ICLR

SNU IMLAB
Dept. of Industrial Engineering
Seoul National University

MambaSL : Exploring Single-Layer Mamba for Time Series Classification

Yoo-Min Jung, pamela7384@gmail.com
Leekyung Kim klk97@snu.ac.kr

Department of Industrial Engineering, Seoul National University

Motivation

Motivation

1. Underexplored in time series classification (TSC) using Mamba
   - in time series forecasting (TSF)
     - TimeMachine (Ahamed & Cheng, 2024), S-Mamba (Wang et al., 2025), …
       - showed promising performance with efficiency
   - in TSC
     - poor performance with vanilla Mamba (Wang et al., 2024)
     - TSCMamba (Ahamed & Cheng, 2025)
       - only variant that is tested on extensive benchmark datasets

→ Find out **why** naively applying Mamba to TSC is not working

Motivation

2. Issues in benchmarking TSC results

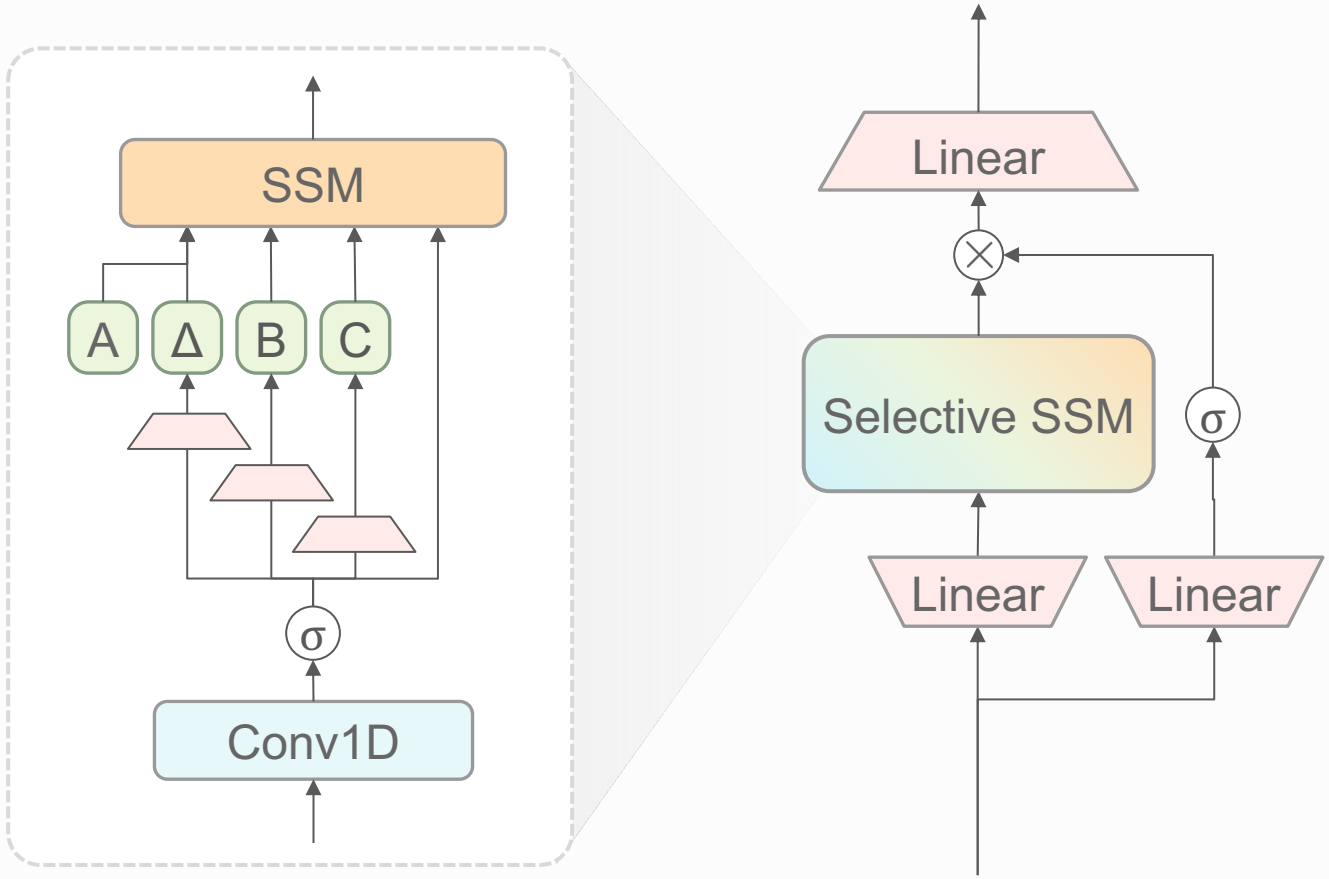
- coverage
  - each study uses a different subset of the UEA dataset (Bagnall et al., 2018)
- fairness
  - uneven hyperparameter tuning (especially for TSF-origin model)
- reproducibility
  - reported results often fail to reproduce original results (Eldele et al., 2024)

→ establish fair, reproducible, and full UEA 30 benchmarking

Background

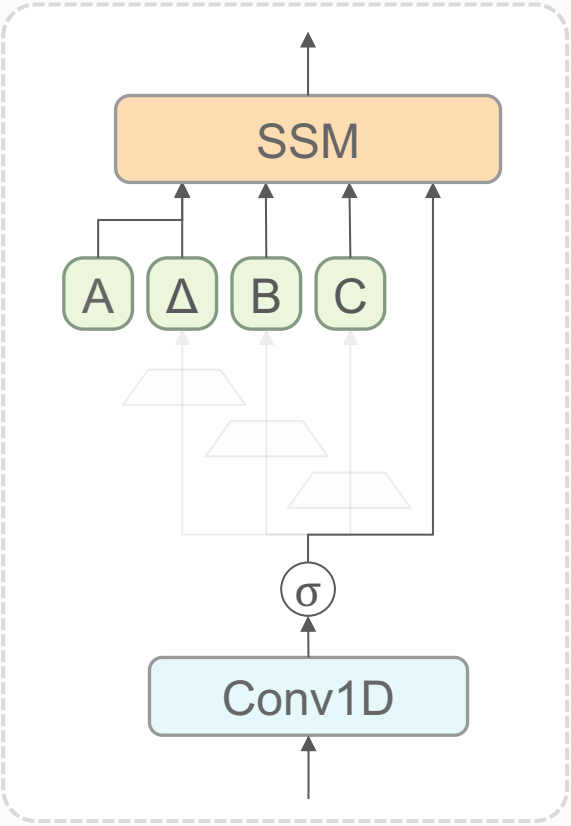
Background

- **Mamba** = Selective SSM + Gated Linear Unit



Background

- SSM (state space model)



✓ $\boldsymbol{s}(t)$: state vector, $u(t), y(t)$: input/output scalar

$$\dot{\boldsymbol{s}}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{B}u(t), \quad y(t) = \boldsymbol{C}\boldsymbol{s}(t) + \boldsymbol{D}u(t), \tag{5}$$

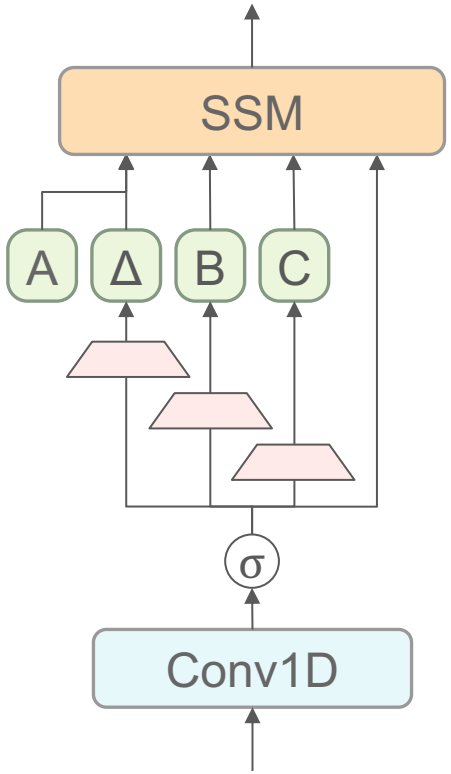
$$\boldsymbol{s}_t = \bar{\boldsymbol{A}}\boldsymbol{s}_{t-1} + \bar{\boldsymbol{B}}u_t, \quad y_t = \boldsymbol{C}\boldsymbol{s}_t + \boldsymbol{D}u_t, \tag{6}$$

Discretize
(fixed step $\Delta > 0$)

- Characteristics
  - LTI (Linear Time-Invariant) System
  - Channel Independent

Background

- Selective SSM



✓ $\tilde{x}_t$: given input at time $t \rightarrow \boldsymbol{\Delta}_t, \boldsymbol{B}_t, \boldsymbol{C}_t$

$$
\begin{aligned}
\boldsymbol{\Delta}_t &= \phi_\Delta(\tilde{\boldsymbol{x}}_t) = \zeta\big(\mathrm{Linear}^{\mathrm{bias}}_{d_\mathrm{m}}\big(\mathrm{Linear}_{d_\mathrm{r}}(\tilde{\boldsymbol{x}}_t)\big)\big) &&\in \mathbb{R}^{d_\mathrm{m}}_{>0}, \\
\boldsymbol{B}_t &= \phi_B(\tilde{\boldsymbol{x}}_t) = \mathrm{Linear}_{d_\mathrm{s}}(\tilde{\boldsymbol{x}}_t) &&\in \mathbb{R}^{d_\mathrm{s}\times 1}, \qquad (7) \\
\boldsymbol{C}_t &= \phi_C(\tilde{\boldsymbol{x}}_t) = \mathrm{Linear}_{d_\mathrm{s}}(\tilde{\boldsymbol{x}}_t)^\top &&\in \mathbb{R}^{1\times d_\mathrm{s}}.
\end{aligned}
$$

- Characteristics
  - Time-Varying (TV) System
  - Channel Mixed

Background

- On time variance in Δ, B, and C

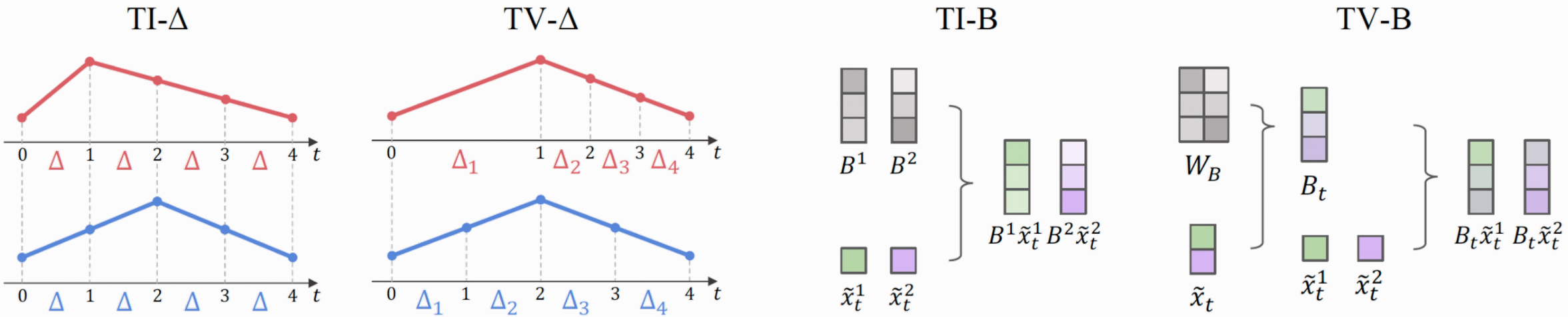
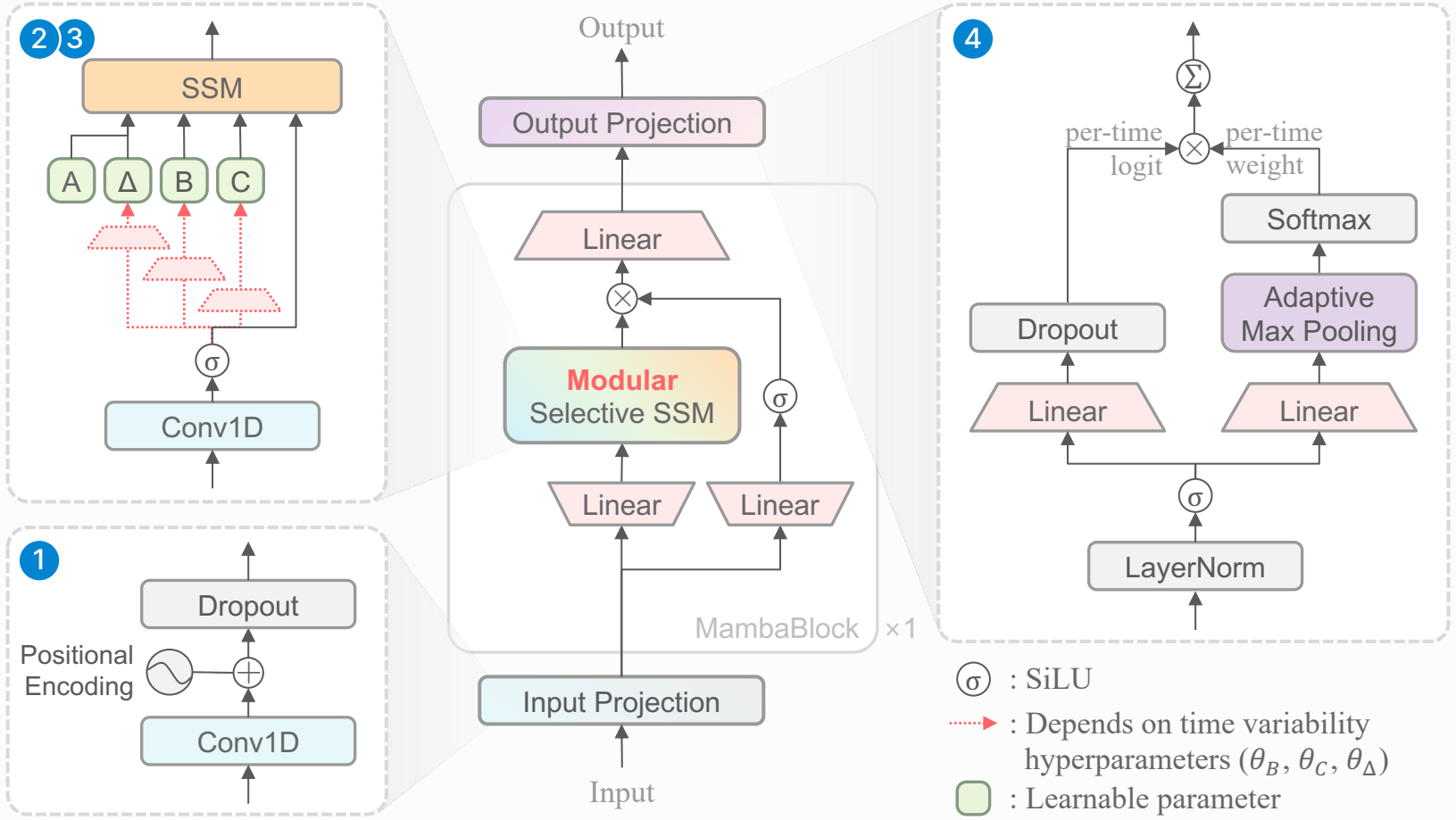


Figure 1: TI/TV parameterization of (left) Δ and (right) $\boldsymbol{B}$ in the SSM. TI-Δ fixes the update rate, while TV-Δ adapts it to align sequences with varying speeds. TI-$\boldsymbol{B}$ preserves channel independence, whereas TV-$\boldsymbol{B}$ introduces input-dependent mixing. $\boldsymbol{C}$ follows the same pattern as $\boldsymbol{B}$ at the output stage, highlighting the temporal pacing of Δ versus the spatial routing of $\boldsymbol{B}$ and $\boldsymbol{C}$.

Method & Results

Method

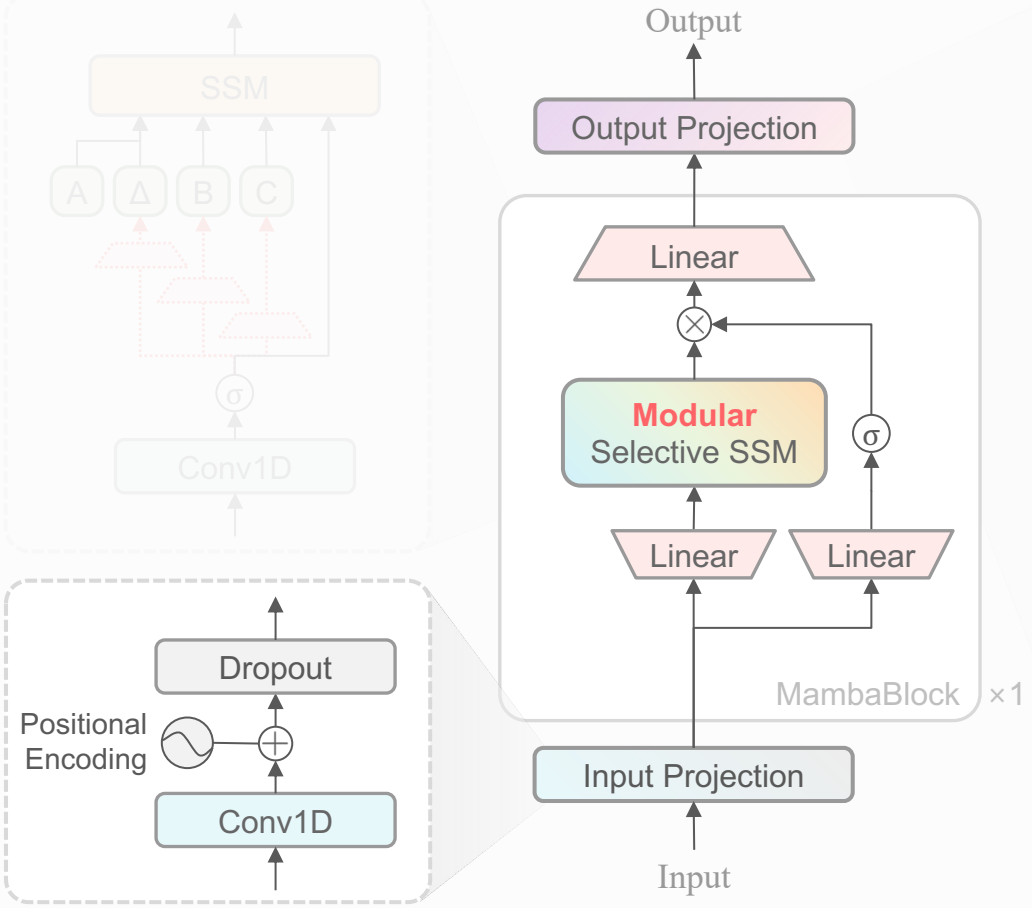
- **MambaSL** : TSC framework with a single-layer Mamba



1. Scale input projection
2. Modularize time (in)variance
3. Remove skip connection
4. Aggregate via adaptive pooling

Method

- (H1) Scale Input Projection

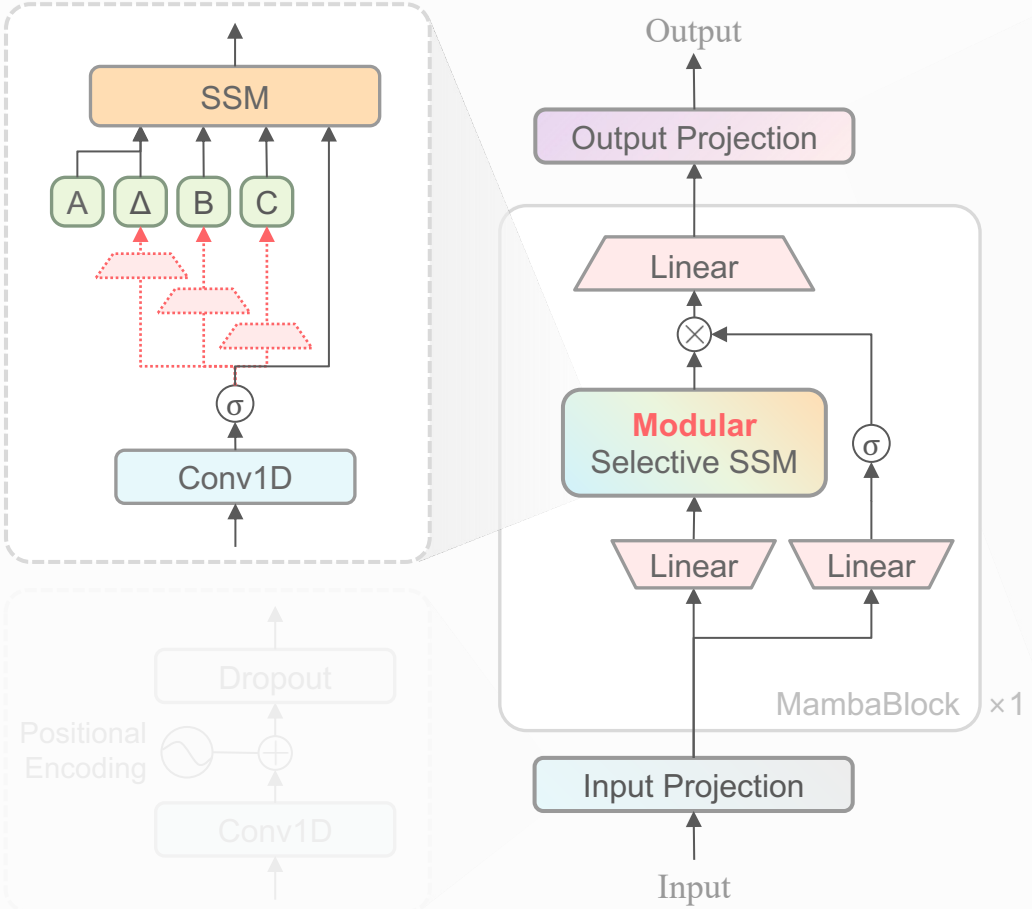


$$k = \max(k_{\min}, \lfloor \lambda L \rfloor) \qquad (10)$$

- Mamba's gating mechanism modulates the SSM output based on input projection.
- time series dataset may require larger receptive fields to generate token embedding
  - especially for dataset with high sampling rate
  - especially for dataset with long sequence length

Method

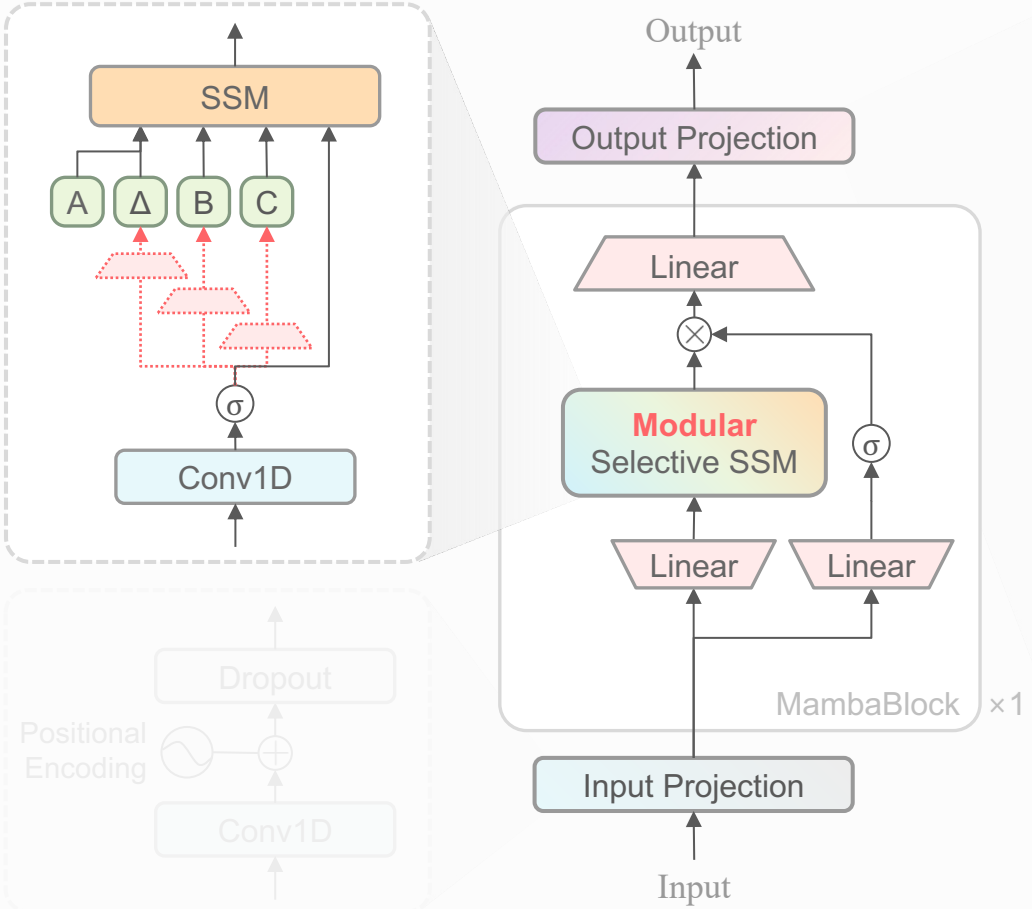
- (H2) Modularize time (in)variance



- decouple time variance of Δ, B, and C as a hyperparameter → $\theta_\Delta, \theta_B, \theta_C \in \{0, 1\}$
- 8 combinations
  - if all $\theta = 0$ → LTI & channel-independent system
  - if all $\theta = 1$ → TV & channel-mixed system
- Ablation results showed that simpler, near time-invariant settings can be more effective.

Method

- (H3) Remove skip connection



- remove D from Mamba block

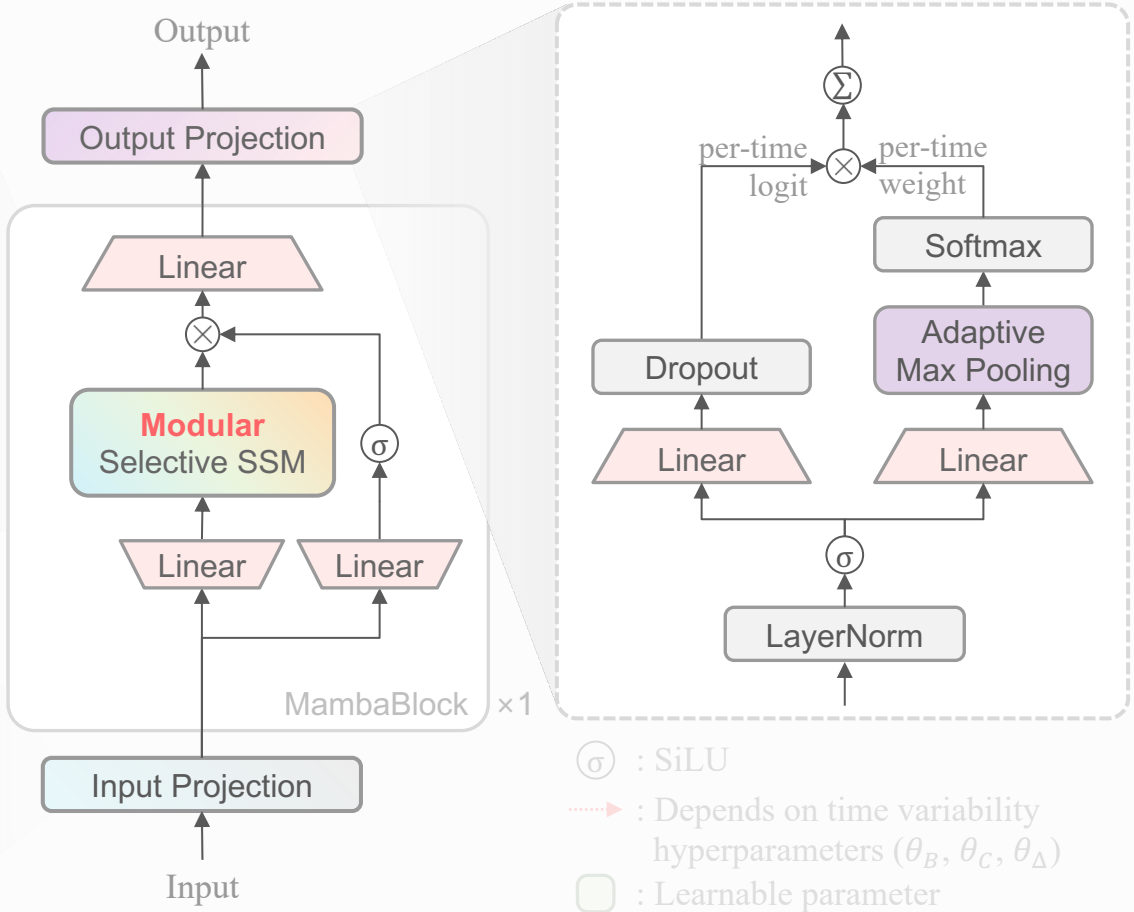
$$f_t^{(j)} = \boldsymbol{C}_t \boldsymbol{s}_t^{(j)} \quad \text{(no skip term)} \quad (12)$$

- Most time series classification models are in shallow layers (1 – 6 layers)

- InceptionTime (Chen et al., 2015) showed that the skip connection makes little difference across 85 UCR datasets..

Method

- (H4) Aggregate via adaptive pooling



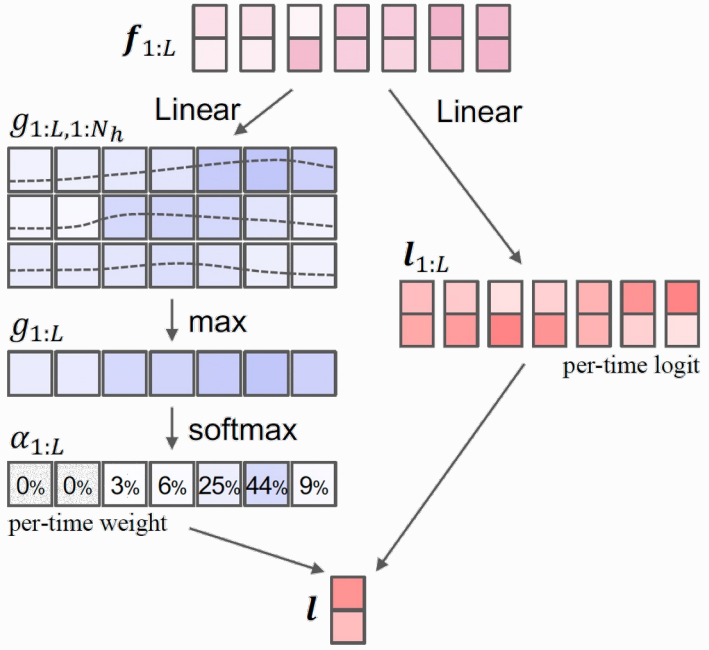
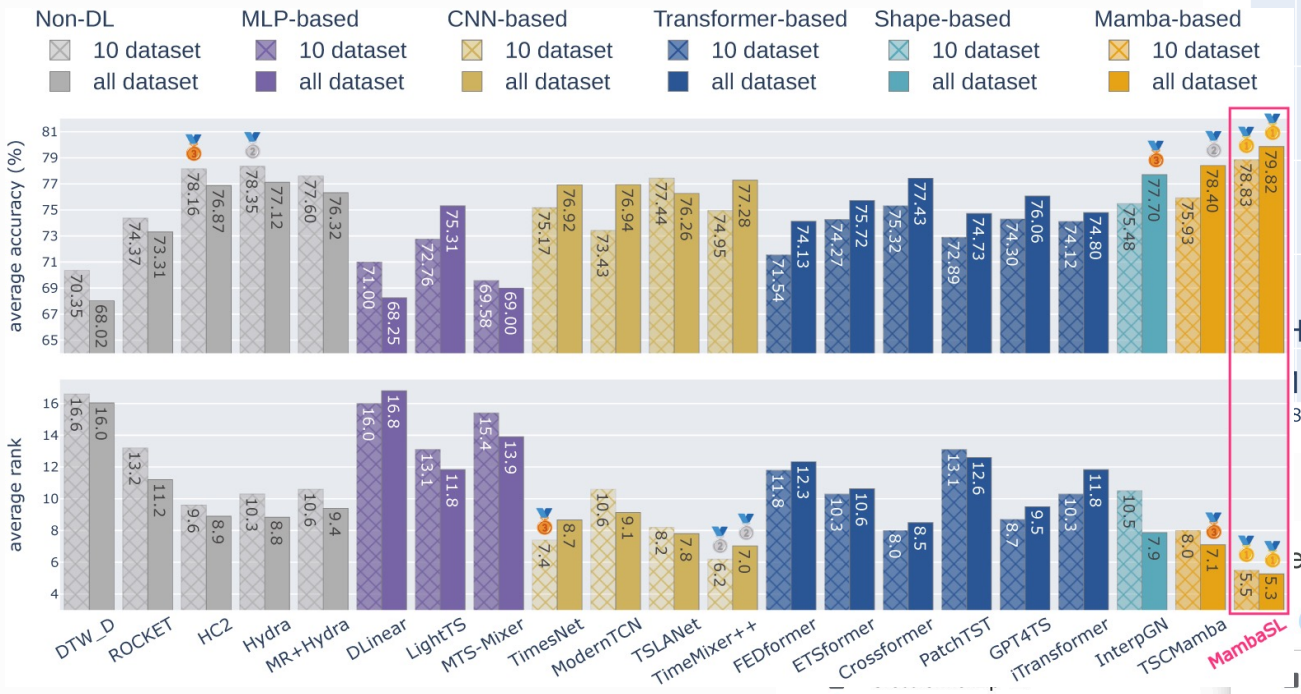
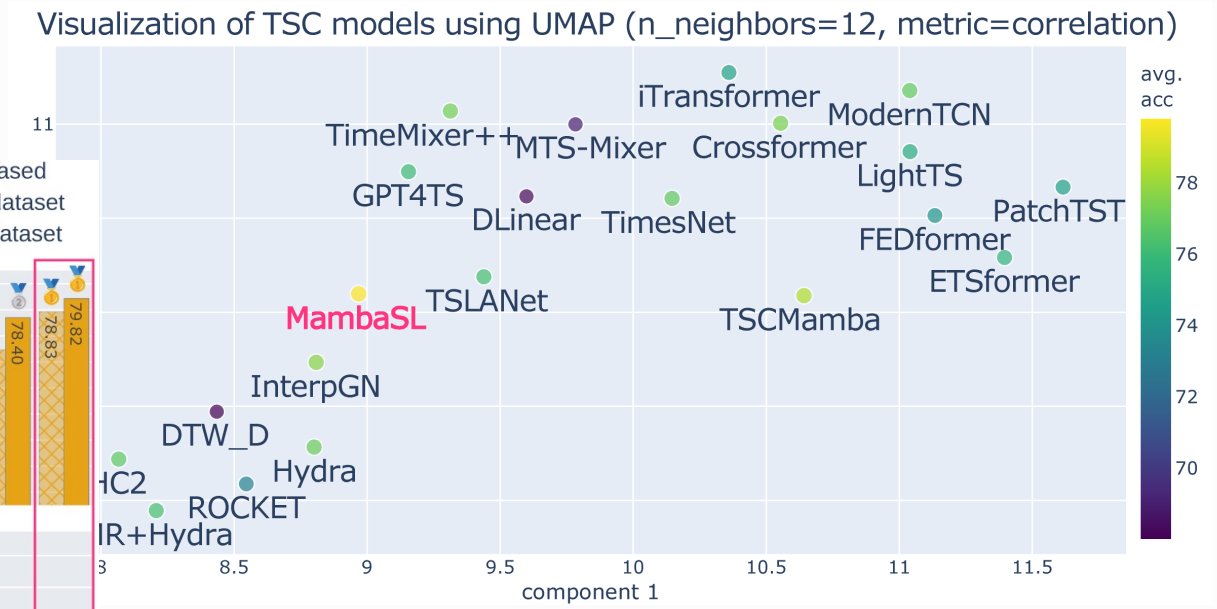


Figure 3: Illustration of proposed multi-head adaptive pooling with $(L, d_{\mathrm{m}}, N_h, d_{\mathrm{y}}) = (7, 2, 3, 2)$.

Results





checkpoint_best

- DLinear.zip
- ETSformer.zip
- FEDformer.zip
- GPT4TS.zip
- InterpretGN.zip
- iTransformer.zip
- LightTS.zip
- MambaSL (ADFTD).zip
- MambaSL (FLAAP).zip
- MambaSL (inceptiontime-setting).zip
- MambaSL (multilayer).zip
- MambaSL (UCR, inceptiontime-setting).zip
- MambaSL.zip

checkpoint_best

이름

- ModernTCN.zip
- MTSMixer.zip
- PatchTST.zip
- TimeMixerPP.zip
- TimesNet.zip
- TSCMamba.zip
- TSLANet.zip

ICLR

SNU IMLAB
Dept. of Industrial Engineering
Seoul National University

Thank you

Paper

Code

MambaSL : Exploring Single-Layer Mamba for Time Series Classification