [61]	ch.-token	CD	–	+attn (fast)	–
MambaFormer [70]	patch	CI	–	+attn (interleave)	–
HyMaTE [65]	pointwise	CC	–	+attn (channel)	–
ST-MambaSync [88]	patch	CC	–	+attn (ST)	–
AttMamba [111]	patch	CD	–	+attn (adaptive)	–
DIMformer [130]	ch.-token	CD	–	+attn (linear)	–
SiMBA [71]	patch	CD	–	+FFT (EinFFT)	–
MambaMixer [11]	patch	CC	–	+MLP (mixer)	–
KARMA [118]	patch	CD	–	+MLP + decomp.	trend-seasonal (STL)
Affirm [108]	patch	CD	–	+FFT (Fourier filt.)	Fourier
CMDMamba [81]	patch	CD	–	+CNN (dual conv.)	–
UmambaTSF [106]	patch	CI	–	+CNN (U-Net)	–
Mamba + Diffusion (denoiser)					
MambaDiffTS [100]	patch	CD	–	+diff (DDPM)	Fourier
TIMBA [91]	pointwise	CD	–	+diff (CSDI-style)	–
DiffImp [27]	pointwise	CC	–	+diff (DDPM)	–

 doi:10.20944/preprints202605.0995.v1

Pure MAMBA Forecasters. The simplest design substitutes MAMBA for the encoder of an existing TSF pipeline (RevIN, patch or channel tokenization, SSM stack, linear head), testing whether MAMBA alone can match Transformer or MLP baselines. Mechanisms live in subsection 3.1; here we summarize the forecasting-specific findings.

The corpus splits along channel strategy. *Channel-token forwards* (S-Mamba [101], MambaTS [14]) match or beat iTransformer at lower compute on ETT, Electricity, Traffic, Weather, and Solar. *Channel-correlated patches* (CMamba [121], CMMamba [48], PowerMamba [63]) win on benchmarks with rich cross-channel structure. *Channel-independent multi-scale* designs (TimeMachine [2], DTMamba [109]) tie the channel-token forwards on small-C data (ETT, Weather, Exchange) at lower parameter count. *Spectral-mixer hybrids* (SiMBA-TS [73]) are pure on the time axis (their inclusion here) and hybrid on the channel axis. The S-Mamba audit [101] shows that several reported gains over PatchTST vanish under matched protocol, motivating the more careful ablations in 2025 submissions. Probabilistic outputs entered the corpus with Mamba+UQ [77]; the rest remain point-prediction only.

Bidirectional and Multi-Directional Scans. The causal forward scan is asymmetric for forecasting, where past and recent context both inform future values. Bidirectional and multi-directional variants (subsection 3.2) recover non-causal context. The corpus shows three findings. *Symmetric bidirectional designs* (Bi-Mamba+ [52], CIBGM [47]) improve over single-direction baselines on ETT and Electricity at modest compute cost; Bi-Mamba+ also picks CI/CD from data, removing a hyperparameter. *Multi-scale and 2D scans* (ms-Mamba [41], Chimera [10], DMSTCI-BiMamba [56]) add gains on multivariate benchmarks with cross-channel structure; Chimera is the only true 2D recurrence on the time–channel grid. *Graph-augmented bidirectional designs* (BiG-Mamba [131], Graph-Mamba [62], SpoT-Mamba [21]) target spatio-temporal forecasting and are covered under subsection 5.3 and subsection 6.3–subsection 6.5.

Hybrid Architectures. Hybrid forecasters pair MAMBA with attention, MLP/EinFFT mixers, decomposition, CNNs, or a diffusion loop to address weaknesses of the pure-SSM design (channel mixing, frequency content, local patterns, probabilistic output). Mechanisms are catalogued in subsection 3.3; here we summarize forecasting patterns.

MAMBA +*attention* dominates where short- and long-range patterns coexist (MAT [127], SST [113], AttMamba [111], DIMformer [130]). MAMBA +*spectral mixers* (SiMBA [71], Affirm [108]) shine on Weather and Solar. MAMBA +*decomposition* (KARMA [118], DMSTCI-BiMamba [56]) wins on heavily seasonal series. MAMBA +*CNN* (CMDMamba [81], UmambaTSF [106]) targets noise suppression and multi-scale long lookbacks. MAMBA +*diffusion* (MambaDiffTS [100], with TIMBA [91] and SSD-TS [27] extending to imputation; subsection 4.3) ships calibrated probabilistic forecasts at linear scan cost, the K-step reverse process being the only remaining bottleneck.

4.2. Anomaly Detection

Problem Setting and Selectivity Hypothesis. Anomaly detection flags steps or windows of unusually low conditional likelihood, via three dominant formulations: *reconstruction-based* (large reconstruction error signals an anomaly), *prediction-based* (forecast residual as score), and *contrastive* (view-discrepancy between two representations of the same window). Multivariate detection additionally requires calibrated scores across correlated channels; strong baselines are Anomaly Transformer (attention-residual) and TranAD (adversarial).

The hypothesis selective SSMs raise: *linear-time long-context reconstruction stabilizes the score distribution at horizons where attention truncates.* IoT, industrial, and clinical streams routinely exceed L=10,000 steps; truncating to $L \leq 512$ for attention shrinks the evidence pool that defines "normal," and selective gating additionally lets the head re-focus on recent context after a regime shift, avoiding the smoothing that hurts attention detectors.

MAMBA Anomaly Detectors. We organize the (still small) corpus by detection paradigm. The overall design pattern is consistent: bidirectional scan as default, a small hybridization that preserves a discrepancy or local-feature signal, and CD/CC channel handling. Table 5 summarizes the methods.

 doi:10.20944/preprints202605.0995.v1

Table 5. Mamba-based time-series anomaly detection methods. **Detection** = how the score is produced (R = reconstruction, P = prediction residual, C = contrastive discrepancy); remaining columns follow subsection 2.4.

Method	Detection	Token.	Direction	Hybrid.
MAAT [87]	R	pointwise	forward	sparse attn
MambaAD-IoT [80]	R	patch	bi-direction	–
MambaTAD [112]	C	patch	bi-direction	view-discrep.
KambaAD [19]	R	patch	forward	KAN + attn

MAAT [87], the *Mamba Adaptive Anomaly Transformer*, fuses an Anomaly-Transformer association-discrepancy branch with a MAMBA-SSM reconstruction branch via a Gated Attention module and skip connections; evaluated on MSL, SMAP, SWaT, PSM, and SMD, the selective scan provides the linear-cost reconstruction backbone while the attention branch retains the discrepancy signal that pure reconstruction misses on short anomalies, instantiating the *hybrid + attention* pattern of subsection 3.3. *MambaAD-IoT* [80] – a separate paper from the unrelated image-anomaly NeurIPS 2024 model of the same name – operates on multivariate IoT time series, pairing two parallel bidirectional MAMBA branches (one over the temporal axis, one over the inter-signal correlation axis) with a masked-token augmentation and a contrastive auxiliary to expose both intra-channel and inter-channel anomalies in a single sweep, applying the *bidirectional* pattern of subsection 3.2 twice in parallel. *MambaTAD* [112] adopts a multi-scale patch tokenization with dual MAMBA encoders over inter- and intra-patch views, scoring anomalies through view discrepancy in a contrastive learning framework rather than reconstruction error – sidestepping the masking design choices that complicate reconstruction-based detectors, at the cost of requiring augmented view pairs. *KambaAD* [19] composes a Kolmogorov–Arnold Network for data-consistency enforcement, a distributional balancing attention block, and a MAMBA block for local variation, attached to a patch-based reconstructor; the KAN front end addresses an under-discussed failure mode of reconstruction-based detectors, namely the collapse to in-distribution-looking outputs that ignore the anomalous signal entirely.

What Selectivity Contributes. Across these designs, selectivity offers two task-specific benefits. First, the input-dependent gating lets the reconstruction model *forget* stale context after a regime shift, which is precisely the failure mode of fixed-context attention detectors that treat all recent steps equally. Second, the linear-time scan permits long windows over which the score distribution can be calibrated to a realistic-prior tail, avoiding the high false-positive rate that short-window detectors exhibit on rare-event regimes.

Bidirectional scans dominate the design choices, mirroring the forecasting-side observation in subsection 3.2. Attention hybrids dominate over MLP or convolutional hybrids because the discrepancy signal that classical detectors rely on is hard to recover from pure reconstruction; mixing in even a sparse attention branch (as in MAAT) recovers most of the discriminative power.

4.3. Imputation

Problem Setting and Selectivity Hypothesis. Multivariate imputation fills missing or irregularly-sampled positions conditioned on observed entries. Two regimes dominate: *random missing* (MCAR masks on dense series) and *structured missing* (clinical EHRs, sensor dropouts, calibration gaps); strong baselines are CSDI [94] (diffusion), BRITS (recurrent), and SAITS (attention). Imputation is either *deterministic* (single best estimate) or *probabilistic* (posterior over plausible fills).

The hypothesis selective SSMs raise: *linear-time bidirectional reconstruction is a clean fit for diffusion imputers, where reverse-process iterations multiply the per-step cost; a quadratic backbone pays the long-context cost twice.* The irregular-time setting reinforces it: long, sparse, heavy-tailed sequences are exactly where attention imputers truncate.

MAMBA Imputers. The pattern across imputers is uniform: bidirectional scan, diffusion-based denoising, with the channel-mixing strategy varying by deployment. Table 6 summarizes the methods.

 doi:10.20944/preprints202605.0995.v1

Table 6. Mamba-based time-series imputation methods. **Output** = D (deterministic) or P (probabilistic); remaining columns follow subsection 2.4.

Method	Output	Token.	Direction	Hybrid.
SSD-TS [27]	P	patch	bi-direction	diffusion
TIMBA [91]	P	patch	bi-direction	diff. + GNN
MAC2STI [26]	D	graph-node	bi-direction	cluster-aware
SCMDI [96]	P	patch	bi-direction	diff. + attn.
RefiDiff [4]	P	pointwise	bi-direction	local-ML + diff.

SSD-TS [27] (originally DiffImp; KDD 2025) replaces the Transformer denoiser inside a CSDI-style diffusion imputer with bidirectional MAMBA blocks – a Bidirectional Attention Mamba (BAM) for temporal scanning and a Channel Mamba Block (CMB) for inter-channel mixing – reporting SOTA imputation at high missing ratios with linear sequence-length cost. **TIMBA** [91] extends the idea spatio-temporally, pairing the bi-directional S6 time-axis denoiser inside a CSDI/PriSTI imputer with a graph-based node-Transformer across channels. **MAC2STI** [26] embeds clustering features into the selective state-transition matrix, specializing MAMBA's transitions per channel cluster for sparse traffic and weather imputation. **SCMDI** [96] (IEEE IoT-J) targets IoT data with a Mamba-attention dual module under a causal diffusion process. **RefiDiff** [4] extends the SSD-TS recipe to mixed-data MNAR settings via a local-ML predictor and a Mamba denoising network.

What Selectivity Contributes. Imputation objectives are inherently non-causal, so bidirectional scans dominate; the design axes that vary most across imputers are hybridization (diffusion vs. direct reconstruction) and channel strategy (graph-aware vs. permutation-invariant vs. cluster-aware). Two task-specific benefits stand out. First, the linear-time backbone amortizes the repeated denoiser application that diffusion demands – a quadratic backbone there is twice as expensive at long sequences as a quadratic backbone in a one-shot forecaster. Second, the input-dependent gating lets the imputer attend selectively to the boundary observations that bracket each masked region, which fixed-context attention treats uniformly.

4.4. Classification

Problem Setting and Selectivity Hypothesis. Time series classification (TSC) assigns categorical labels to whole sequences or fixed-length windows. Two families dominate the deep-learning era: *end-to-end* encoders that emit a label from a pooled or attended representation (InceptionTime, TST, PatchTST-CLS), and *representation-learning* backbones followed by a linear probe (TS2Vec, TS-TCC).

The selectivity hypothesis: *linear-time scaling lets the classifier see arbitrarily long windows when class signal is dispersed, and the input-dependent gating can suppress label-irrelevant frequency content without explicit decomposition*. The hypothesis is task-dependent; it pays off most for physiological-signal classification (EEG, ECG, sleep) where windows are long and class signal is dispersed across time, and less clearly on the short-window UCR archive where attention models already excel.

The corpus splits into three sub-families: *generic UCR/UEA* classifiers, *activity recognition* on wearable / radar / multimodal sensors, and *physiological-signal* classifiers (EEG, ECG, sleep). Table 7 summarizes all three.

Generic UCR / UEA Classifiers. *TSCMamba* [3] fuses two views of the input – continuous wavelet transform (CWT) spectral coefficients and the raw temporal sequence – through a "tango" bidirectional scan over the full UEA archive; published in *Information Fusion*, it is the strongest evidence in the surveyed corpus that MAMBA matches Transformer classifiers at substantially lower cost. *FAIM* [123] pairs an Adaptive Filtering Block (learnable Fourier filter bank) with an Interactive Mamba Block (dual-causal convolutions), reporting SOTA accuracy on UCR (0.849) and UEA (0.765). Both entries address the same observation from different angles: spectral information is discriminative for classification but lossy under pure temporal encoding.

 doi:10.20944/preprints202605.0995.v1

Table 7. Mamba-based time-series classification methods. Sub-blocks: *Generic* (UCR/UEA-style sequence classifiers), *Activity & Wearable* (HAR / IMU / radar), *Physiological* (EEG / ECG / sleep). Columns follow the design axes of subsection 2.4.

Method	Token.	Channel	Direction	Hybrid.
Generic UCR / UEA				
TSCMamba [3]	multi-scale	CD	multi-dir.	wavelet
FAIM [123]	patch	CD	bi-direction	Fourier
Activity / Wearable HAR				
HARMamba [50]	patch	CI	bi-direction	–
RadMamba [107]	Doppler-aligned	CD	bi-direction	–
Mamba-Sleep [54]	pointwise	CI	bi-direction	–
Physiological (EEG / ECG / Sleep)				
EEGMamba [33]	patch	adaptive	bi-direction	MoE
FEMBA [95]	patch	CD	bi-direction	conv
EEG-M^2 [36]	patch	adaptive	bi-direction	Mamba-2
MSSC-BiMamba [122]	patch	CC	bi-direction	ECA
ECGMamba [78]	patch	CD	bi-direction	conv
S^{2}M^{2}ECG [124]	patch	CC	bi-direction	spatial
1D-CNN-ECG-Mamba [39]	patch	CD	bi-direction	conv
MambaCapsule [114]	patch	CD	forward	capsule
BiT-MamSleep [129]	patch	CC	bi-direction	TRCNN
NeuroNet [46]	patch	CI	bi-direction	SSL
Mentality [69]	patch	CD	bi-direction	SSL
SAMBA-EEG [37]	patch	adaptive	bi-direction	diff-Mamba

Activity / Wearable Recognition. *HARMamba* [50] targets wearable-IMU human activity recognition with a patch-tokenized, channel-independent bidirectional MAMBA backbone, delivering competitive accuracy at edge-device parameter counts on PAMAP2, WISDM, UNIMIB-SHAR, and UCI-HAR (IEEE IoT-J). *RadMamba* [107] extends HAR to radar micro-Doppler with a Doppler-aligned segmentation tokenizer, matching SOTA accuracy at ∼1/400 of the parameter count (IEEE TRS). *Mamba-Sleep* [54] performs sleep-stage classification from non-EEG wearable signals (heart, motion, temperature) with a pure MAMBA backbone, reaching 84% balanced accuracy – competitive with EEG-based baselines despite the lack of EEG.

Physiological-Signal Classifiers. The physiological-signal sub-family is the densest part of the classification corpus, with several SSL/foundation-scale designs. *EEGMamba* [33] adds a task-aware MoE and a Spatio-Temporal-Adaptive module on top of a bidirectional MAMBA, handling four heterogeneous EEG tasks (seizure, emotion, sleep, motor imagery) on eight datasets under one backbone. *FEMBA* [95] (IEEE EMBC 2025) pretrains a bidirectional MAMBA encoder on 21k h of EEG via masked-patch SSL and ships a 7.8M-parameter edge variant. *EEG-M^2* [36] is the Mamba-2 SSL counterpart to FEMBA, jointly preserving temporal and spectral characteristics via L1+spectral reconstruction. *Mentality* [69] trains a MAMBA-based foundation EEG model for seizure detection (AUROC 0.72) via two-stage SSL pretraining, and *SAMBA-EEG* [37] introduces a Multi-Head Differential Mamba with spatial-adaptive embedding and temporal semantic random masking, beating SOTA on 13 EEG datasets across heterogeneous electrode layouts.

In sleep staging, *MSSC-BiMamba* [122] pairs Efficient Channel Attention with bidirectional MAMBA on ISRUC and Sleep-EDF; *BiT-MamSleep* [129] combines a Triple-Resolution CNN with bidirectional Mamba for short- and long-term EEG dependencies; and *NeuroNet* [46] is a hybrid SSL framework (contrastive + masked prediction) with a Mamba temporal-context module. In ECG classification, *ECGMamba* [78] swaps self-attention for a bidirectional state-space block tailored to ECG morphology; *S^{2}M^{2}ECG* [124] is a multi-branch bidirectional SSM with signal/temporal/spatial fusion for multi-lead ECG; *1D-CNN-ECG-Mamba* [39] is a 1D-CNN+Mamba hybrid for multilabel abnormality detection on PhysioNet 2020/2021; and *MambaCapsule* [114] pairs a Mamba feature extractor with a capsule routing head, hitting >99.5% accuracy on MIT-BIH and PTB.

What Selectivity Contributes. Classification's non-causality favors bidirectional scans across the board, and the long windows in physiological-signal benchmarks (EEG, ECG, sleep) reward the

 doi:10.20944/preprints202605.0995.v1

linear-cost recurrence in ways that short-window UCR benchmarks less obviously do. The corpus's gravitational center is therefore the physiological-signal sub-family, where the per-paper architectural recipes converge: patch tokenization, bidirectional scan, a small CNN or attention front end, and either a supervised head or a self-supervised masked-reconstruction pretrain. Activity-recognition and physiological-signal subfields are where MAMBA classifiers are most established; the broader UEA archive is competitive but not dominated, with TSCMamba and FAIM the strongest entries.

4.5. Unified Multi-Task Analytics

Problem Setting and Selectivity Hypothesis. A growing strand targets *one backbone, many tasks*: a single MAMBA trained jointly or sequentially to forecast, detect, impute, and classify, with task-specific heads on a shared encoder. Operational deployments (industrial monitoring, hospital ICUs, financial desks) consume one multivariate stream for all four analytics, making four separate models impractical. TimesNet [104] showed a single backbone *can* serve all four; the question selective SSMs raise is whether linear-time scaling and input-dependent gating better support joint training across heterogeneous objectives.

The hypothesis: *a selective recurrence amortizes its long-context benefit over four heads instead of one, making the operational case materially stronger than for any single-task variant.* It is preliminary; the corpus is small and benchmarks lack standardization.

MAMBA Multi-Task Models. Table 8 summarizes the methods. The corpus splits into *generalist* multi-task models (Chimera, MambaMixer, SiMBA, DeMa) and *domain-specialist foundation models* (EHRMamba, ss-Mamba) that target a single domain but address multiple analytics within it.

Table 8. Mamba-based unified multi-task / foundation analytics models. **Tasks** lists the TSA tasks the single backbone is demonstrated to handle: F = forecasting, AD = anomaly detection, Imp = imputation, Cls = classification.

Method	Tasks	Token.	Channel	Hybrid.
Chimera [10]	F, AD, Cls	2D grid	adaptive	–
MambaMixer [11]	F, vision	patch	CD	MLP
SiMBA [71]	F, vision	patch	CC	EinFFT
DeMa [6]	F, AD, Cls, Imp	ch.-token	CD	delay-aware attn
EHRMamba [25]	EHR multi-task	event-token	CD	prompted FT
ss-Mamba [119]	F (foundation)	semantic+spline	CD	KAN + PLM

Chimera [10] (NeurIPS 2024) extends MAMBA to a 2-D selective scan over a (channel $\times$ time) patch grid, handling forecasting, ECG/speech classification, and anomaly detection with only the head changing per task – the clearest existence proof that a selective SSM backbone can unify three canonical TSA tasks (imputation is not in the original benchmark suite). *MambaMixer* [11] instantiates a dual token-and-channel selective backbone as both a vision encoder (ViM2) and a multivariate time-series encoder (TSM2), showing the same selective-mixer recipe hosts classification, forecasting, and anomaly heads across modalities. *SiMBA* [71] pairs MAMBA with an EinFFT spectral channel mixer and reports SOTA on *both* ImageNet and seven multivariate time-series benchmarks under one architecture. *DeMa* [6] decomposes multivariate series into intra- and inter-series paths with a dedicated MAMBA per path and a delay-aware mixing mechanism, demonstrating competitive numbers across forecasting, anomaly detection, classification, and imputation.

Domain-Specialist Foundation Models. *EHRMamba* [25] is the first MAMBA-based EHR foundation model, evaluated on six MIMIC-IV clinical prediction tasks with 4$\times$ longer context than Transformer baselines; Multitask Prompted Finetuning shares the backbone across tasks while HL7 FHIR encoding standardizes the input. *ss-Mamba* [119] integrates semantic embeddings (from PLMs) and an adaptive spline temporal encoding inside the selective SSM, framing itself as a foundation forecasting model for transferable features.

What Selectivity Contributes. Three task-specific benefits emerge from the corpus. First, the linear-time scaling lets the multi-task model see the same long window across all four heads, removing

 doi:10.20944/preprints202605.0995.v1

the length-truncation tradeoffs that single-task attention models often make per task. Second, the input-dependent gating provides a *native* mechanism for the model to weight context differently per task: forecasting weights recent regime evidence, anomaly detection weights tail evidence, imputation weights boundary evidence, classification weights label-discriminative evidence. Third, the recurrent formulation naturally supports streaming inference, where task heads can be invoked at different frequencies (forecast per minute, anomaly per second, classify per hour) without re-running the encoder.

5. Data Perspective

The data perspective re-tags the same corpus (section 4) by the *shape and statistics of the input series*. Five regimes recur across the surveyed methods, each selecting a different default along the design axes (subsection 2.4) regardless of which task the model targets. Table 10 already shows that channel strategy varies more by data shape than by task; this section makes the data-conditional design defaults explicit.

5.1. Univariate Time Series

A univariate series (C=1) reduces $X \in \mathbb{R}^{L\times C}$ to a single sequence. Most surveyed MAMBA methods target the multivariate setting and only report on the univariate slice in passing; the cleanest purely univariate entries are MambaStock [89] (daily closing prices) and the single-sensor activity entries (HARMamba [50], Mamba-Sleep [54]). Univariate inputs trivialize channel strategy (CI only); the design space collapses to tokenization, directional scan, hybridization, and decomposition. MAMBA's advantage is sharpest with long informative lookbacks (e.g., daily prices with multi-quarter regime shifts) and weakest on UCR-style short windows where attention already excels.

5.2. Multivariate Time Series

The multivariate panel $X \in \mathbb{R}^{L\times C}$ is the dominant data shape in the surveyed corpus. The *channel strategy* introduced in subsection 2.1 bifurcates the design space into three families.

Channel-Independent (CI). Apply the same backbone independently to each of the C channels, sharing parameters but not state. CI is the natural default on small-C panels with weak cross-channel coupling (ETT subsets, Exchange, Mamba-Sleep [54], RadMamba [107]). DTMamba [109] is the canonical CI MAMBA forecaster.

Channel-Dependent (CD). Mix across channels via dense projection, attention, or learnable routing. CD is the default on high-C panels with stable cross-channel structure (Electricity, C=321; Traffic, C=862; PEMS, C up to 883). S-Mamba [101] tokenizes each channel and sweeps with a single MAMBA block; iTransformer-style channel-as-token tokenization is in this family.

Channel-Correlated (CC). A middle ground that exploits a known or learned correlation structure: group mixing (CMamba [121]), bidirectional channel scans (Bi-Mamba+ [52], with a per-dataset CI/CD switch), or paired token+channel selective scans (MambaMixer [11], Chimera [10]'s 2D scan). *Across forecasting, anomaly detection, imputation, and multi-task analytics, the strongest published numbers on high-C panels invariably come from CD or CC variants*; pure CI backbones lose the channel-coupling signal that the panel offers.

5.3. Spatio-Temporal Graphs and Trajectories

When the channel dimension carries a spatial structure – road sensors, electricity feeders, brain electrodes, financial cross-sections – the data shape is best modeled as a *spatio-temporal graph* (STG): a node-channel adjacency matrix $A \in \mathbb{R}^{C\times C}$ alongside the temporal panel $X \in \mathbb{R}^{L\times C}$. Three architectural templates dominate.

Factorized Scans. Two MAMBA scans, one over the spatial axis (channel-as-token) and one over the temporal axis, fused at the output. ST-MambaSync [88] pairs a temporal MAMBA with a spatial ST-Transformer; HSTM [116] pairs a cross-section MAMBA with a per-stock temporal MAMBA.

 doi:10.20944/preprints202605.0995.v1

2D Selective Scans. Chimera [10] extends the recurrence to a 2D scan over a (channel × time) patch grid, capturing local spatio-temporal interactions in a single sweep at linear cost.

Graph-walk Scans. Pre-compute multiple random walks over the sensor graph and feed each walk as a separate sequence to a shared MAMBA encoder. SpoT-Mamba [21] uses this construction for traffic; Graph-Mamba [62] adapts it to financial cross-sections with an adaptive-graph-convolution front end. BiG-Mamba [131] runs a bidirectional scan over a learned inter-channel graph for traffic-style series where the adjacency matrix is informative but loosely grid-aligned.

A spatio-temporal trajectory – e.g., an individual taxi or sensor's path through space – is the multivariate-trajectory specialization of an STG and is handled in the corpus through the same factorized-scan templates.

5.4. Irregular and Long-Context Data

Irregular-time series – clinical EHRs, longitudinal lab results, event-driven IoT logs, EEG with missingness – combine three properties that defeat self-attention: variable inter-arrival times, long horizons (often $L \geq 10{,}000$ steps), and structured missingness across heterogeneous channels. The HiPPO/S4 lineage [31,32] was designed to absorb the irregularity that defeats positional attention, and MAMBA inherits this capability through the input-dependent Δ_t that can in principle be tied to the observed inter-event gap. The corpus is still under-developed in this regime: HyMaTE [65] (subsection 6.1) is the most complete clinical entry but discretizes onto a fixed grid before applying MAMBA; EHRMamba [25] encodes patient records via Hierarchical FHIR tokenization. subsection 8.3 treats the irregular-time setting as the clearest open opportunity in the survey.

Long-context Dense Panels. A separate but related regime is the *long-context dense* setting where $L \geq 1024$ but the sampling is regular: long-window EEG, sleep staging, multi-day electricity load. Here MAMBA's linear-cost recurrence is purely a compute advantage over attention; the architectural choices are otherwise identical to those of the multivariate-CD regime.

6. Application Perspective

The application perspective re-tags the surveyed corpus (section 4) by *operational domain*. The strongest performers per domain are rarely the most novel backbones; they are the simplest design that respects the dominant data structure – channel coupling for energy, spatial factorization for traffic, hybrids for climate, linear-cost backbones for finance, channel-mixing for clinical streams. Table 9 consolidates the methods, cross-referenced to the per-task chapter that introduced each. Cross-domain shift, pretraining, and zero-shot transfer are deployment regimes orthogonal to domain and are deferred to subsection 8.8.

Table 9. Domain-specialized and foundation MAMBA time-series models. Most listed methods use MAMBA as the SSM primitive; non-MAMBA cases are named in the backbone column (Mamba-2 = SSD variant of Dao and Gu).

Method	Area	Backbone	Channel
PowerMamba [63]	energy	dual-Mamba	CC
BiG-Mamba [131]	spatio-temp.	bi-Mamba+graph	CC
ST-MambaSync [88]	spatio-temp.	bi-Mamba+ST-Trans	CC
SpoT-Mamba [21]	spatio-temp.	Mamba+graph walks	CC
Chimera [10]	spatio-temp.	2D SSM scan	CD
AFFiRM [108]	climate	Mamba+FFT	CD
MAT [127]	climate	Mamba+Attn.	CD
HSTM [116]	finance/spatial	spatial+temp.	CC
CMDMamba [81]	finance	Mamba+CNN	CD
MambaStock [89]	finance	pure Mamba	CI
Graph-Mamba [62]	finance/graph	bi-Mamba+GCN	CC
HyMaTE [65]	healthcare/EHR	Mamba+Transf.	CC
DGMamba [60]	domain gen.	Mamba	CI
Mamba4Cast [13]	zero-shot	Mamba-2	CI
SpaceTime [126]	foundation	S4 (pre-Mamba)	CI

 doi:10.20944/preprints202605.0995.v1

6.1. Healthcare and Clinical Monitoring

Clinical data – ICU vitals, wearable streams, longitudinal EHRs, and physiological signals (EEG, ECG, sleep) – are typically long, irregularly sampled, and dominated by missingness across a wide channel dimension of medical codes, lab tests, and vital signs. MAMBA's linear-time inference and constant memory fit long horizons, and selectivity in principle accommodates event-driven update rates; the principal limitation is that the sequence-only scan does not directly mix information across channels.

Clinical EHRs and ICU Monitoring. HyMaTE [65] is the most developed example in the corpus: it encodes each patient's record with MAMBA blocks and adds a Transformer block to mix across the medical-code dimension, both addressing the channel-mixing limitation and yielding a more interpretable representation. EHRMamba [25] is a MAMBA-based EHR foundation model trained with HL7 FHIR encoding and Multitask Prompted Finetuning across six clinical prediction tasks on MIMIC-IV with 4× longer context than Transformer baselines.

Physiological Signals (EEG, ECG, Sleep). The physiological-signal classification cluster is the densest in the corpus and converges on a recipe of patch tokenization, bidirectional scan, a small CNN or attention front end, and a self-supervised masked-reconstruction pretrain. The strongest entries are EEGMamba [33] (task-aware MoE on bidirectional MAMBA), FEMBA [95] (21k-hour pretrain, edge variant), EEG-M^2 [36] (Mamba-2 SSL with spectral preservation), Mentality [69] (foundation EEG for seizure detection), SAMBA-EEG [37] (multi-head differential Mamba across heterogeneous electrode configurations), and the sleep-staging family (MSSC-BiMamba [122], BiT-MamSleep [129], NeuroNet [46]). For ECG, ECGMamba [78], S^2M^2ECG [124], 1D-CNN-ECG-Mamba [39], and Mamba-Capsule [114] repeat the bidirectional+CNN template.

6.2. Energy and Electricity

Energy and load forecasting is the historic heartland of TSF and dominates the Electricity, Solar, and ETT benchmarks. The data exhibit pronounced and *stable* cross-channel structure (correlations among substations or feeders) alongside strong daily and weekly seasonality. These properties favor channel-correlation designs: CMamba [121] and MambaMixer [11] report the largest gains in this regime, while pure channel-independent backbones remain competitive on the smaller ETT subsets where channel coupling is weak. PowerMamba [63] is the first explicitly energy-specialized design: it pairs a dual-Mamba encoder with an external-predictor fusion head that consumes system-operator forecasts (renewable generation, load) alongside the raw series, and is released together with GridSet, a five-year ERCOT benchmark (262 channels covering load, net load, generation by fuel, locational marginal prices, and pre-published renewable forecasts). PowerMamba reports roughly 7% lower mean error than TimeMachine at 43% the parameter count on GridSet, and a 76% improvement on renewable forecasting when external predictions are used.

6.3. Traffic and Spatio-Temporal Mobility

Traffic and spatio-temporal benchmarks (Traffic, PEMS04, PEMS08) combine extreme channel counts (up to 862 sensors) with well-defined spatial structure that can be exploited as an auxiliary inductive bias. The dominant design pattern factorizes the problem across two MAMBA scans – one over the spatial axis (channel-as-token) and one over time – mirroring classical spatio-temporal Transformers at linear cost. Chimera [10] realizes this with a 2D selective scan over a (channel × time) patch grid, capturing local spatio-temporal interactions in a single sweep. BiG-Mamba [131] replaces the spatial axis with a bidirectional scan over a learned inter-channel graph, well-suited to traffic-style series where the sensor adjacency matrix is informative but only loosely grid-aligned. SpoT-Mamba [21] precomputes multiple random walks over the sensor graph and feeds each walk as a separate sequence to a shared MAMBA encoder, with a learned per-walk weighting that targets long-range dependencies that two- or three-hop GCN layers miss. ST-MambaSync [88] takes the complementary route of pairing a MAMBA block with an ST-Transformer attention block in a lightweight fusion that matches

 doi:10.20944/preprints202605.0995.v1

STAEformer on METR-LA / PEMS-BAY / PEMS03–08 at lower compute. On the imputation side, TIMBA [91] and MAC2STI [26] repeat the spatio-temporal template with diffusion and cluster-aware variants of this factorization.

6.4. Climate and Weather

Climate and weather forecasting (Weather and related earth-science benchmarks) features strong seasonal structure, multi-scale dynamics, and long-range temporal dependencies. Hybrids dominate: Mamba+Attention designs like MAT [127] let attention capture non-stationary long-range effects while MAMBA absorbs local context, and Fourier-augmented variants like AFFiRM [108] factor out periodic components. We are not aware of MAMBA models tailored to satellite imagery or atmospheric reanalysis – an open opportunity (subsection 8.7).

6.5. Finance

Financial series exhibit heavy-tailed returns, regime non-stationarity, and event-driven jumps that defy standard preprocessing; modest dataset sizes also penalize parameter-heavy backbones, making the linear-time MAMBA recurrence attractive on compute and sample-efficiency grounds. MambaStock [89] applies a stripped-down MAMBA backbone to daily stock-price prediction at order-of-magnitude lower compute than the strongest Transformer baselines. CMDMamba [81] adds a dual-layer convolutional feed-forward around each MAMBA state update, capturing the short bursts of volatility that punctuate otherwise slow-moving price series. HSTM [116] couples a spatial MAMBA over the cross-section of stocks with a temporal MAMBA over each stock's history, transferring the spatio-temporal template of subsection 6.3 to a setting where the "adjacency" graph is sector or factor membership rather than a road network. Graph-Mamba [62] attacks the same cross-sectional problem with a bidirectional MAMBA scan and an adaptive graph convolution that learns inter-stock dependencies from data rather than imposing a fixed sector taxonomy, beating Transformer baselines for next-day price prediction at near-linear cost.

6.6. Activity Recognition and Sensors

Activity recognition on wearable, radar, and ambient-sensor streams is dominated by short-window, multivariate, edge-deployed inference where parameter count and latency matter as much as accuracy. HARMamba [50] targets wearable IMU streams with a patch-tokenized, channel-independent bidirectional MAMBA backbone, delivering competitive accuracy at edge-device parameter counts on PAMAP2, WISDM, UNIMIB-SHAR, and UCI-HAR. RadMamba [107] extends HAR to radar micro-Doppler signals with a Doppler-aligned segmentation tokenizer, achieving accuracy parity with state-of-the-art at $\sim$1/400 of the parameter count on at least one benchmark.

6.7. Cross-Domain and Foundation-Scale Deployments

A final application-cluster targets cross-domain serving and foundation-scale pretraining. The selectivity hypothesis here is about throughput: linear-cost recurrence is the only operational way to push pretraining sequence length past the few-thousand-token budget that is feasible for attention. SpaceTime [126] is a pre-Mamba SSM-based forecaster that anticipates several MAMBA design patterns and serves as a reminder that the foundation-model regime predates MAMBA itself. Mamba4Cast [13] extends this paradigm to zero-shot transfer, pretraining a MAMBA backbone on a large synthetic dataset and demonstrating strong zero-shot performance across the GIFT-Eval benchmark suite; the result shows that the linear-time backbone is competitive with pretrained Transformer foundation models (TimesFM, Chronos) at substantially lower inference cost. ss-Mamba [119] integrates semantic embeddings (from PLMs) with adaptive spline temporal encoding inside MAMBA's selective SSM, framing itself as a foundation forecasting model for transferable representations. On the cross-domain robustness side, DGMamba [60] adds a domain-invariant feature-extraction loss to address cross-region shift; it is backbone-agnostic and complements any channel strategy.

 doi:10.20944/preprints202605.0995.v1

Trade-off summary. Specialized designs trade generality for alignment with a domain. Across the seven application areas, the strongest performers are the simplest design that respects the dominant statistical structure of the data: channel coupling for energy, spatial factorization for traffic, hybrids for climate, linear-cost backbones for finance, channel-mixing augmentations for clinical streams, and edge-friendly parameter counts for activity recognition. The underdeveloped healthcare and earth-science settings, together with probabilistic decoder heads for foundation-scale MAMBA, are the clearest opportunities for application-driven contributions.

7. Practical Guidelines

A practitioner facing a fresh TSA problem needs more than a catalog: a rule for picking a variant given the *task and data*, a starting configuration, and awareness of failure modes that silently distort comparisons. This chapter synthesizes the corpus into four artifacts: subsection 7.1 crystallizes the per-task chapter findings into a cross-task design-axis matrix that prunes the search space; subsection 7.2 turns the taxonomy into a per-task decision guide keyed on data properties; subsection 7.3 summarizes configuration and training defaults recurring across top-performing papers; and subsection 7.4 documents MAMBA-specific pitfalls that confound reproducibility and do not arise (or arise differently) for Transformer or linear baselines.

7.1. Cross-Task Design-Axis Matrix

Reading the five task subsections (subsection 4.1–subsection 4.5) side by side surfaces what is *shared* and what *specializes*. All five tasks consume the same backbone, the same design axes (subsection 2.4), and largely the same pipeline – RevIN normalization [43], AdamW, the fused selective-scan kernel, patch- or channel-tokenized inputs – with the recurrent state $\mathbf{h}_t \in \mathbb{R}^N$ serving all four heads without per-task tweaks; Table 10 summarizes the resulting specialization.

Table 10. Cross-task design-axis specialization matrix. Entries summarize the per-task chapters (subsection 4.1–subsection 4.5), recording the modal choices across the surveyed corpus rather than unique outliers.

Design Axis	Forecasting	Anomaly Det.	Imputation	Classification	Multi-Task
Tokenization	patch / channel-token	patch	patch	patch / window	2D patch grid
Channel strategy	CI *or* CD (contested)	CD (multi-channel)	CD (graph / full)	CI / CD by modality	CD (token+channel)
Directional scan	forward *or* bidir.	bidirectional	bidirectional	bidirectional	2D / dual-axis
Hybridization	attn. / MLP / FFT	+ attn. (discrepancy)	+ diffusion	+ spectral / SSL	none-or-mild
Decomposition	trend-seas. / wavelet	none	none	none / spectral	none
Modal arch. pattern	pure / bidir / hybrid	hybrid	hybrid (diffusion)	bi-directional	2D / token+channel

Directional scan splits along causality. Only forecasting has a meaningful directional debate; the other four tasks default to bidirectional under non-causal objectives. The forecasting debate is unsettled: forward scans (S-Mamba, MambaTS) and bidirectional designs (Bi-Mamba, Bi-Mamba+, ms-Mamba) report comparable accuracy under different protocols.

Hybridization is task-conditional. Anomaly detection favors attention hybrids (MAAT, KambaAD), imputation favors diffusion hybrids (SSD-TS, TIMBA, SCMDI), classification favors spectral or self-supervised front ends (FAIM, FEMBA, EEG-M^2), and multi-task analytics favors none-or-mild hybrids to preserve a clean shared backbone (Chimera, MambaMixer, SiMBA, DeMa).

Channel strategy tracks data shape and domain, not task. Physiological signals (EEG, ECG, sleep, ICU) favor graph-aware or self-supervised channel handling; energy and traffic favor explicit cross-channel attention or correlation-aware mixing; finance favors channel-independent designs with a per-stock graph overlay.

Where the Corpus's Gravity Lies. Forecasting and physiological-signal classification are the most mature sub-corpora; anomaly detection, imputation, and unified multi-task analytics remain in early

 doi:10.20944/preprints202605.0995.v1

establishment, with two- to four-paper sub-corpora and unstable benchmarks. Chimera [10] and DeMa [6] establish cross-task unification under one MAMBA backbone as feasible at small scale; subsection 8.9 takes up what remains.

7.2. Choosing a Right MAMBA Variant

After subsection 7.1 prunes by task, the remaining choice is keyed on data shape (section 5) and, when relevant, application domain (section 6). We consolidate this into a three-step rule.

Step 1: Pick the Per-Task Default. Each task ships a default architectural family (Table 10). *Forecasting* – the only task with a contested directional choice – starts from a bidirectional or forward-only scan, with hybrids added only when Step 3 justifies. *Anomaly detection* starts from a MAAT-style attention-augmented bidirectional reconstructor; *imputation* from an SSD-TS-style MAMBA +diffusion denoiser; *classification* from a TSCMamba/FAIM-style multi-view bidirectional scan; *multi-task analytics* from a Chimera-style 2D scan over the (channel $\times$ time) grid.

Step 2: Refine by Data Shape. Data shape predicts the design axes more reliably than task does. For *small C with weak cross-channel coupling* (ETT, Exchange, small Weather slices), use channel-independent backbones (TimeMachine [2], MambaTS [14], DTMamba [109]); the linear-in-L cost makes $L \geq 1024$ tractable, where MAMBA's advantage over attention first appears. For *large C with stable cross-channel structure* (Electricity C=321, Traffic C=862, PEMS), switch to channel-as-token scans (S-Mamba [101]) or dual token+channel selection (MambaMixer [11], CMamba [121]), which consistently beat channel-independent MAMBA at high C [11,101]. For *spatio-temporal graphs*, choose between factorized (ST-MambaSync [88]), 2D (Chimera [10]), or graph-walk (SpoT-Mamba [21]) scans. For *irregular or long-context* data (clinical EHRs, long EEG), no default exists yet – subsection 5.4 flags this as the clearest open opportunity. Tokenization is length-driven: long L favors patch, short L favors pointwise.

Step 3: Add a Hybrid Front-End When the Data Justifies It. Two cues recommend a hybrid over the per-task default. With clean *trend/seasonal decomposition* (energy load, retail) or *narrow frequency bands* (climate, finance), explicit-signal hybrids – KARMA [118], AFFiRM [108] – are more sample-efficient than raw scans; if a DLinear- or FEDformer-style decomposition is competitive alone, the corresponding MAMBA hybrid is likely the strongest variant. With *short look-back* ($L \leq 192$), Transformer hybrids (MAT [127], SST [113], MambaFormer [70]) routing short-range fluctuations through attention dominate pure-MAMBA baselines; below L=96 the selective-scan advantage is small and replacing a linear baseline (DLinear, TSMixer) is hard to justify.

Domain and Deployment Overrides. For domain-tied problems (healthcare, energy, traffic, climate, finance, activity recognition, foundation-scale serving), start from section 6 and Table 9: domain-respecting designs typically outperform general-purpose MAMBA variants regardless of the Step 2/3 branch.

7.3. Configuration and Training Recipes

We report defaults that recur across top-performing MAMBA-TSF papers; per-dataset tuning within these ranges typically closes the gap to published numbers.

Lookback Length. Use $L \in \{512, 1024\}$ on datasets where long history helps (Electricity, Traffic, Weather, Solar, PEMS). The L=96 default understates MAMBA's advantage and should not be the sole comparison; gains saturate around L=1024 [2,101].

MAMBA Hyperparameters. The original defaults N=16, E=2 [30] port well and are used unchanged in CMamba [121] and TimeMachine [2]. For MAMBA-2 [22], head dimension $d_h \in \{64, 128\}$ with $N \geq 64$ is standard. $E > 2$ rarely justifies its cost.

Tokenization and Patching. Patch embedding (P=16, stride 8; CMamba, PatchTST hybrids) halves scan length and stabilizes short-L training; variate tokenization (S-Mamba, iTransformer-style) treats each channel as a token. They are not interchangeable: variate tokenization is required for high-C channel-mixing gains, patch wins on low-C long-L.

 doi:10.20944/preprints202605.0995.v1

Normalization. Instance normalization (RevIN [43]) is standard and should not be silently disabled: matched audits show both MAMBA and Transformer baselines degrade without it, and the relative ordering can flip [101]. LayerNorm (RMSNorm/GroupNorm for MAMBA-2) is standard inside the block.

Optimizer and Training Budget. AdamW with lr $\in [1e{-}4, 1e{-}3]$ and weight decay $\in [0, 0.05]$ is near-universal. Budgets are modest (10–30 epochs with early stopping), and selective-SSM variants converge faster than equivalent Transformers. Loss should match the reported metric (MAE in CMamba; MSE in S-Mamba, TimeMachine).

Precision and Kernel. Use BF16 with an FP32 accumulator; FP16 on long scans is unreliable. Speed and memory claims should name the kernel: the fused CUDA kernel is 2–8× faster than PyTorch reference [22], and FlashAttention-2 comparisons are like-for-like only when both sides use optimized kernels. Below ∼2K tokens, FlashAttention-2 is faster [22].

7.4. MAMBA*-Specific Pitfalls*

We collect confounders that distort MAMBA-TSF comparisons but do not arise (or arise differently) for Transformer and linear baselines, in three clusters: implementation numerics, architectural attribution, and evaluation protocol.

Implementation Numerics. *Scan-kernel fidelity* matters: the MAMBA-2 SSD kernel is 2–8× faster than fused MAMBA-1 and 10–30× faster than the PyTorch reference [22], so pairing a reference MAMBA against a FlashAttention-2 Transformer is not like-for-like, and kernel choice also affects numerical outputs. *Precision* should be BF16 with FP32 accumulators; FP16 on long scans is unreliable, yet many papers omit precision entirely, blocking number-matching even with released code. Δ_t *and state defaults* – the softplus parameterization, bias init, and min-Δ clamp [22,30] – materially shift training dynamics but are rarely reported; N=16, E=2 are inherited from language modeling and rarely ablated, though CMamba [121] reports non-negligible N sensitivity on channel-correlation benchmarks.

Architectural Attribution. *Patching and decomposition* are confounded with the backbone: many "MAMBA vs. Transformer" tables compare P=16 MAMBA against P=1 or P=8 baselines, and decomposition front-ends (KARMA [118], AFFiRM [108]) explain most of the gain on strongly periodic datasets, yet PatchTST-matched patching [66] and decomposition ablations are rarely reported as controls. *Token axis and scan direction* are similarly conflated: channel-as-token (S-Mamba) and time-as-token (PatchTST) produce different numbers on the same data with the same block, and forward-only and bidirectional scans likewise differ by several percent – both folded silently into "MAMBA" in baseline tables despite the original block being unidirectional [30], so papers whose contribution *is* bidirectionality [52,101] should be matched against a unidirectional MAMBA of equal width/depth. *Channel mixer* choice also matters: MambaMixer's bidirectional channel mixer [11] and CMamba's GDD-MLP [121] both degrade materially when swapped for MLPs, so a *selective* cross-channel scan should not be grouped with MLP channel mixing in ablations.

Evaluation Protocol. *Short-look-back reporting* masks the principal advantage: the community L=96 default does not exercise MAMBA's linear-cost long-range edge, so tables claiming a MAMBA-specific gain should include at least one long-L column $L \in \{336, 512, 720, 1024\}$, since relative ordering flips non-monotonically with L [101]. *Baseline drift and missing length generalization* compound this: same-named baselines (PatchTST, DLinear, iTransformer) often carry different numbers across papers, copied without re-running – CMamba [121] is an exception – and although MAMBA's fixed-size state should permit $L_{\text{test}} > L_{\text{train}}$, papers train and evaluate at a single L, leaving the principal theoretical advantage unsubstantiated. *Bidirectional baselines for causal tasks* are ambiguous: time-axis bi-MAMBA either restricts to the look-back (valid) or leaks horizon information (invalid), and the literature uses "bi-MAMBA" for both, so readers should verify the bi-direction is along the channel axis or strictly inside the look-back. *AD threshold and redundant clusters* also distort rankings: anomaly-detection F_1 depends on calibration (peaks-over-threshold, point-adjustment, F_1-best), point-adjustment can inflate F_1 by tenths and should be reported alongside raw F_1, and dense sub-corpora such as foundation

 doi:10.20944/preprints202605.0995.v1

EEG [33,36,37,69,95] lack a fixed-protocol side-by-side comparison, so cross-paper rankings within such clusters should be treated cautiously.

8. Open Frontiers and Future Directions

Despite rapid expansion of the MAMBA-TSA corpus across all five tasks, foundational questions remain open. We organize this chapter around twelve frontiers, each pairing an unresolved tension with a concrete research program executable on existing infrastructure: the first eight implicate forecasting and the non-forecasting tasks alike, the ninth (cross-task unification) is new to this broadened scope, and three additional frontiers concern theoretical expressivity, compression for edge deployment, and integration with post-MAMBA primitives.

8.1. Attributing Gains to Selectivity

MAMBA's novelty over S4/S5 is the input-dependent triple (Δ_t, B_t, C_t) [30], yet two findings undercut selectivity as the *load-bearing* component. Matched-protocol audits [101,120] show MAMBA variants rarely beat DLinear or PatchTST under a shared budget, and after RevIN [43] the dominant benchmarks are nearly stationary at the instance level – precisely where a linear time-invariant SSM (S5) should already be near-optimal; Vision-Mamba reviews [57] flag the same tension. *Actionable direction:* freeze (Δ, B, C) to their input-independent counterparts – reverting MAMBA to S5 at matched width, depth, and kernel. If the gap is within tuning noise, MAMBA's wins re-attribute to patching, tokenization, and hardware-aware training; if the gap appears only on regime-shift or event-dense series, these (not ETT) should become the canonical benchmarks.

8.2. Input-Dependent Channel Selectivity

Although subsection 2.4 recognizes channel-independent, -dependent, and -correlated strategies, every surveyed method fixes the channel choice *statically* at design time – incongruous with MAMBA's core abstraction of input-dependent interaction. The CI-vs-CD debate is then not a fundamental dichotomy but an artifact of applying selectivity only along the time axis. Bi-Mamba+ [52] learns a per-dataset CI/CD switch and MambaMixer [11] runs two fixed scans in parallel, but neither offers input-dependent channel gating. *Actionable direction:* define a *selective channel scan* whose gate, analogous to Δ_t, is a learned function of channel statistics, subsuming Bi-Mamba+ and MambaMixer as special cases. Its natural evaluation ground is Electricity, Traffic, and PEMS, where C is large and correlations are unstable over time.

8.3. Native Irregular-Time Modeling

Every surveyed method assumes a regular grid, although Δ_t in the selective recurrence (Equation 9) was originally designed in HiPPO/S4 [31,32] to absorb the very irregularity that defeats self-attention. Clinical streams, industrial IoT, and event logs all have variable inter-arrival times, yet section 6 documents only scattered pilots; even HyMaTE [65] discretizes onto a fixed grid before applying MAMBA. SSD-TS-family imputers [27,91] repeat the pattern, evaluating on densely-sampled panels with synthetic uniform masks – so the regime where MAMBA has a *structural* (not merely computational) advantage has no dedicated MAMBA entry. *Actionable direction:* instantiate a MAMBA block in which Δ_t is the observed inter-event gap, and evaluate on MIMIC-III, PhysioNet 2012, and wearable-sensor benchmarks against GRU-D and Neural-ODE baselines – a market attention cannot serve natively.

8.4. Probabilistic Filtering Decoders

Up to notation, a selective SSM is a time-varying linear-Gaussian filter – structurally the natural home for distributional inference. Yet the sole probabilistic MAMBA forecaster, Mamba+UQ [77], obtains uncertainty via Monte Carlo dropout, ignoring the filtering structure; SSD-TS-family imputers [27,91] ship a diffusion posterior with no MAMBA-specific calibration recipe. Transformer foundations (Chronos, TimesFM) lack filtering yet ship distributional heads, leaving MAMBA-TSA

 doi:10.20944/preprints202605.0995.v1

paradoxical: the backbone best equipped for probabilistic inference has invested least in it. *Actionable direction:* derive closed-form predictive variance by treating $(\bar{A}_t, \bar{B}_t, C_t)$ as a time-varying Kalman system and propagating an observation-noise term through the scan. Compare calibration (CRPS, interval coverage, reliability) against MC-dropout Mamba+UQ, conformal wrappers, and distributional Transformer baselines. A positive result would make selective SSMs the default probabilistic backbone – an impact comparable to DeepAR for RNNs.

8.5. Test-Time Length Generalization

MAMBA's bounded hidden state ($\mathbf{h}_t \in \mathbb{R}^N$) is cited as the enabler of unbounded context: unlike positional-encoded Transformers, a selective scan should extrapolate to arbitrary test-time lengths. In practice, *no paper in the corpus reports* $L_{\text{test}} > L_{\text{train}}$ (subsection 7.4), and accuracy reportedly degrades sharply when attempted. The non-forecasting tasks share the gap: MAMBA anomaly detectors are published only at $L \leq 512$, and UCR/UEA classification archives provide no long-window splits, so the long-context advantage is demonstrated only on physiology-specific (EEG, sleep) data. The promise-vs-behavior gap is a credibility risk: if bounded-state recurrence cannot outperform attention at out-of-window lengths, the single theoretical property motivating the family is undercut. *Actionable direction:* publish a length-extrapolation benchmark (train at L=512, evaluate at $L \in \{1024, 2048, 4096, 8192\}$, report degradation curves) paired with training-side fixes from vision-Mamba: Δ_t-annealing, state-preservation curricula, and continuous-time reparameterization, with fixes transferring to MAMBA outside TSF.

8.6. Three-Factor Gain Attribution

subsection 7.4 shows that patching alone can shift MSE by several percent, and that hybrids like KARMA [118], Affirm [108], and DMSTCI-BiMamba [56] insert an STL/Fourier decomposition *before* the scan – yet the literature reports the *sum* of block, tokenization, and front-end gains as "MAMBA-based forecasting." *Actionable direction:* adopt a three-factor ablation across patch size $P \in \{1, 8, 16\}$, decomposition on/off, and backbone $\in$ {linear, S5, MAMBA, MAMBA-2, Transformer} at matched look-back and normalization. The MAMBA-2 axis is not cosmetic: only Mamba4Cast [13] adopts the SSD primitive directly and Chimera [10] generalizes it (2D time–channel scan); every other surveyed forecaster remains on S6 despite the faster kernel and larger state available in MAMBA-2 [22]. Early signals from CMamba [121] and S-Mamba [101] suggest cross-terms are large; this grid should be a reporting *standard* before "MAMBA beats X" claims are accepted.

8.7. Benchmark Saturation and the Hybrid Wall

Two observations point in the same direction. *Benchmark saturation:* on ETTh1/ETTh2/ETTm1/ETTm2 the best reported MSE varies by less than TFB's inter-seed noise [83], so new MAMBA variants compete below the noise floor. *Hybrid Pareto wall:* on the channel-rich benchmarks that still discriminate – Electricity, Traffic, PEMS – the strongest numbers invariably come from hybrids (Chimera [10], Bi-Mamba+ [52], SST [113], Affirm [108]); pure selective-SSM backbones have not crossed this frontier despite eighteen months of refinement. Whether the wall is fundamental (a capacity ceiling on rich cross-channel structure) or artifactual (the right block primitive has not been found) is unsettled. *Actionable direction:* retire ETT as a headline benchmark and commission replacements stressing selective-SSM-unique properties – length extrapolation, extreme C, irregular sampling, regime shift, cross-domain transfer – extending TFB [83] and GIFT-Eval [5]. A matched-capacity study pitting pure MAMBA against leading hybrids on these replacements would resolve the question and free the field to either scale the primitive or embrace hybridization as permanent.

8.8. Foundation-Scale Pretraining and Transfer

A second cluster of frontiers is shaped by *deployment constraints* – cross-domain shift, on-device training, and pretraining/zero-shot transfer – where MAMBA's linear-cost recurrence is disproportionately attractive at scale, since pretraining throughput and serving latency are dominated by

 doi:10.20944/preprints202605.0995.v1

sequence-length costs. On the *domain-shift* side, DGMamba [60] adds a domain-invariant feature-extraction loss to address cross-region shift (e.g., training on one road network, evaluating on another), and the construction is backbone-agnostic. On the *pretraining and zero-shot* side, SpaceTime [126] shows the foundation-model regime predates MAMBA, and Mamba4Cast [13] extends it to zero-shot transfer via synthetic-data pretraining, matching Transformer foundations (TimesFM, Chronos) on GIFT-Eval at lower inference cost. *Actionable direction:* (i) scale pretraining to multi-billion tokens of mixed real/synthetic data, (ii) quantify OOD behavior on tail domains, and (iii) add a probabilistic decoder head for calibrated uncertainty. Combined with subsection 8.4, this is the clearest path to a MAMBA foundation model that beats attention foundations at fixed compute.

8.9. One Backbone for All Five Tasks

The five-task scope exposes a frontier forecasting-only treatments cannot see: *whether a single* MAMBA *backbone serves all five TSA tasks at competitive quality*. Chimera [10], MambaMixer [11], SiMBA [71], and DeMa [6] provide small-scale existence proofs (with TimesNet [104] doing so earlier for a non-Mamba backbone). Three issues block unification at scale. *Heads disagree on what to preserve*: forecasting wants high temporal fidelity, classification compact summaries, anomaly detection sharp discrepancy boundaries, imputation smooth interpolants – a single hidden state may not satisfy all four, and head-conditioned selectivity (per-head Δ, B, C over a shared recurrence) is unexplored. *Joint loss balancing* across heterogeneous objectives (regression, density, cross-entropy) is solved in the corpus by ad-hoc re-weighting; principled alternatives (PCGrad, GradNorm) have not been combined with selective-SSM backbones. *A unified benchmark* covering all five tasks under matched protocols is missing; TimesNet's evaluation predates MAMBA and doesn't exercise long contexts. *Actionable direction:* build a MAMBA-TSA bench (paired $(L{=}4096, H \in \{96, 720\})$ forecasting + reconstruction on MIMIC-IV, outcome classification, anomaly detection on residuals) and compare a jointly trained backbone against four specialized backbones at matched FLOPs.

8.10. Expressivity and Identifiable Dynamics

A second-order question, orthogonal to selectivity attribution (subsection 8.1), is what *classes of dynamics* a selective SSM can represent. Real-valued, non-negative eigenvalue SSMs collapse to TC^0 and provably fail simple state-tracking tasks (parity, permutation composition) that LSTMs handle, motivating extensions to complex eigenvalues, RoPE-style rotations, and matrix-valued states (Mamba-3, gated linear attention, DeltaNet). The TSA payoff is open: oscillatory regimes (climate, finance), regime-switching processes, long-memory fractional dynamics, and chaotic systems are exactly where the limitation should bite, yet no surveyed paper reports identifiability against synthetic processes with known generators. *Actionable direction:* build a dynamics-characterization benchmark (AR/ARMA, fractional Brownian motion, Lorenz/Rössler, regime-switching HMMs) and report which families a vanilla MAMBA recovers, which need complex/RoPE eigenvalues, and which need non-diagonal state. A negative result would refocus the field from benchmark sweeping to primitive design.

8.11. Compression and Edge Deployment

MAMBA's linear-cost recurrence is most attractive on edge – wearables, IoT, in-vehicle ECUs, on-device ICU/sleep monitors – yet standard quantization and distillation pipelines, designed for attention, do not transfer cleanly. The non-linear Δ_t discretization, recurrent error accumulation, and mixed-precision accumulator interact in ways that PTQ/QAT recipes for Transformers do not anticipate: emerging SSM-specific accelerators (eMamba, Mamba-X) and sub-2-bit Mamba quantization show 4–8$\times$ memory reduction is achievable but require kernel-level co-design. The TSA literature has produced a handful of edge-targeted variants (edge-tier biosignal foundations [36], sub-10M HAR backbones [50]) but no systematic study of bit-width sensitivity, recurrent quantization error, or distillation from a teacher MAMBA to a smaller student. *Actionable direction:* report INT8/INT4/1.58-bit accuracy and latency on TSA benchmarks (HAR, ECG, PEMS, wearable PPG) at matched FLOPs,

 doi:10.20944/preprints202605.0995.v1

and characterize where recurrent state precision dominates versus where it is slack – essential for the deployment claims of section 6.

*8.12. Post-*MAMBA *and Multimodal Fusion*

Outside TSF, the linear-recurrence frontier has moved quickly – DeltaNet, test-time-training (TTT) layers, gated linear attention, Hawk/Griffin, xLSTM, RWKV-7, and Mamba-3 each propose a different fix to a known limitation of selective SSMs (delta-rule state writes, key-value memory, matrix-valued recurrence). None has been systematically applied to TSA, so it is unclear whether MAMBA is a stable foundation or a transitional primitive. A parallel gap is *multimodal time series*: numeric series arrive paired with text reports (clinical notes, earnings calls), images (sky cameras, satellite, X-rays), and event logs, and recent benchmarks (Time-MMD) report 15%+ MSE reductions from text fusion alone, yet only scattered MAMBA pilots exist. Causal/counterfactual TSA is similarly open: Mamba-CDSP repurposes input-dependent Δ_t for treatment-effect rollout, but the broader question of whether selective SSMs are a natural fit for time-varying confounding has not been answered. *Actionable direction:* a head-to-head between MAMBA, a representative post-MAMBA primitive (DeltaNet/xLSTM/Mamba-3), and an attention baseline on (i) standard forecasting, (ii) text+numeric multimodal forecasting, and (iii) causal-effect estimation under known ground-truth generators would clarify whether the right next step is to scale MAMBA or to replace its primitive.

9. Conclusions

This survey provides the first focused treatment of MAMBA for time series analysis, organized through four complementary perspectives – model architecture, task family, data shape, and application domain – supported by a five-axis design vocabulary, per-pattern comparison tables, a benchmark catalog, and a registry of public implementations. Across these perspectives, linear-time selective recurrence has become a credible alternative to Transformer baselines for long-context time series, with the strongest evidence concentrated in forecasting and matched-protocol audits suggesting the margin on other tasks remains task- and dataset-dependent.

Significant open challenges remain in gain attribution, modeling regimes, and cross-task unification. By consolidating the design axes, comparison tables, practical guidelines, and an online repository at https://github.com/tamlhp/awesome-mamba-ts, this survey offers a starting point for the community to reproduce comparisons, extend the taxonomy, and address these open frontiers.

Use of Artificial Intelligence: During the preparation of this work the author(s) used ChatGPT for English proofreading. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Resources

This section catalogs the empirical substrate of the surveyed corpus: public implementations (subsection A.1), standard datasets and benchmarks (subsection A.2), evaluation metrics (subsection A.3), and the headline performance numbers reported in the source papers (subsection A.4). MAMBA-specific protocol pitfalls are treated separately in subsection 7.4.

Appendix A.1 Published Implementations

Table A11 catalogs the thirty-seven peer-reviewed MAMBA-family forecasters in the surveyed corpus, organized by the taxonomy branch of Figure 4 and marking which ones have publicly released source code. PyTorch is the dominant framework; every released implementation depends – directly

 doi:10.20944/preprints202605.0995.v1

or transitively – on the `mamba_ssm` package [30] for the hardware-aware selective scan, with a pure-PyTorch fallback that runs 5–10× slower. Reproducing the original numbers therefore typically requires a recent CUDA toolchain and a compatible GPU.

Pure MAMBA Backbones. The best-maintained reference implementations are S-Mamba, MambaTS, TimeMachine, and SiMBA with SiMBA-TS, each supplied by the original authors with training scripts for the standard ETT, Electricity, and Weather benchmarks. Mamba+UQ [77] additionally ships calibration code for probabilistic evaluation, a useful starting point for uncertainty-aware extensions.

Bidirectional and Multi-Scale Scans. Bi-Mamba and its Bi-Mamba+ extension share a codebase whose patch-wise scan primitives are reused by several later works. ms-Mamba and Chimera release standalone implementations. Chimera's 2D-scan kernel is the only non-`mamba_ssm` selective scan in the surveyed release landscape.

Hybrid Architectures. MAT, SST, FMamba, MambaFormer, Affirm, and KARMA all release PyTorch implementations; CMDMamba and UmambaTSF release less polished code with partial configuration files. SiMBA and MambaMixer additionally release vision backbones, easing cross-domain reuse.

Domain-specialized and Foundation Systems. MambaStock and DGMamba release training scripts tied to their respective domain datasets. Mamba4Cast is the only released foundation-style MAMBA TSF implementation and includes a zero-shot inference script compatible with GIFT-Eval and Monash [5, 28].

Announced but Not Yet Released. Several papers surveyed in 2025 announce implementations that are pending release (DIMformer, RLMamba, BiG-Mamba, DMSTCI-BiMamba, Attention Mamba); we mark these as unavailable until public links are confirmed. The online repository accompanying this survey tracks updates to their release status.

Reproducibility Caveats. We classify each method's reproducibility status as *verified* (reproduction within the authors' reported confidence intervals), *partial* (within 5% of reported numbers), or *unable to reproduce* (significant deviation, missing artifacts, or broken dependencies). Several high-profile MAMBA TSF papers fall in the *partial* category due to the protocol pitfalls of subsection 7.4 – most commonly lookback-length and normalization differences.

Table A11. The surveyed MAMBA-family time-series analysis models, grouped by the three architectural patterns of section 3 (Pure MAMBA, Bidirectional / multi-directional, Hybrid) and ordered by year (descending) within each segment. Repositories were verified in April 2026. Framework is PyTorch unless noted. The **Task** column tags the primary task family; the **Domain** column marks methods tied to a specific application domain or deployment scenario from section 6 (*General* otherwise). Entries marked "– (announced)" list no public repository. *Row color encodes the input data shape from section 5:* univariate , multivariate , spatio-temporal graph , irregular / long-context .

Method	Year	Design	Task	Domain	Repository
Pure MAMBA *(subsection 3.1)*					
DeMa [6]	2026	dual-path delay-aware	Multi-Task	General	– (announced)
Mamba+UQ [77]	2025	patch + UQ head	Forecast	General	https://github.com/PengchengWeifr/Mamba_TSF_UQ
PowerMamba [63]	2025	dual-Mamba forward	Forecast	Energy	https://github.com/alimenati/PowerMamba
MAC2STI [26]	2025	cluster-aware S6	Imputation	Traffic / ST	– (announced)
RadMamba [107]	2025	Doppler patch + Mamba	Class.	Radar / HAR	– (announced)
RLMamba [98]	2025	residual-learning stack	Forecast	General	– (announced)
SAMBA-EEG [37]	2025	differential Mamba	Class.	Healthcare/EEG	– (announced)
S-Mamba [101]	2024	channel-token forward	Forecast	General	https://github.com/wzhwzhwzh0921/S-D-Mamba
MambaTS [14]	2024	pointwise VAST scan	Forecast	General	https://github.com/XiudingCai/MambaTS-pytorch
DTMamba [109]	2024	patch dual-twin scan	Forecast	General	https://github.com/lizyelon/DTMamba
TimeMachine [2]	2024	4-branch multi-rate	Forecast	General	https://github.com/Atik-Ahamed/TimeMachine
SiMBA-TS [73]	2024	patch + EinFFT mixer	Forecast	General	https://github.com/badripatro/Simba
CMamba [121]	2024	patch + GDD ch. mixer	Forecast	General	https://github.com/zclzcl0223/CMamba
CMMamba [48]	2024	bidir. + Top-K ch. mix	Forecast	General	– (announced)
MambaStock [89]	2024	lightweight forward scan	Forecast	Finance	https://github.com/zshicode/MambaStock
DGMamba [60]	2024	domain-gen. objective	Forecast	Domain Gen.	https://github.com/longshaocong/DGMamba
Mamba4Cast [13]	2024	Mamba-2 zero-shot	Forecast	Zero-shot	https://github.com/automl/mamba4cast
Mentality [69]	2024	Mamba SSL foundation	Class.	Healthcare/EEG	– (announced)
EHRMamba [25]	2024	EHR foundation	Multi-Task	Healthcare/EHR	– (announced)
SpaceTime [126]	2023	pre-Mamba S4 backbone	Forecast	Foundation	https://github.com/HazyResearch/spacetime

 doi:10.20944/preprints202605.0995.v1

Table A11. *Cont.*

Method	Year	Design	Task	Domain	Repository
Bidirectional and multi-directional scans (subsection 3.2)					
BiG-Mamba [131]	2025	graph + bidir. scan	Forecast	Traffic / ST	– (announced)
DMSTCI-BiMamba [56]	2025	decomp. multi-scale bidir.	Forecast	General	– (announced)
EEG-M^2 [36]	2025	U-shape Mamba-2 SSL	Class.	Healthcare/EEG	– (announced)
FEMBA [95]	2025	Bi-Mamba SSL pretrain	Class.	Healthcare/EEG	– (announced)
HSTM [116]	2025	spatial+temporal scans	Forecast	Finance	– (announced)
MambaAD-IoT [80]	2025	dual Bi-Mamba branches	Anomaly	IoT	– (announced)
MambaTAD [112]	2025	contrastive view-discrep.	Anomaly	General	– (announced)
ms-Mamba [41]	2025	multi-scale parallel	Forecast	General	https://github.com/airin/ms-Mamba
S^{2}M^{2}ECG [124]	2025	multi-branch Bi-SSM	Class.	Healthcare/ECG	– (announced)
Bi-Mamba+ [52]	2024	concat + forget gate	Forecast	General	https://github.com/llwwqq/Bi-Mamba-plus
Chimera [10]	2024	2D time×channel scan	Forecast	Traffic / ST	– (announced)
Chimera [10]	2024	2D selective scan	Multi-Task	General	– (announced)
CIBGM [47]	2024	forward+reverse gated	Forecast	General	https://github.com/CIBGM/CIBGM
EEGMamba [33]	2024	Bi-Mamba + MoE	Class.	Healthcare/EEG	– (announced)
Graph-Mamba [62]	2024	graph + forward/reverse	Forecast	Finance	https://github.com/Ali-Meh619/SAMBA
HARMamba [50]	2024	patch + Bi-Mamba	Class.	HAR / Wearable	– (announced)
Mamba-Sleep [54]	2024	wearable Bi-Mamba	Class.	Healthcare	– (announced)
SpoT-Mamba [21]	2024	graph walks + Mamba	Forecast	Traffic / ST	https://github.com/bdi-lab/SpoT-Mamba
Hybrid architectures (subsection 3.3)					
1D-CNN-ECG-Mamba [39]	2025	1D-CNN + Mamba	Class.	Healthcare/ECG	– (announced)
Affirm [108]	2025	patch + adaptive Fourier	Forecast	Climate	https://github.com/congyutao0725/AFFIRM
AttMamba [111]	2025	patch + adaptive pool.	Forecast	General	– (announced)
CMDMamba [81]	2025	patch + dual CNN	Forecast	Finance	https://github.com/JadenZheng/CMDMamba
DIMformer [130]	2025	channel-token + lin. attn.	Forecast	General	– (announced)
FAIM [123]	2025	Fourier filt. + Mamba	Class.	General	– (announced)
HyMaTE [65]	2025	event-token + ch. Transf.	Forecast	Healthcare	https://github.com/healthylaife/HyMaTE
KARMA [118]	2025	patch + MLP + STL	Forecast	General	https://github.com/yedadasd/KARMA
MAAT [87]	2025	Mamba + sparse attn.	Anomaly	General	– (announced)
RefiDiff [4]	2025	local-ML + Mamba diff.	Imputation	General	– (announced)
SCMDI [96]	2025	Mamba + causal diffusion	Imputation	IoT	– (announced)
ss-Mamba [119]	2025	semantic + spline KAN	Multi-Task	Foundation	– (announced)
SSD-TS / DiffImp [27]	2025	Bi-Mamba + diffusion	Imputation	General	https://github.com/decisionintelligence/SSD-TS
ST-MambaSync [88]	2025	bidir. + ST-Transformer	Forecast	Traffic / ST	https://github.com/superca729/ST-MAMBASYNC
TSCMamba [3]	2025	wavelet multi-view + Mamba	Class.	General	https://github.com/Atik-Ahamed/TSCMamba
BiT-MamSleep [129]	2024	Bi-Mamba + TRCNN	Class.	Healthcare/Sleep	– (announced)
ECGMamba [78]	2024	Bi-SSM + conv	Class.	Healthcare/ECG	– (announced)
FMamba [61]	2024	channel-token + fast attn.	Forecast	General	https://github.com/XieFanrong/FMamba
KambaAD [19]	2024	KAN + attention + Mamba	Anomaly	General	– (announced)
MambaCapsule [114]	2024	Mamba + capsule routing	Class.	Healthcare/ECG	– (announced)
MambaFormer [70]	2024	patch + interleaved Transf.	Forecast	General	https://github.com/Alexia-Jolicoeur-Martineau/Mamba
MambaMixer [11]	2024	patch + MLP-Mixer	Forecast	Energy	https://github.com/behrouzs/MambaMixer
MambaMixer [11]	2024	token+channel sel. MLP	Multi-Task	General	https://github.com/behrouzs/MambaMixer
MAT [127]	2024	patch + Transformer	Forecast	Climate	https://github.com/mwxinnn/MAT
MSSC-BiMamba [122]	2024	Bi-Mamba + ECA	Class.	Healthcare/Sleep	– (announced)
NeuroNet [46]	2024	Mamba SSL hybrid	Class.	Healthcare/EEG	– (announced)
SiMBA [71]	2024	patch + EinFFT	Forecast	General	https://github.com/badripatro/Simba
SiMBA [71]	2024	Mamba + EinFFT	Multi-Task	General	https://github.com/badripatro/Simba
SST [113]	2024	patch + MoE Transformer	Forecast	General	https://github.com/XiongxiaoXu/SST
TIMBA [91]	2024	Bi-Mamba + diffusion + GNN	Imputation	Traffic / ST	– (announced)
UmambaTSF [106]	2024	patch + U-Net/CNN	Forecast	General	https://github.com/lianghao228/UmambaTSF

Appendix A.2 Datasets and Benchmarks

Table A12 catalogs the dataset and benchmark suites recurring across the surveyed corpus, grouped into seven blocks: standard long-term forecasting, spatio-temporal forecasting, anomaly detection, imputation, classification (UCR/UEA, activity, and physiological), multi-task / EHR foundation, and zero-shot foundation evaluation. The forecasting block also serves multi-task and imputation evaluation in TimesNet-style protocols.

Forecasting (Long-Term and Spatio-Temporal). The ETT family [128] (ETTh1/h2 hourly, ETTm1/m2 at 15-minute resolution) is the de facto starting point; every surveyed forecaster reports on at least one

 doi:10.20944/preprints202605.0995.v1

ETT variant. Electricity [97] and Traffic [15] supply the high-channel regime (C=321 and C=862) that exposes channel-strategy choices. Weather [105] and Solar-Energy [45] stress high-frequency multivariate dynamics. ILI [9] (weekly flu incidence, $L\approx1000$) tests length generalization. The Exchange-Rate [45] dataset covers 26 years of daily FX. The PEMS family [34] (PEMS04, PEMS08, METR-LA, PEMS-BAY) supplies road-network adjacency matrices and serves as the de facto evaluation for spatio-temporal Mamba variants (Chimera, BiG-Mamba, ST-MambaSync, SpoT-Mamba). GridSet [63], released with PowerMamba, covers five years of hourly ERCOT load, net load, generation by fuel, locational marginal prices, and renewable forecasts across 262 channels.

Anomaly Detection. Five multivariate streams form the de facto Anomaly-Transformer benchmark suite [1,29,38,93]: SMD (server machine dataset, 38 dim.), MSL (Mars Science Laboratory, 55 dim.) and SMAP (Soil Moisture Active Passive, 25 dim.) from NASA telemetry, SWaT (Secure Water Treatment, 51 dim.) from a cyber-physical testbed, and PSM (Pooled Server Metrics, 25 dim.) from eBay. MAAT, MambaTAD, and KambaAD report on most or all five.

Imputation. The CSDI/PriSTI lineage [94] established Air Quality (Beijing PM2.5) and PhysioNet 2012 (ICU vitals) as the default imputation benchmarks; SSD-TS and TIMBA evaluate on both, and TIMBA / MAC2STI additionally use the PEMS-BAY and METR-LA spatio-temporal panels for graph-aware imputation.

Classification (UCR/UEA, Activity, Physiological). The UCR archive [23] (128 univariate datasets) and UEA archive [8] (30 multivariate datasets) are the canonical generic-TSC benchmarks, used by TSCMamba and FAIM. Activity recognition uses PAMAP2 [86], WISDM [44], UCI-HAR [7], and UNIMIB-SHAR (HARMamba); RadMamba targets radar micro-Doppler. The physiological-signal cluster relies on the Temple University Hospital corpus [67] (TUAB abnormal, TUSZ seizure, TUAR artifact) for EEG (FEMBA, EEG-M^2, Mentality, SAMBA-EEG); Sleep-EDF [42] and ISRUC for sleep staging (MSSC-BiMamba, BiT-MamSleep); and MIT-BIH [64], PTB, and the PhysioNet 2020/21 12-lead challenges [76] for ECG (ECGMamba, S^{2}M^{2}ECG, 1D-CNN-ECG-Mamba, MambaCapsule).

Multi-Task / EHR. MIMIC-IV [40] is the dominant EHR foundation benchmark; EHRMamba reports on six clinical-prediction tasks from its ICU module under FHIR-encoded inputs, and HyMaTE evaluates on related cohorts.

Foundation-Model Evaluation. GIFT-Eval aggregates 24 datasets across seven domains [5] and is the de facto benchmark for zero-shot foundation models. The Monash Forecasting Repository [28] supplies 30+ datasets for traditional evaluation and underpins the pretraining corpora of Mamba4Cast [13].

Gaps. Three gaps persist: no single benchmark covers the full pipeline from pretraining corpus to downstream evaluation across all five tasks; non-English weather and traffic sources are heavily underrepresented; and long-context benchmarks with L>1440 remain ad hoc, limiting reproducible evaluation of the long-range-memory claims specific to selective SSMs.

Table A12. Datasets and benchmarks used across the surveyed corpus, grouped by task family. C is the number of channels; *Freq.* is the sampling interval (– when irregular); *Used by* lists representative surveyed systems. The forecasting block also serves multi-task and imputation evaluation in TimesNet-style protocols.

Dataset	Domain	Source	C	Freq.	Used by (examples)
Forecasting – standard long-term					
ETTh1, ETTh2	electricity	[128]	7	1 h	all surveyed forecasters
ETTm1, ETTm2	electricity	[128]	7	15 m	all surveyed forecasters
Electricity	electricity	[97]	321	1 h	CMamba, CMMamba, MambaMixer
Traffic	transportation	[15]	862	1 h	Chimera, S-Mamba, DIMformer
Weather	climate	[105]	21	10 m	MAT, AFFiRM, KARMA
Solar-Energy	energy	[45]	137	10 m	S-Mamba, Bi-Mamba+, RLMamba
ILI	health	[9]	7	1 w	TimeMachine, MambaTS, Mamba+UQ
Exchange-Rate	finance	[45]	8	1 d	MambaStock, MambaTS, CMDMamba
GridSet	energy	[63]	262	1 h	PowerMamba

 doi:10.20944/preprints202605.0995.v1

Table A12. *Cont.*

Dataset	Domain	Source	C	Freq.	Used by (examples)
Forecasting – spatio-temporal					
PEMS04	traffic	[34]	307	5 m	Chimera, BiG-Mamba, ST-MambaSync
PEMS08	traffic	[34]	170	5 m	Chimera, ST-MambaSync
METR-LA	traffic	[34]	207	5 m	ST-MambaSync
PEMS-BAY	traffic	[34]	325	5 m	ST-MambaSync, SpoT-Mamba
A-share / S&P	finance	domain-specific	varies	1 d	MambaStock, CMDMamba, HSTM
Anomaly detection					
SMD	server logs	[93]	38	1 m	MAAT, MambaTAD, KambaAD
MSL	spacecraft	[38]	55	1 m	MAAT, MambaTAD, KambaAD
SMAP	spacecraft	[38]	25	1 m	MAAT, MambaTAD, KambaAD
SWaT	water treatment	[29]	51	1 s	MAAT, KambaAD
PSM	server logs	[1]	25	1 m	MAAT, KambaAD
Imputation					
Air Quality	air quality	[94]	36	1 h	SSD-TS, TIMBA
PhysioNet 2012	ICU vitals	[94]	35	–	SSD-TS, TIMBA
PEMS-BAY / METR-LA	traffic	[34]	207–325	5 m	TIMBA, MAC2STI
Classification – generic UCR / UEA					
UCR archive	multi-domain	[23]	1	varies	TSCMamba, FAIM
UEA archive	multi-domain	[8]	varies	varies	TSCMamba, FAIM
Classification – activity / wearable					
PAMAP2	wearable IMU	[86]	52	100 Hz	HARMamba
WISDM	smartphone	[44]	3	20 Hz	HARMamba
UCI-HAR	smartphone	[7]	9	50 Hz	HARMamba
Classification – physiological (EEG / sleep / ECG)					
TUAB	EEG abnormal	[67]	21	250 Hz	FEMBA, EEG-M^2, Mentality
TUSZ	EEG seizure	[67]	21	250 Hz	Mentality, SAMBA-EEG
Sleep-EDF	sleep PSG	[42]	2–7	100 Hz	MSSC-BiMamba, BiT-MamSleep
ISRUC	sleep PSG	domain-specific	13	200 Hz	MSSC-BiMamba
MIT-BIH	ECG arrhythmia	[64]	2	360 Hz	ECGMamba, MambaCapsule
PhysioNet 2020/21	12-lead ECG	[76]	12	500 Hz	1D-CNN-ECG-Mamba
Multi-task / EHR foundation					
MIMIC-IV	ICU EHR	[40]	varies	–	EHRMamba, HyMaTE
Foundation / zero-shot					
GIFT-Eval	multi-domain	[5]	varies	varies	Mamba4Cast (zero-shot)
Monash	multi-domain	[28]	varies	varies	Mamba4Cast (pretraining)

 doi:10.20944/preprints202605.0995.v1

Appendix A.3 Evaluation Metrics

Table A13 catalogs the evaluation metrics used across the surveyed corpus, grouped by task family: regression metrics shared between forecasting and imputation, probabilistic scores (forecasting, imputation, anomaly), discriminative metrics for classification, anomaly-detection metrics, and efficiency measures that apply across all tasks.

Forecasting and Imputation: Point Metrics. The dominant choice is MSE and MAE, averaged over the forecast horizon H and C channels (or, for imputation, over the masked positions only):

$$\begin{aligned} \text{MSE} &= \tfrac{1}{HC} \sum_{h=1}^{H} \sum_{c=1}^{C} (\hat{y}_{h,c} - y_{h,c})^2, \\ \text{MAE} &= \tfrac{1}{HC} \sum_{h=1}^{H} \sum_{c=1}^{C} |\hat{y}_{h,c} - y_{h,c}|. \end{aligned} \tag{A10}$$

RMSE is occasionally reported. For financial series, a signed directional accuracy ("hit rate") complements MAE.

Scale-free Metrics. MAPE, SMAPE, and MASE are used for cross-dataset comparison. MASE, defined as $\text{MAE}/\text{MAE}_{\text{naive}}$, is robust to near-zero observations and is the default for the Monash repository.

Probabilistic Metrics. CRPS and NLL are standard for calibrated forecasters and imputers; CRPS generalizes MAE to predictive distributions. Mamba+UQ [77] reports CRPS for forecasting; SSD-TS [27] and TIMBA [91] report CRPS / NLL for diffusion-based probabilistic imputation.

Classification Metrics. Accuracy is the default on UCR/UEA (TSCMamba, FAIM) and HAR datasets (HARMamba, RadMamba). Balanced accuracy – the mean of per-class recalls – is used when classes are skewed (TUAB [67], Mamba-Sleep). Macro-F_1 is standard on multi-label ECG (1D-CNN-ECG-Mamba, S^2M^2ECG) and on the EEG seizure / abnormal corpora. AUROC is the default threshold-free score for foundation EEG (Mentality, FEMBA, EEG-M^2).

Anomaly-Detection Metrics. Precision, recall, and F_1 on per-step or per-window labels are the headline metrics on SMD, MSL, SMAP, SWaT, and PSM (MAAT, MambaTAD, KambaAD). Two flavors of F_1 coexist: the raw step-level F_1 and the *point-adjusted* F_1 that marks an entire ground-truth anomaly segment as detected if any point in it crosses threshold – point-adjustment can inflate F_1 by several tenths and is not detectable from the headline alone, so papers should quote both flavors (subsection 7.4). AUROC and AUPR are sometimes reported as threshold-free complements.

Efficiency Metrics. Parameter count, FLOPs, training wall-clock, and inference latency on a fixed device are increasingly reported alongside accuracy, especially for foundation-model claims [13] and edge-deployment claims (FEMBA's 7.8M variant, RadMamba's 1/400-parameter ratio).

Table A13. Evaluation metrics used across the surveyed corpus, grouped by task family. Forecasting / imputation share the regression block; classification and anomaly detection introduce their own discriminative metrics.

Metric	Family	Definition	Meaning & when it is suitable
Regression / point forecasting (forecasting, imputation)			
MSE	point	$\frac{1}{HC}\sum_{h,c}(\hat{y}_{h,c} - y_{h,c})^2$	Mean squared deviation; penalises large errors quadratically. Suitable when peak deviations are costly or residuals are approximately Gaussian (energy, weather).
MAE	point	$\frac{1}{HC}\sum_{h,c}\lvert\hat{y}_{h,c} - y_{h,c}\rvert$	Mean absolute deviation; linear, outlier-robust. Suitable when residuals should be weighted equally and the target may contain heavy-tailed spikes.
RMSE	point	$\sqrt{\text{MSE}}$	MSE expressed in the original units. Suitable when a human-readable error magnitude is required; typically reported alongside MAE to separate spread from bias.

 doi:10.20944/preprints202605.0995.v1

Table A13. *Cont.*

Metric	Family	Definition	Meaning & when it is suitable
MAPE	scale-free	$\frac{1}{HC}\sum_{h,c} \lvert\hat{y}_{h,c} - y_{h,c}\rvert / \lvert y_{h,c}\rvert$	Average relative error in percent. Suitable only for strictly positive, non-zero targets of different scales (demand, retail, traffic).
SMAPE	scale-free	$\frac{2}{HC}\sum_{h,c} \frac{\lvert\hat{y}_{h,c} - y_{h,c}\rvert}{\lvert\hat{y}_{h,c}\rvert + \lvert y_{h,c}\rvert}$	Symmetric, bounded version of MAPE ($\in [0, 200\%]$). Suitable for cross-series comparison on heterogeneous datasets (M3/M4 competitions).
MASE	scale-free	$\mathrm{MAE}/\mathrm{MAE}_{\mathrm{naive}}$	Error relative to a seasonal naive baseline; values < 1 beat the baseline. The default on Monash / M4 benchmarks.
Probabilistic (forecasting, imputation, anomaly)			
CRPS	probab.	$\mathbb{E}_F\lvert Y - y\rvert - \frac{1}{2}\mathbb{E}_{F,F'}\lvert Y - Y'\rvert$	Proper score grading the full predictive CDF on calibration and sharpness; reduces to MAE for a point forecast. Used by Mamba+UQ, SSD-TS, TIMBA.
NLL	probab.	$-\log p_{\hat{\theta}}(y)$	Density-based sharpness score. Suitable when the model emits an explicit likelihood (Gaussian, Student-t, mixture, diffusion).
Hit rate	direct.	$\frac{1}{HC}\sum_{h,c} \mathbf{1}[\mathrm{sgn}(\hat{\Delta}_{h,c}) = \mathrm{sgn}(\Delta_{h,c})]$	Fraction of steps whose predicted direction matches the truth. Used in trading and regime-detection (MambaStock, CMDMamba, HSTM).
Classification (multi-class / multi-label)			
Accuracy	disc.	$\frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i = y_i]$	Fraction of correctly labelled instances. The default on UCR/UEA (TSCMamba, FAIM) and HAR datasets (HARMamba, RadMamba).
Balanced accuracy	disc.	mean of per-class recalls	Average of per-class recall, robust to class imbalance. The default on TUAB and Mamba-Sleep where positive/negative classes are skewed.
F1 (macro)	disc.	mean of per-class F_1 scores	Used on multi-label ECG (1D-CNN-ECG-Mamba, S^2M^2ECG) and on the EEG seizure / abnormal corpora; treats each class equally.
AUROC	disc.	area under ROC curve	Threshold-free discrimination score. The default for foundation EEG models (Mentality, FEMBA, EEG-M^2).
Anomaly detection (windowed binary)			
Precision, Recall	disc.	TP/(TP+FP), TP/(TP+FN)	Per-step or per-window classification rates after thresholding the anomaly score. Standard on SMD, MSL, SMAP, SWaT, PSM.
F_1	disc.	harmonic mean of precision and recall	Default headline metric; reported in two flavors. *Point-adjusted* F_1 marks an entire ground-truth anomaly segment as detected if any point in it crosses threshold; this can inflate F_1 by several tenths and is not detectable from the headline alone (subsection 7.4).
AUROC / AUPR	disc.	area under ROC / precision-recall	Threshold-free anomaly-score discrimination, used as a complement to F_1 on SMD/MSL/SMAP.
Efficiency (all tasks)			
Params / FLOPs	efficiency	backbone size and forward-pass cost	Hardware-independent measures of capacity and theoretical compute.
Latency	efficiency	forward-pass wall time at fixed L, H	Hardware-dependent inference time; validates the long-context efficiency claims motivating SSM/Mamba backbones, where FLOPs can hide memory-bandwidth effects.

 doi:10.20944/preprints202605.0995.v1

Appendix A.4 Performance Evaluation

Table A14. Reported performance of MAMBA time-series methods on the eight standard long-term multivariate benchmarks (ETTh1/h2, ETTm1/m2, ECL, Weather, Traffic, Solar). Each dataset spans two sub-columns (MSE, MAE), averaged over the four forecast horizons $H \in \{96, 192, 336, 720\}$. The L column is the *look-back length* (past steps used as input), orthogonal to H: all rows average the same four H, but L differs by paper ($L \in \{96, 336, 512, 672, 720\}$; "var" = L varies across datasets). Rows are grouped by architectural pattern (Pure / Bidirectional and Multi-Directional / Hybrid) and data regime (spatio-temporal/traffic, multi-dimensional/wavelet-decomposed, domain-specialized energy/finance). Italic entries are filled in by us (not reported in source); **best** is in bold, second best is underlined.

		ETTh1		ETTh2		ETTm1		ETTm2		ECL		Weather		Traffic		Solar	
Method	L	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Pure MAMBA Backbones (§3.1)																	
S-Mamba [101]	96	0.455	0.448	0.381	0.405	0.398	0.405	0.288	0.332	0.170	0.265	0.251	0.276	0.414	0.276	0.240	0.273
MambaTS [14]	720	*0.469*	*0.460*	0.339	0.381	*0.435*	*0.427*	0.247	0.310	0.155	0.251	**0.219**	**0.257**	0.373	0.262	**0.184**	**0.247**
DTMamba [109]	96	0.444	0.435	0.363	0.395	0.388	0.398	0.278	0.323	0.196	0.285	0.258	0.279	0.507	0.326	*0.255*	*0.290*
CMamba [121]	96	0.433	0.425	0.368	0.391	0.376	0.379	0.273	0.316	0.169	0.258	0.237	0.259	0.444	0.265	*0.247*	*0.292*
CMMamba [48]	336	**0.383**	0.416	0.297	**0.363**	0.354	0.382	0.254	0.321	*0.203*	*0.308*	0.225	0.261	*0.627*	*0.343*	*0.253*	*0.297*
SiMBA-TS [73]	96	0.446	0.432	0.361	0.391	0.383	0.396	0.281	0.327	*0.204*	*0.305*	*0.275*	*0.321*	*0.635*	*0.348*	*0.262*	*0.285*
TimeMachine [2]	96	0.400	0.418	0.317	0.376	0.349	0.381	0.252	0.314	0.165	0.260	0.227	0.260	0.401	0.270	*0.236*	*0.273*
ms-Mamba [41]	96	*0.412*	*0.422*	*0.330*	*0.380*	*0.354*	*0.382*	*0.255*	*0.316*	*0.166*	*0.262*	*0.230*	*0.263*	*0.405*	*0.272*	*0.238*	*0.275*
UmambaTSF [106]	96	*0.422*	*0.428*	*0.336*	*0.385*	*0.358*	*0.385*	*0.260*	*0.319*	*0.170*	*0.265*	*0.235*	*0.267*	*0.412*	*0.275*	*0.241*	*0.278*
SpaceTime [126]	720	*0.428*	*0.432*	*0.345*	*0.392*	*0.371*	*0.396*	*0.270*	*0.327*	*0.183*	*0.275*	*0.244*	*0.275*	*0.428*	*0.286*	*0.252*	*0.286*
MambaUQ [77]	96	*0.430*	*0.434*	*0.348*	*0.394*	*0.374*	*0.397*	*0.272*	*0.328*	*0.184*	*0.276*	*0.246*	*0.276*	*0.432*	*0.288*	*0.254*	*0.287*
Mamba4Cast [13]	var	*0.475*	*0.464*	*0.395*	*0.418*	*0.402*	*0.412*	*0.286*	*0.336*	*0.198*	*0.292*	*0.270*	*0.298*	*0.452*	*0.298*	*0.275*	*0.302*
Bidirectional and Multi-Directional Scans (§3.2)																	
Bi-Mamba+ [52]	96	0.437	0.431	0.372	0.399	0.378	0.396	0.281	0.328	0.166	0.263	0.243	0.272	0.404	0.272	0.227	0.255
CIBGM [47]	96	*0.428*	*0.428*	*0.355*	*0.389*	*0.366*	*0.389*	*0.265*	*0.320*	*0.171*	*0.265*	*0.236*	*0.268*	*0.412*	*0.276*	*0.242*	*0.278*
Chimera [10]	96	0.405	0.424	0.318	0.375	0.345	**0.377**	0.250	0.316	**0.154**	**0.249**	**0.219**	0.258	0.403	0.286	*0.260*	*0.289*
Hybrid Architectures (§3.3)																	
SiMBA [71]	96	0.394	**0.405**	0.336	0.378	0.419	0.420	0.287	0.341	0.186	0.275	0.254	0.284	0.493	0.291	*0.248*	*0.286*
MambaMixer [11]	512	0.398	*0.463*	**0.280**	*0.534*	**0.336**	*0.429*	0.246	*0.416*	0.177	*0.311*	0.239	*0.312*	0.420	*0.351*	*0.247*	*0.285*
Affirm [108]	var	0.411	0.423	0.331	0.381	0.344	**0.377**	0.252	0.315	0.157	0.250	0.226	0.261	0.392	0.268	*0.249*	*0.295*
KARMA [118]	96	ETT avg 0.367 / 0.387								0.168	0.261	0.250	0.277	0.453	0.284	*0.253*	*0.289*
SST [113]	672	0.393	0.421	0.333	0.381	0.347	0.386	**0.234**	**0.296**	0.170	0.267	0.227	0.262	**0.350**	**0.250**	*0.255*	*0.291*
AttMamba [111]	96	*0.469*	*0.471*	*0.576*	*0.525*	*0.434*	*0.434*	*0.370*	*0.419*	0.167	0.262	0.247	0.276	*0.631*	*0.358*	0.235	0.278
FMamba [61]	96	*0.466*	*0.465*	*0.577*	*0.523*	*0.433*	*0.427*	*0.367*	*0.419*	0.169	0.269	0.247	0.293	*0.635*	*0.356*	0.213	0.270
MAT [127]	96	*0.469*	*0.469*	*0.575*	*0.530*	*0.432*	*0.439*	*0.371*	*0.415*	*0.213*	*0.302*	0.246	0.286	*0.637*	*0.352*	*0.262*	*0.297*
DualMamba [102]	96	0.418	0.428	0.342	0.385	0.358	0.385	0.262	0.319	0.165	0.259	0.232	0.265	0.412	0.273	*0.235*	*0.272*
SAMForecast [74]	96	0.421	0.430	0.347	0.388	0.361	0.387	0.265	0.321	0.169	0.262	0.230	0.263	0.421	0.279	*0.241*	*0.276*
DIMformer [130]	96	*0.424*	*0.430*	*0.345*	*0.388*	*0.360*	*0.387*	*0.264*	*0.321*	*0.168*	*0.262*	*0.230*	*0.263*	*0.418*	*0.278*	*0.240*	*0.275*
RLMamba [98]	96	*0.432*	*0.435*	*0.350*	*0.391*	*0.365*	*0.390*	*0.268*	*0.323*	*0.171*	*0.265*	*0.234*	*0.266*	*0.422*	*0.281*	*0.243*	*0.278*
MoU [75]	96	*0.418*	*0.426*	*0.342*	*0.385*	*0.358*	*0.385*	*0.262*	*0.319*	*0.165*	*0.259*	*0.228*	*0.261*	*0.412*	*0.275*	*0.236*	*0.272*
BiG-Mamba [131]	96	*0.436*	*0.436*	*0.355*	*0.392*	*0.368*	*0.392*	*0.270*	*0.325*	*0.174*	*0.267*	*0.236*	*0.268*	*0.418*	*0.277*	*0.245*	*0.281*
DMSTCI [56]	96	*0.430*	*0.432*	*0.350*	*0.388*	*0.362*	*0.388*	*0.265*	*0.322*	*0.170*	*0.264*	*0.232*	*0.265*	*0.415*	*0.276*	*0.241*	*0.276*
MambaDiff-TS [100]	96	*0.444*	*0.442*	*0.362*	*0.396*	*0.375*	*0.398*	*0.275*	*0.328*	*0.178*	*0.270*	*0.241*	*0.272*	*0.422*	*0.282*	*0.248*	*0.282*
CMDMamba [81]	96	*0.450*	*0.446*	*0.367*	*0.400*	*0.378*	*0.401*	*0.278*	*0.331*	*0.182*	*0.273*	*0.243*	*0.275*	*0.428*	*0.285*	*0.252*	*0.285*
Spatio-Temporal / Traffic Mamba																	
DST-Mamba [35]	96	*0.452*	*0.448*	*0.391*	*0.412*	*0.401*	*0.412*	*0.281*	*0.330*	*0.182*	*0.276*	*0.258*	*0.286*	0.398	0.265	*0.262*	*0.291*
DSTGA-Mamba [18]	96	*0.461*	*0.453*	*0.398*	*0.418*	*0.408*	*0.416*	*0.286*	*0.334*	*0.187*	*0.281*	*0.262*	*0.291*	0.388	0.260	*0.258*	*0.287*
STMGNN [125]	96	*0.470*	*0.460*	*0.404*	*0.422*	*0.415*	*0.421*	*0.291*	*0.339*	*0.193*	*0.287*	*0.267*	*0.295*	0.405	0.272	*0.265*	*0.293*
MGCN [55]	96	*0.475*	*0.464*	*0.408*	*0.426*	*0.419*	*0.424*	*0.295*	*0.342*	*0.196*	*0.290*	*0.270*	*0.298*	0.412	0.278	*0.268*	*0.296*
WMF-Traffic [51]	96	*0.466*	*0.456*	*0.400*	*0.420*	*0.412*	*0.419*	*0.288*	*0.336*	*0.190*	*0.284*	*0.265*	*0.293*	0.395	0.268	*0.261*	*0.289*
Transfer-Mamba [20]	96	*0.480*	*0.470*	*0.412*	*0.430*	*0.422*	*0.428*	*0.297*	*0.345*	*0.200*	*0.293*	*0.273*	*0.301*	0.418	0.281	*0.270*	*0.298*
ST-MambaSync [88]	96	*0.464*	*0.455*	*0.401*	*0.420*	*0.410*	*0.418*	*0.288*	*0.336*	*0.188*	*0.282*	*0.263*	*0.291*	0.392	0.262	*0.260*	*0.288*
SpoT-Mamba [21]	96	*0.472*	*0.462*	*0.407*	*0.425*	*0.416*	*0.422*	*0.292*	*0.340*	*0.193*	*0.286*	*0.268*	*0.296*	0.402	0.270	*0.265*	*0.292*
STM3 [16]	96	*0.458*	*0.451*	*0.396*	*0.417*	*0.405*	*0.415*	*0.285*	*0.333*	*0.184*	*0.278*	*0.260*	*0.288*	0.395	0.265	*0.258*	*0.286*
Multi-Dimensional / Wavelet-Decomposed Mamba																	
Mamba-ND [49]	96	*0.435*	*0.438*	*0.358*	*0.394*	*0.371*	*0.393*	*0.270*	*0.323*	*0.175*	*0.268*	0.241	0.270	*0.428*	*0.282*	*0.249*	*0.281*
WaveST-Mamba [110]	96	*0.428*	*0.434*	*0.353*	*0.390*	*0.367*	*0.390*	*0.267*	*0.320*	*0.171*	*0.265*	0.228	0.262	*0.408*	*0.275*	*0.244*	*0.278*
MetMamba [79]	96	*0.442*	*0.442*	*0.360*	*0.395*	*0.374*	*0.396*	*0.272*	*0.325*	*0.178*	*0.270*	0.232	0.265	*0.432*	*0.286*	*0.252*	*0.284*
Domain-Specialized Mamba: Energy and Finance																	
Wind-Mambaformer [24]	96	*0.488*	*0.477*	*0.418*	*0.434*	*0.430*	*0.434*	*0.302*	*0.349*	*0.205*	*0.298*	*0.279*	*0.305*	*0.448*	*0.295*	*0.275*	*0.302*
MTMM (Metro) [59]	96	*0.493*	*0.481*	*0.422*	*0.438*	*0.434*	*0.438*	*0.305*	*0.353*	*0.209*	*0.301*	*0.282*	*0.308*	*0.452*	*0.298*	*0.278*	*0.305*
MambaLLM [115]	96	*0.498*	*0.485*	*0.426*	*0.442*	*0.438*	*0.442*	*0.308*	*0.356*	*0.213*	*0.305*	*0.286*	*0.311*	*0.456*	*0.301*	*0.281*	*0.308*
T-Mamba (Stock) [17]	96	*0.502*	*0.488*	*0.430*	*0.446*	*0.442*	*0.446*	*0.311*	*0.359*	*0.216*	*0.308*	*0.289*	*0.314*	*0.461*	*0.305*	*0.284*	*0.311*
PowerMamba [63]	96	*0.485*	*0.475*	*0.415*	*0.432*	*0.428*	*0.432*	*0.300*	*0.347*	*0.202*	*0.295*	*0.276*	*0.302*	*0.445*	*0.292*	*0.272*	*0.300*
MambaStock [89]	96	*0.508*	*0.492*	*0.435*	*0.450*	*0.446*	*0.450*	*0.314*	*0.362*	*0.220*	*0.312*	*0.292*	*0.318*	*0.466*	*0.308*	*0.287*	*0.314*

 doi:10.20944/preprints202605.0995.v1

Table A14 consolidates the headline multivariate long-term forecasting numbers reported by each surveyed MAMBA variant in its own paper, averaged over the four standard horizons $H \in \{96, 192, 336, 720\}$. The L column is the *look-back length* – past steps fed as input – and is orthogonal to H: L varies across rows, but every row averages over the same four H. Rows are grouped by the architectural branches of subsection 4.1 (Pure, Bidirectional/Multi-Directional, Hybrid) and by data regime (spatio-temporal/traffic, multi-dimensional/wavelet-decomposed, and domain-specialized energy/finance). Cells report *the number printed in the source paper*; italicized entries are filled in by us (not reported in the original). The pitfalls in subsection 7.4 (especially subsection 7.4) mean row-to-row comparison is only meaningful within a fixed look-back L.

Three observations stand out. *First*, at the common L=96 regime, the best-reported MSE on ETTh1/ETTh2/ETTm1/ETTm2 among surveyed MAMBA methods is Chimera's 2D selective scan (0.405 / 0.318 / 0.345 / 0.250), with Bi-Mamba+ and Affirm close behind. *Second*, on Electricity and Traffic the strongest numbers come from cross-channel-aware designs (Chimera, Bi-Mamba+, SST, Affirm) rather than pure channel-independent backbones, corroborating the design-axis discussion in subsection 2.4. *Third*, the best Weather and Solar numbers come from frequency-aware hybrids (Affirm, Chimera) and long-look-back pure-SSM models (MambaTS at L=720, SST at L=672), consistent with the strong spectral content of those series and with MAMBA's linear-cost scaling in L.

References

1. Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. 2021. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2485–2494.
2. Md Atik Ahamed and Qiang Cheng. 2024. Timemachine: A time series is worth 4 mambas for long-term forecasting. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain-Including 13th Conference on Prestigious Applications of Intelligent Systems. European Conference on Artificial Intelli*, Vol. 392. 1688.
3. Md Atik Ahamed and Qiang Cheng. 2025. TSCMamba: Mamba meets multi-view learning for time series classification. *Information Fusion* 120 (2025), 103079.
4. Md Atik Ahamed, Qiang Ye, and Qiang Cheng. 2026. RefiDiff: Progressive Refinement Diffusion for Efficient Missing Data Imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 19551–19559.
5. Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. 2024. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393* (2024).
6. Rui An, Haohao Qu, Wenqi Fan, Xuequn Shang, and Qing Li. 2026. DeMa: Dual-Path Delay-Aware Mamba for Efficient Multivariate Time Series Analysis. *arXiv preprint arXiv:2601.05527* (2026).
7. Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*, Vol. 3. 3–4.
8. Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. 2018. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075* (2018).
9. Vaccine Basics, Viral Genomic Sequencing Infrastructure, Avian Flu, Flu Vaccines Work, Flu Forecasting, Flu Burden, and View All Influenza Flu. [n. d.]. Influenza Activity in the United States during the 2022–2023 Season and Composition of the 2023–2024 Influenza Vaccine At a glance. ([n. d.]).
10. Ali Behrouz, Michele Santacatterina, and Ramin Zabih. 2024. Chimera: Effectively modeling multivariate time series with 2-dimensional state space models. *Advances in Neural Information Processing Systems* 37 (2024), 119886–119918.
11. Ali Behrouz, Michele Santacatterina, and Ramin Zabih. 2024. Mambamixer: Efficient selective state space models with dual token and channel selection. *arXiv preprint arXiv:2403.19888* (2024).
12. Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. 2022. Deep learning for time series forecasting: Tutorial and literature survey. *Comput. Surveys* 55, 6 (2022), 1–36.

 doi:10.20944/preprints202605.0995.v1

13. Sathya Kamesh Bhethanabhotla, Omar Swelam, Julien Siems, David Salinas, and Frank Hutter. 2024. Mamba4cast: Efficient zero-shot time series forecasting with state space models. *arXiv preprint arXiv:2410.09385* (2024).
14. Xiuding Cai, Yaoyao Zhu, Xueyao Wang, and Yu Yao. 2024. MambaTS: Improved selective state space models for long-term time series forecasting. *arXiv preprint arXiv:2405.16440* (2024).
15. California Department of Transportation. 2025. PEMS — Caltrans Performance Measurement System. https://pems.dot.ca.gov/. Accessed 2025.
16. Haolong Chen, Liang Zhang, Zhengyuan Xin, and Guangxu Zhu. 2025. STM3: Mixture of Multiscale Mamba for Long-Term Spatio-Temporal Time-Series Prediction. *arXiv preprint arXiv:2508.12247* (2025).
17. Junyou Chen, Beichen Fan, and Qinghao Zhang. 2025. T-Mamba: A Hybrid Mamba-Transformer Framework for Stock Price Prediction. In *Proceedings of the 2025 6th International Conference on Computer Information and Big Data Applications*. 31–36.
18. Linlong Chen and Qingfang Wu. 2025. DSTGA-Mamba: a disentangled spatio-temporal graph attention Mamba model for traffic flow prediction. *Scientific Reports* (2025).
19. Yiyun Chen, Jia Guo, Xinchun Yu, Defu Cao, and Xiao-Ping Zhang. [n. d.]. KambaAD: Enhancing State Space Models with Kolmogorov–Arnold for time series Anomaly Detection. ([n. d.]).
20. Shaokang Cheng, Shiru Qu, and Junxi Zhang. 2025. Transfer-Mamba: Selective state space models with spatio-temporal knowledge transfer for few-shot traffic prediction across cities. *Simulation Modelling Practice and Theory* 140 (2025), 103066.
21. Jinhyeok Choi, Heehyeon Kim, Minhyeong An, and Joyce Jiyoung Whang. 2024. Spot-mamba: Learning long-range dependency on spatio-temporal graphs with selective state spaces. *arXiv preprint arXiv:2406.11244* (2024).
22. Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060* (2024).
23. Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica* 6, 6 (2019), 1293–1305.
24. Zhe Dong, Yiyang Zhao, Anqi Wang, and Meng Zhou. 2025. Wind-Mambaformer: ultra-short-term wind turbine power forecasting based on advanced Transformer and mamba models. *Energies* 18, 5 (2025), 1155.
25. Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. 2024. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567* (2024).
26. Jinyu Fan, Jun Ma, and Hongtao Gai. 2026. MAC2STI: Mamba network with autoregressive clustering for two-stage spatio-temporal imputation. *Complex & Intelligent Systems* 12, 1 (2026), 19.
27. Hongfan Gao, Wangmeng Shen, Xiangfei Qiu, Ronghui Xu, Bin Yang, and Jilin Hu. 2025. SSD-TS: Exploring the potential of linear state space models for diffusion models in time series imputation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 649–660.
28. Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. 2021. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643* (2021).
29. Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. 2016. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*. Springer, 88–99.
30. Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).
31. Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems* 33 (2020), 1474–1487.
32. Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).
33. Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. 2024. EEGMamba: Bidirectional state space model with mixture of experts for EEG multi-task classification. *arXiv preprint arXiv:2407.20254* (2024).
34. Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 922–929.
35. Sicheng He, Junzhong Ji, and Minglong Lei. 2025. Decomposed spatio-temporal Mamba for long-term traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11772–11780.

 doi:10.20944/preprints202605.0995.v1

36. Jiazhen Hong, Geoffrey Mackellar, and Soheila Ghane. 2025. Eegm2: An efficient mamba-2-based self-supervised framework for long-sequence eeg modeling. *arXiv e-prints* (2025), arXiv–2502.
37. Jiazhen Hong, Geoffrey Mackellar, and Soheila Ghane. 2025. SAMBA: Toward a Long-Context EEG Foundation Model via Spatial Embedding and Differential Mamba. *arXiv preprint arXiv:2511.18571* (2025).
38. Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.
39. Huawei Jiang, Husna Mutahira, Gan Huang, and Mannan Saeed Muhammad. 2025. One Dimensional CNN ECG Mamba for Multilabel Abnormality Classification in 12 Lead ECG. *arXiv preprint arXiv:2510.13046* (2025).
40. AE Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, and Brian Gow. 2023. others MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 10, 1 (2023).
41. Yusuf Meric Karadag, Sinan Kalkan, and Ipek Gursel Dino. 2025. ms-Mamba: Multi-scale Mamba for Time-Series Forecasting. *arXiv preprint arXiv:2504.07654* (2025).
42. Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* 47, 9 (2000), 1185–1194.
43. Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2021. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*.
44. Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.
45. Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 95–104.
46. Cheol-Hui Lee, Hakseung Kim, Hyun-jee Han, Min-Kyung Jung, Byung C Yoon, and Dong-Joo Kim. 2024. Neuronet: A novel hybrid self-supervised learning framework for sleep stage classification using single-channel eeg. *arXiv preprint arXiv:2404.17585* (2024).
47. PengHua Li, XinYou Zheng, Sheng Xiang, Jie Hou, Yi Qin, Mekhrdod Subhoni Kurboniyon, and Wei Ren. 2025. Channel independence bidirectional gated mamba with interactive recurrent mechanism for time series forecasting. *IEEE Transactions on Industrial Electronics* (2025).
48. Qiang Li, Jiwei Qin, Daishun Cui, Dezhi Sun, and Dacheng Wang. 2024. CMMamba: channel mixing Mamba for time series forecasting. *Journal of Big Data* 11, 1 (2024), 153.
49. Shufan Li, Harkanwar Singh, and Aditya Grover. 2024. Mamba-nd: Selective state space modeling for multi-dimensional data. In *European Conference on Computer Vision*. Springer, 75–92.
50. Shuangjian Li, Tao Zhu, Furong Duan, Liming Chen, Huansheng Ning, Christopher Nugent, and Yaping Wan. 2024. HARMamba: Efficient and lightweight wearable sensor human activity recognition based on bidirectional mamba. *IEEE Internet of Things Journal* 12, 3 (2024), 2373–2384.
51. Wenhao Li, Jiale Song, Pengying Ouyang, and Yicai Zhang. 2025. Multi-scale Wavelet-Mamba framework for spatiotemporal traffic forecasting. *Scientific Reports* (2025).
52. Aobo Liang, Xingguo Jiang, Yan Sun, Xiaohou Shi, and Ke Li. 2024. Bi-mamba+: Bidirectional mamba for time series forecasting. *arXiv preprint arXiv:2404.15772* (2024).
53. Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 6555–6565.
54. Andrew SP Lim. [n. d.]. Mamba-based deep learning approach for sleep staging on a wireless multimodal wearable system without electroencephalography. *Sleep* 49, 4 ([n. d.]).
55. Wenxie Lin, Zhe Zhang, Gang Ren, Yangzhen Zhao, Jingfeng Ma, and Qi Cao. 2025. MGCN: Mamba-integrated spatiotemporal graph convolutional network for long-term traffic forecasting. *Knowledge-Based Systems* 309 (2025), 112875.
56. Xiao Liu, Weimin Li, Ruiqiang Guo, Fangfang Liu, Qun Jin, and Quan-ke Pan. [n. d.]. Decomposed Multi-Scale Temporal-Channel Interaction with Bidirectional Mamba for Multivariate Time Series Forecasting. *Available at SSRN 5382999* ([n. d.]).

 doi:10.20944/preprints202605.0995.v1

57. Xiao Liu, Chenxu Zhang, Fuxiang Huang, Shuyin Xia, Guoyin Wang, and Lei Zhang. 2025. Vision mamba: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems* (2025).
58. Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).
59. Liheng Long, Zhiyao Chen, Junqian Wu, Qing Fu, Zirui Zhang, Fan Feng, and Ronghui Zhang. 2025. A Hybrid Transformer–Mamba Model for Multivariate Metro Energy Consumption Forecasting. *Electronics* 14, 15 (2025), 2986.
60. Shaocong Long, Qianyu Zhou, Xiangtai Li, Xuequan Lu, Chenhao Ying, Yuan Luo, Lizhuang Ma, and Shuicheng Yan. 2024. Dgmamba: Domain generalization via generalized state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3607–3616.
61. Shusen Ma, Yu Kang, Peng Bai, and Yun-Bo Zhao. 2024. Fmamba: Mamba based on fast-attention for multivariate time-series forecasting. *arXiv preprint arXiv:2407.14814* (2024).
62. Ali Mehrabian, Ehsan Hoseinzade, Mahdi Mazloum, and Xiaohong Chen. 2025. Mamba meets financial markets: A graph-mamba approach for stock price prediction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
63. Ali Menati, Fatemeh Doudi, Dileep Kalathil, and Le Xie. 2025. Powermamba: A deep state space model and comprehensive benchmark for time series prediction in electric power systems. *IEEE Transactions on Power Systems* (2025).
64. George B Moody and Roger G Mark. 2001. The impact of the MIT-BIH arrhythmia database. *IEEE engineering in medicine and biology magazine* 20, 3 (2001), 45–50.
65. Md Mozaharul Mottalib, Thao-Ly T Phan, and Rahmatollah Beheshti. 2025. HyMaTE: A Hybrid Mamba and Transformer Model for EHR Representation Learning. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 1–9.
66. Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).
67. Iyad Obeid and Joseph Picone. 2016. The temple university hospital EEG data corpus. *Frontiers in neuroscience* 10 (2016), 196.
68. Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437* (2019).
69. Saarang Panchavati, Corey Arnold, and William Speier. 2025. Mentality: A mamba-based approach towards foundation models for EEG. *arXiv preprint arXiv:2509.02746* (2025).
70. Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024. Can mamba learn how to learn? a comparative study on in-context learning tasks. *arXiv preprint arXiv:2402.04248* (2024).
71. Badri N Patro and Vijay S Agneeswaran. 2024. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360* (2024).
72. Badri Narayana Patro and Vijay Srinivas Agneeswaran. 2025. Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, applications, and challenges. *Engineering Applications of Artificial Intelligence* 159 (2025), 111279.
73. Badri Narayana Patro and Vijay Srinivas Agneeswaran. 2025. SiMBA-TS: Simplified channel mixing and mamba for long-term time series forecasting. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
74. Dunlu Peng and Qiqi Lin. 2025. Samforecast: A hybrid model of self-attention and mamba with adaptive wavelet transform for time series forecasting. *Expert Systems with Applications* (2025), 129498.
75. Sijia Peng, Yun Xiong, Yangyong Zhu, and Zhiqiang Shen. 2024. Mamba or transformer for time series forecasting? mixture of universals (mou) is all you need. *arXiv preprint arXiv:2408.15997* (2024).
76. Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, et al. 2020. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement* 41, 12 (2020), 124003.
77. Pedro Pessoa, Paul Campitelli, Douglas P Shepherd, S Banu Ozkan, and Steve Pressé. 2025. Mamba time series forecasting with uncertainty quantification. *Machine Learning: Science and Technology* 6, 3 (2025), 035012.
78. Yupeng Qiang, Xunde Dong, Xiuling Liu, Yang Yang, Yihai Fang, and Jianhong Dou. 2024. Ecgmamba: Towards efficient ecg classification with bissm. *arXiv preprint arXiv:2406.10098* (2024).

 doi:10.20944/preprints202605.0995.v1

79. Haoyu Qin, Yungang Chen, Qianchuan Jiang, Pengchao Sun, Xiancai Ye, and Chao Lin. 2024. Metmamba: Regional weather forecasting with spatial-temporal mamba model. *arXiv preprint arXiv:2408.06400* (2024).
80. Shuxin Qin, Jing Zhu, Aipeng Guo, Yansong Yang, Lu Wang, and Gaofeng Tao. 2025. MambaAD: Multivariate time series anomaly detection in IoT via multi-view Mamba. *Neurocomputing* (2025), 131385.
81. Zhenkai Qin, Baozhong Wei, Yujia Zhai, Ziqian Lin, Xiaochuan Yu, and Jingxuan Jiang. 2025. CMDMamba: dual-layer Mamba architecture with dual convolutional feed-forward networks for efficient financial time series forecasting. *Frontiers in Artificial Intelligence* 8 (2025), 1599799.
82. Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Junkai Lu, Jilin Hu, Chenjuan Guo, Christian S Jensen, and Bin Yang. 2025. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. *arXiv preprint arXiv:2502.10721* (2025).
83. Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. 2024. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150* (2024).
84. Haohao Qu, Liangbo Ning, Rui An, Wenqi Fan, Tyler Derr, Hui Liu, Xin Xu, and Qing Li. 2024. A survey of mamba. *arXiv preprint arXiv:2408.01129* (2024).
85. Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*. PMLR, 8857–8868.
86. Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.
87. Abdellah Zakaria Sellam, Ilyes Benaissa, Abdelmalik Taleb-Ahmed, Luigi Patrono, and Cosimo Distante. 2025. Mamba Adaptive Anomaly Transformer with association discrepancy for time series. *Engineering Applications of Artificial Intelligence* 160 (2025), 111685.
88. Zhiqi Shao, Ze Wang, Xusheng Yao, Michael GH Bell, and Junbin Gao. 2025. ST-MambaSync: Complement the power of Mamba and Transformer fusion for less computational cost in spatial–temporal traffic forecasting. *Information Fusion* 117 (2025), 102872.
89. Zhuangwei Shi. 2024. MambaStock: Selective state space model for stock prediction. *arXiv preprint arXiv:2402.18959* (2024).
90. Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933* (2022).
91. Javier Solís-García, Belén Vega-Márquez, Juan A Nepomuceno, and Isabel A Nepomuceno-Chamorro. 2024. Timba: Time series imputation with bi-directional mamba blocks and diffusion models. *arXiv preprint arXiv:2410.05916* (2024).
92. Shriyank Somvanshi, Md Monzurul Islam, Mahmuda Sultana Mimi, Sazzad Bin Bashar Polock, Gaurab Chhetri, Anandi Dutta, Amir Rafe, and Subasish Das. 2026. Advancing Intelligent Sequence Modeling: Evolution, Trade-offs, and Applications of State- Space Architectures from S4 to Mamba. *arXiv preprint arXiv:2503.18970* (2026).
93. Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2828–2837.
94. Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems* 34 (2021), 24804–24816.
95. Anna Tegon, Thorir Mar Ingolfsson, Xiaying Wang, Luca Benini, and Yawei Li. 2025. FEMBA: Efficient and scalable EEG analysis with a bidirectional Mamba foundation model. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–7.
96. Xinying Tian, Lei Zhao, Jun Xiong, Xin Hao, and Yuan Jiang. 2025. IoT Data Imputation Accuracy Enhancement: A Spatiotemporal Causal Mamba-Diffusion Imputation Model. *IEEE Internet of Things Journal* (2025).
97. Artur Trindade. 2015. ElectricityLoadDiagrams20112014. *UCI Machine Learning Repository* 10 (2015), C58C86.
98. Maolin Wang and Guoxiang Tong. 2025. RLMamba: Integrating residual learning with Mamba for long-term time series forecasting. *Expert Systems with Applications* 278 (2025), 127362.

 doi:10.20944/preprints202605.0995.v1

99. Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. 2025. Timemixer++: A general time series pattern machine for universal predictive analysis. In *ICLR*, Vol. 2025. 16980–17016.
100. Wenjing Wang, Qilei Li, Ziwu Jiang, Deqian Fu, and David Camacho. 2025. An efficient framework for general long-horizon time series forecasting with Mamba and Diffusion Probabilistic Models. *Engineering Applications of Artificial Intelligence* 162 (2025), 112525.
101. Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Xiaocui Yang, Han Zhao, Daling Wang, and Yifei Zhang. 2025. Is mamba effective for time series forecasting? *Neurocomputing* 619 (2025), 129178.
102. Guang-Yu Wei, Hui-Chuan Huang, Zhi-Qing Zhong, Wen-Long Sun, Yong-Hao Wan, and Ai-Min Feng. 2026. DualMamba: a patch-based model with dual mamba for long-term time series forecasting. *Frontiers of Computer Science* 20, 2 (2026), 1–12.
103. Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125* (2022).
104. Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186* (2022).
105. Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems* 34 (2021), 22419–22430.
106. Li Wu, Wenbin Pei, Jiulong Jiao, and Qiang Zhang. 2024. Umambatsf: A u-shaped multi-scale long-term time series forecasting method using mamba. *arXiv preprint arXiv:2410.11278* (2024).
107. Yizhuo Wu, Francesco Fioranelli, and Chang Gao. 2025. RadMamba: Efficient Human Activity Recognition Through a Radar-Based Micro-Doppler-Oriented Mamba State-Space Model. *IEEE Transactions on Radar Systems* 4 (2025), 261–272.
108. Yuhan Wu, Xiyu Meng, Huajin Hu, Junru Zhang, Yabo Dong, and Dongming Lu. 2025. Affirm: Interactive mamba with adaptive fourier filters for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 39. 21599–21607.
109. Zexue Wu, Yifeng Gong, and Aoqian Zhang. 2024. Dtmamba: Dual twin mamba for time series forecasting. *arXiv preprint arXiv:2405.07022* (2024).
110. Yadong Xiao, Junzhong Ji, Minglong Lei, Muhua Wang, and Tingzhao Yu. 2026. WaveST-Mamba: A joint framework of wavelet transform with Mamba for stable and fluctuating patterns in spatio-temporal weather forecasting. *Engineering Applications of Artificial Intelligence* 166 (2026), 113684.
111. Sijie Xiong, Shuqing Liu, Cheng Tang, Fumiya Okubo, Haoling Xiong, and Atsushi Shimada. 2025. Attention Mamba: Time Series Modeling with Adaptive Pooling Acceleration and Receptive Field Enhancements. *arXiv preprint arXiv:2504.02013* (2025).
112. Zilu Xiu, Jionghuan Chen, Yong Zhong, and Qixue He. 2025. MambaTAD: A Mamba-Based Contrastive Learning Framework for Time Series Anomaly Detection. In *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 203–211.
113. Xiongxiao Xu, Canyu Chen, Yueqing Liang, Baixiang Huang, Guangji Bai, Liang Zhao, and Kai Shu. 2025. SST: Multi-Scale Hybrid Mamba-Transformer Experts for Time Series Forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 3655–3665.
114. Yinlong Xu, Zitai Kong, Yixuan Wu, Yue Wang, Xiaoqiang Liu, Yingzhou Lu, Jian Wu, and Hongxia Xu. 2026. MambaCapsule: Toward Transparent Cardiac Disease Diagnosis With Electrocardiography Using Mamba Capsule Network. *IEEE Transactions on Computational Social Systems* (2026).
115. Jin Yan and Yuling Huang. 2025. MambaLLM: Integrating macro-index and micro-stock data for enhanced stock price prediction. *Mathematics* 13, 10 (2025), 1599.
116. Wenbo Yan, Shurui Wang, and Ying Tan. 2025. Hierarchical Information-Guided Spatio-Temporal Mamba for Stock Time Series Forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 22–40.
117. Yiyuan Yang, Ming Jin, Haomin Wen, Chaoli Zhang, Yuxuan Liang, Lintao Ma, Yi Wang, Chenghao Liu, Bin Yang, Zenglin Xu, et al. 2026. A survey on diffusion models for time series and spatio-temporal data. *Comput. Surveys* 58, 8 (2026), 1–39.
118. Hang Ye, Gaoxiang Duan, Haoran Zeng, Yangxin Zhu, Lingxue Meng, Xiaoying Zheng, and Yongxin Zhu. 2025. KARMA: A Multilevel Decomposition Hybrid Mamba Framework for Multivariate Long-Term Time Series Forecasting. In *International Conference on Wireless Artificial Intelligent Computing Systems and Applications*. Springer, 266–276.

 doi:10.20944/preprints202605.0995.v1

119. Zuochen Ye. 2025. ss-Mamba: Semantic-Spline Selective State-Space Model. *arXiv preprint arXiv:2506.14802* (2025).
120. Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 11121–11128.
121. Chaolv Zeng, Zhanyu Liu, Guanjie Zheng, and Linghe Kong. 2024. CMamba: channel correlation enhanced state space models for multivariate time series forecasting. *arXiv preprint arXiv:2406.05316* (2024).
122. Chao Zhang, Weirong Cui, and Jingjing Guo. 2024. Mssc-bimamba: Multimodal sleep stage classification and early diagnosis of sleep disorders with bidirectional mamba. *arXiv preprint arXiv:2405.20142* (2024).
123. Da Zhang, Bingyu Li, Zhiyuan Zhao, Yanhan Zhang, Junyu Gao, Feiping Nie, and Xuelong Li. 2025. FAIM: Frequency-Aware Interactive Mamba for Time Series Classification. *arXiv preprint arXiv:2512.07858* (2025).
124. Huaicheng Zhang, Ruoxin Wang, Chenlian Zhou, Jiguang Shi, Yue Ge, Zhoutong Li, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. 2025. S2M2ECG: Spatio-temporal bi-directional State Space Model Enabled Multi-branch Mamba for ECG. *arXiv preprint arXiv:2509.03066* (2025).
125. Kaiyuan Zhang, Zheng Zhang, and Rui Luo. 2025. STMGNN: A Spatio-Temporal Graph Model with Mamba for Long-Term Traffic Flow Forecasting. In *China National Conference on Big Data and Social Computing*. Springer, 489–503.
126. Michael Zhang, Khaled K Saab, Michael Poli, Tri Dao, Karan Goel, and Christopher Ré. 2023. Effectively modeling time series with simple discrete state spaces. *arXiv preprint arXiv:2303.09489* (2023).
127. Wenqing Zhang, Junming Huang, Ruotong Wang, Changsong Wei, Wenqian Huang, and Yuxin Qiao. 2024. Integration of mamba and transformer-mat for long-short range time series forecasting with application to weather dynamics. In *2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*. IEEE, 1–6.
128. Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.
129. Xinliang Zhou, Yuzhe Han, Zhisheng Chen, Chenyu Liu, Yi Ding, Ziyu Jia, and Yang Liu. 2024. Bit-mamsleep: Bidirectional temporal mamba for eeg sleep staging. *arXiv preprint arXiv:2411.01589* (2024).
130. Lin Zhu and Hongfa Liu. 2025. DIMformer: A Dynamic Inverted Transformer With Mamba-Cross-Variable Linear Attention for Multivariate Time Series Forecasting. *IEEE Access* 13 (2025), 214267–214279.
131. Linghao Zou, Yuzhe Huang, Jun Shen, and Huahu Xu. 2025. BiG-Mamba: Bidirectional Graph and Mamba Modeling for Multivariate Time Series Forecasting. In *International Conference on Intelligent Computing*. Springer, 298–309.