BRL Seminar

(2023. 08. 22. Tue)

TSMixer: An all-MLP Architecture for Time Series Forecasting

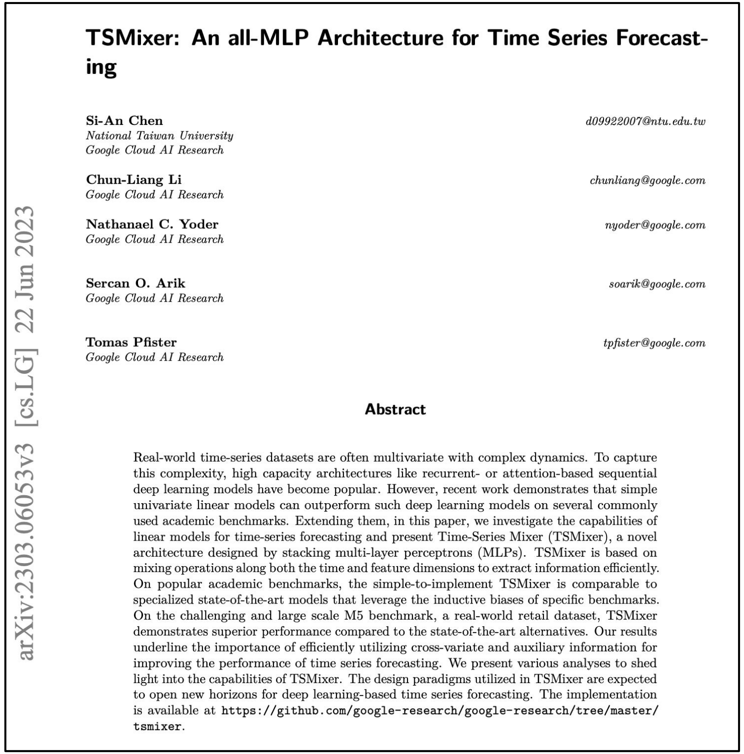
통합과정 6학기 이승한

Paper

TSMixer: An all-MLP Architecture for Time Series Forecasting

(TMLR under review)

- https://arxiv.org/pdf/2303.06053.pdf
- https://github.com/google-research/google-research/tree/master/tsmixer



TSMixer: An all-MLP Architecture for Time Series Forecasting

Si-An Chen d09922007@ntu.edu.tw
National Taiwan University
Google Cloud AI Research

Chun-Liang Li chunliang@google.com
Google Cloud AI Research

Nathanael C. Yoder nyoder@google.com
Google Cloud AI Research

Sercan O. Arik soarik@google.com
Google Cloud AI Research

Tomas Pfister tpfister@google.com
Google Cloud AI Research

arXiv:2303.06053v3 [cs.LG] 22 Jun 2023

Abstract

Real-world time-series datasets are often multivariate with complex dynamics. To capture this complexity, high capacity architectures like recurrent- or attention-based sequential deep learning models have become popular. However, recent work demonstrates that simple univariate linear models can outperform such deep learning models on several commonly used academic benchmarks. Extending them, in this paper, we investigate the capabilities of linear models for time-series forecasting and present Time-Series Mixer (TSMixer), a novel architecture designed by stacking multi-layer perceptrons (MLPs). TSMixer is based on mixing operations along both the time and feature dimensions to extract information efficiently. On popular academic benchmarks, the simple-to-implement TSMixer is comparable to specialized state-of-the-art models that leverage the inductive biases of specific benchmarks. On the challenging and large scale M5 benchmark, a real-world retail dataset, TSMixer demonstrates superior performance compared to the state-of-the-art alternatives. Our results underline the importance of efficiently utilizing cross-variate and auxiliary information for improving the performance of time series forecasting. We present various analyses to shed light into the capabilities of TSMixer. The design paradigms utilized in TSMixer are expected to open new horizons for deep learning-based time series forecasting. The implementation is available at `https://github.com/google-research/google-research/tree/master/tsmixer`.

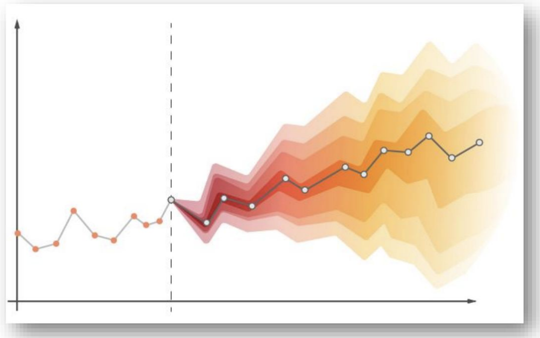
Contents

1. Preliminaries

1. Preliminaries

1. Long-term Time Series Forecasting (LTSF) task
2. Channel Dependence vs. Independence

1. Preliminaries



1. Long-term Time Series Forecasting (LTSF) task

Depends on the **“LENGTH of prediction target”**

- **LONG : Long-term Time Series Forecasting** (Widely-used task)
- SHORT : Short-term Time Series Forecasting

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
Timesteps	52696	17544	26304	966	17420	17420	69680	69680

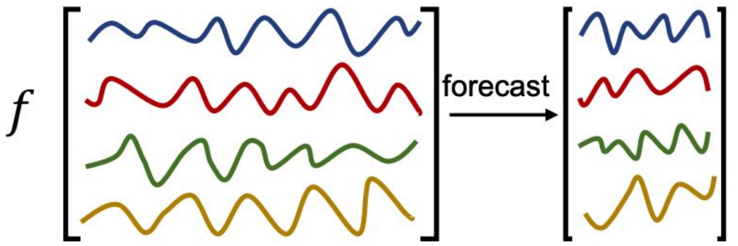
Table 2: Statistics of popular datasets for benchmark.

1. Preliminaries

2. Channel Dependence vs. Independence

Does it captures correlation between the channels (dimensions) ?

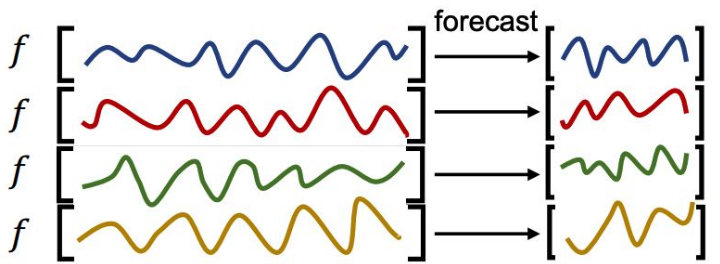
- **CD (Channel Dependence) : YES** !
- **CI (Channel Independence) : NO** !



(a) Channel Dependent (CD) Strategy

$X \in \mathbb{R}^{L \times C}$

$Y \in \mathbb{R}^{H \times C}$



(b) Channel Independent (CI) Strategy

$X \in \mathbb{R}^{L \times 1}$

$Y \in \mathbb{R}^{H \times 1}$

1. Preliminaries

2. Channel Dependence vs. Independence

Does it captures correlation between the channels (dimensions) ?

- **CD (Channel Dependence) : YES** !
- **CI (Channel Independence) : NO** !

C : Number of channels

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} \ell(f(\boldsymbol{X}^{(i)}), \boldsymbol{Y}^{(i)})$$

Loss function of CD

$$\min_{f} \frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} \ell(f(\boldsymbol{x}_c^{(i)}), \boldsymbol{y}_c^{(i)})$$

Loss function of CI

Intuitively, CD seems better!

2. Recent Trends in LTSF

2. Recent Trends in LTSF

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
Timesteps	52696	17544	26304	966	17420	17420	69680	69680

Table 2: Statistics of popular datasets for benchmark.

~ 2022 : **Transfomer-based** methods

- **Informer** (2021)
- **Autoformer** (2022)
- **FEDformer** (2022)

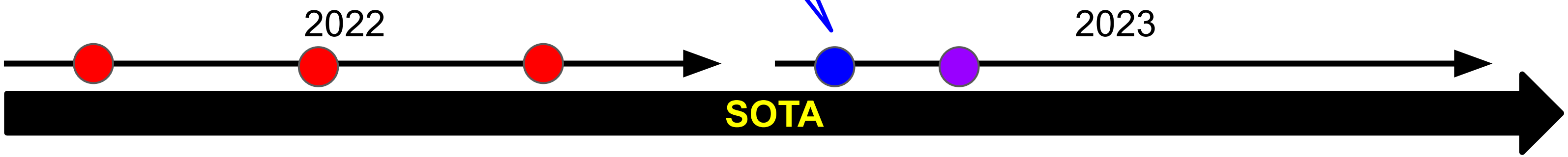
2. Recent Trends in LTSF

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
Timesteps	52696	17544	26304	966	17420	17420	69680	69680

Table 2: Statistics of popular datasets for benchmark.

Are Transformers Effective for Time Series Forecasting?

Ailing Zeng[1]*, Muxi Chen[1]*, Lei Zhang[2], Qiang Xu[1]
[1]The Chinese University of Hong Kong
[2]International Digital Economy Academy (IDEA)
{alzeng, mxchen21, qxu}@cse.cuhk.edu.hk
{leizhang}@idea.edu.cn



~ 2022 : **Transfomer-based** methods

- **Informer** (2021)
- **Autoformer** (2022)
- **FEDformer** (2022)

2023 : **Simple** / **SSL** algorithms

- **Linear models** (2023)
- **PatchTST** (2023)

2. Recent Trends in LTSF

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
							69680	69680

Channel Dependence (CD)

Channel Independence (CI)

SOTA

~ 2022 : **Transfomer-based** methods

- **Informer** (2021)
- **Autoformer** (2022)
- **FEDformer** (2022)

2023 : **Simple** / **SSL** algorithms

- **Linear models** (2023)
- **PatchTST** (2023)

2. Recent Trends in LTSF

~ 2022 : **Transfomer-based** methods

- **Informer** (2021)
- **Autoformer** (2022)
- **FEDformer** (2022)

Are Transformers Effective for Time Series Forecasting? AAAI, 2023

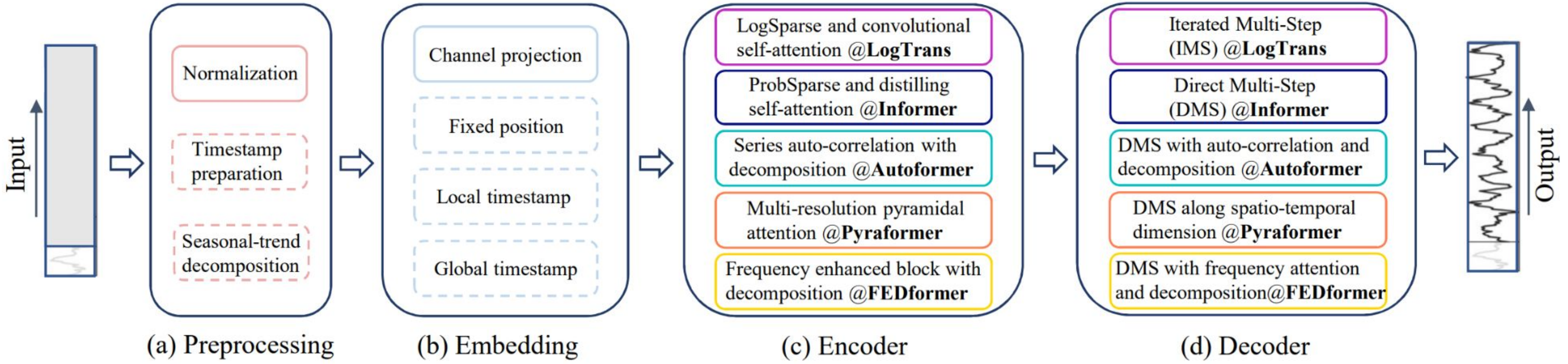


Figure 1. The pipeline of existing Transformer-based TSF solutions. In (a) and (b), the solid boxes are essential operations, and the dotted boxes are applied optionally. (c) and (d) are distinct for different methods [16, 18, 28, 30, 31].

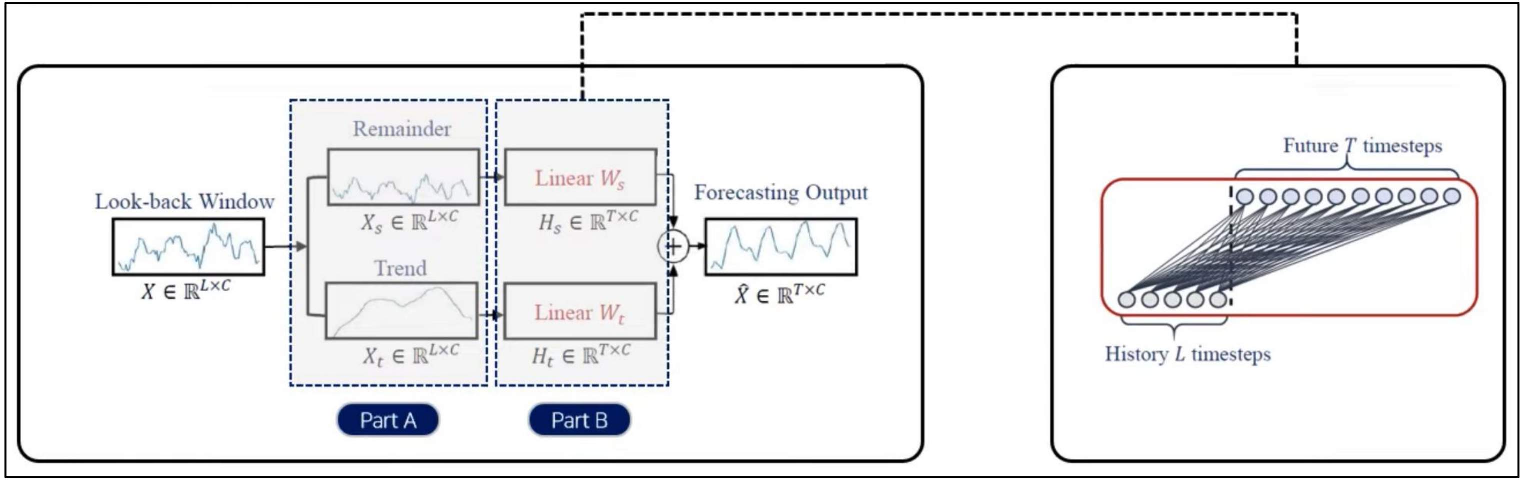
TOO complicated!

2. Recent Trends in LTSF

2023 : **Simple** / SSL algorithms

- **Linear models** (2023)
- PatchTST (2023)

Are Transformers Effective for Time Series Forecasting? AAAI, 2023



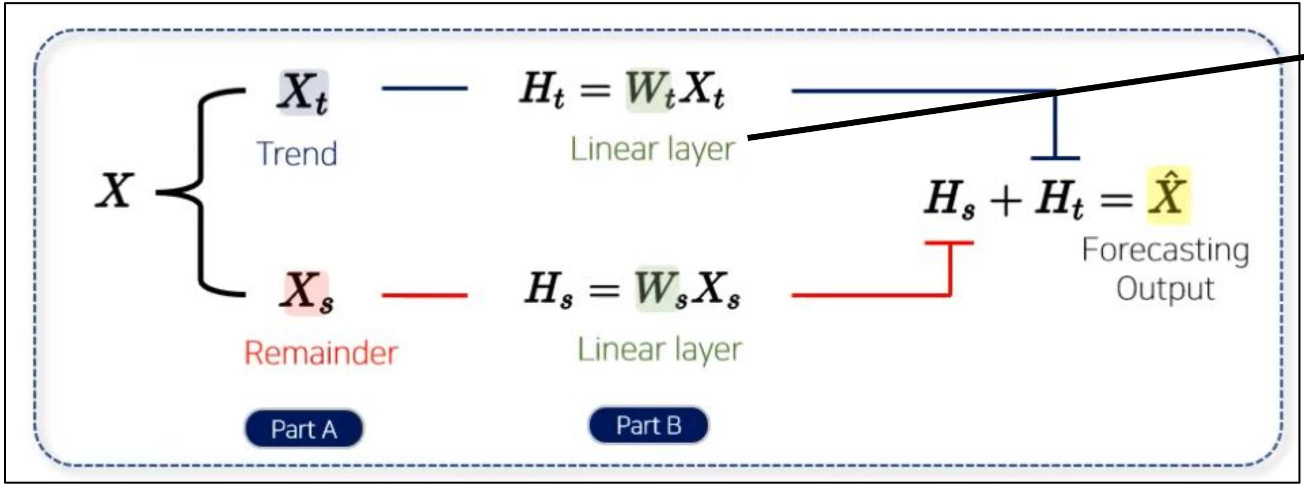
VERY SIMPLE!

2. Recent Trends in LTSF

2023 : **Simple** / SSL algorithms

- **Linear models** (2023)
- PatchTST (2023)

Are Transformers Effective for Time Series Forecasting? AAAI, 2023



Shared for all channels!

(= Channel Independence)

VERY SIMPLE!

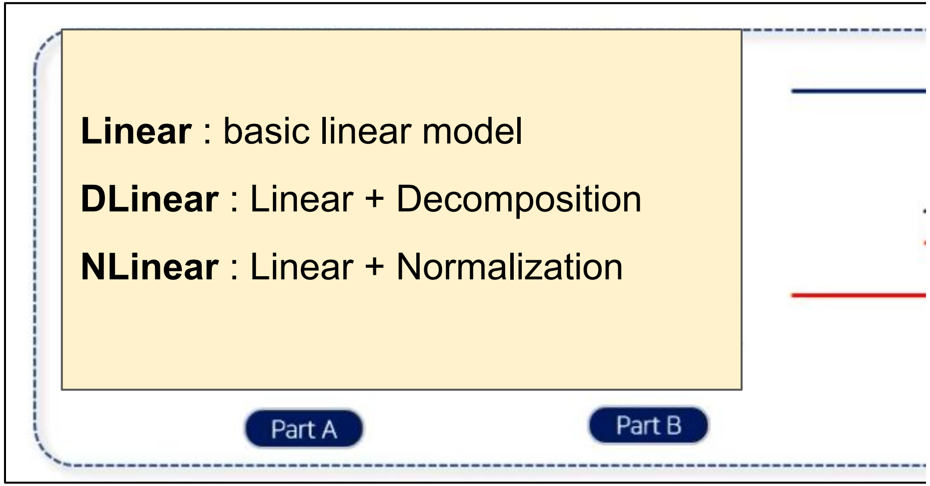
2. Recent Trends in LTSF

Are Transformers Effective for Time Series Forecasting? AAAI,

Linear : basic linear model

DLinear : Linear + Decomposition

NLinear : Linear + Normalization



	Proposed						SOTA	
Methods	IMP.	Linear*		NLinear*		DLinear*		FEDformer	
Metric	MSE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity 96	27.40%	**0.140**	**0.237**	0.141	**0.237**	**0.140**	**0.237**	0.193	0.308
Electricity 192	23.88%	**0.153**	0.250	0.154	**0.248**	**0.153**	0.249	0.201	0.315
Electricity 336	21.02%	**0.169**	0.268	0.171	**0.265**	**0.169**	0.267	0.214	0.329
Electricity 720	17.47%	**0.203**	0.301	0.210	**0.297**	**0.203**	0.301	0.246	0.355
Exchange 96	45.27%	0.082	0.207	0.089	0.208	**0.081**	0.203	0.148	0.278
Exchange 192	42.06%	0.167	0.304	0.180	0.300	**0.157**	0.293	0.271	0.380
Exchange 336	33.69%	0.328	0.432	0.331	0.415	**0.305**	0.414	0.460	0.500
Exchange 720	46.19%	0.964	0.750	1.033	0.780	**0.643**	**0.601**	1.195	0.841
Traffic 96	30.15%	**0.410**	0.282	**0.410**	**0.279**	**0.410**	0.282	0.587	0.366
Traffic 192	29.96%	**0.423**	0.287	**0.423**	**0.284**	**0.423**	0.287	0.604	0.373
Traffic 336	29.95%	0.436	0.295	**0.435**	**0.290**	0.436	0.296	0.621	0.383
Traffic 720	25.87%	0.466	0.315	**0.464**	**0.307**	0.466	0.315	0.626	0.382
Weather 96	18.89%	**0.176**	0.236	0.182	**0.232**	**0.176**	0.237	0.217	0.296
Weather 192	21.01%	**0.218**	0.276	0.225	**0.269**	0.220	0.282	0.276	0.336
Weather 336	22.71%	**0.262**	0.312	0.271	**0.301**	0.265	0.319	0.339	0.380
Weather 720	19.85%	0.326	0.365	0.338	**0.348**	**0.323**	0.362	0.403	0.428
ILI 24	47.86%	1.947	0.985	**1.683**	**0.858**	2.215	1.081	3.228	1.260
ILI 36	36.43%	2.182	1.036	**1.703**	**0.859**	1.963	0.963	2.679	1.080
ILI 48	34.43%	2.256	1.060	**1.719**	**0.884**	2.130	1.024	2.622	1.078
ILI 60	34.33%	2.390	1.104	**1.819**	**0.917**	2.368	1.096	2.857	1.157
ETTh1 96	0.80%	0.375	0.397	**0.374**	**0.394**	0.375	0.399	0.376	0.419
ETTh1 192	3.57%	0.418	0.429	0.408	**0.415**	**0.405**	0.416	0.420	0.448
ETTh1 336	6.54%	0.479	0.476	**0.429**	**0.427**	0.439	0.443	0.459	0.465
ETTh1 720	13.04%	0.624	0.592	**0.440**	**0.453**	0.472	0.490	0.506	0.507
ETTh2 96	19.94%	0.288	0.352	**0.277**	**0.338**	0.289	0.353	0.346	0.388
ETTh2 192	19.81%	0.377	0.413	**0.344**	**0.381**	0.383	0.418	0.429	0.439
ETTh2 336	25.93%	0.452	0.461	**0.357**	**0.400**	0.448	0.465	0.496	0.487
ETTh2 720	14.25%	0.698	0.595	**0.394**	**0.436**	0.605	0.551	0.463	0.474
ETTm1 96	21.10%	0.308	0.352	0.306	0.348	**0.299**	**0.343**	0.379	0.419
ETTm1 192	21.36%	0.340	0.369	0.349	0.375	**0.335**	**0.365**	0.426	0.441
ETTm1 336	17.07%	0.376	0.393	0.375	0.388	**0.369**	**0.386**	0.445	0.459
ETTm1 720	21.73%	0.440	0.435	0.433	0.422	**0.425**	**0.421**	0.543	0.490
ETTm2 96	17.73%	0.168	0.262	**0.167**	**0.255**	**0.167**	0.260	0.203	0.287
ETTm2 192	17.84%	0.232	0.308	**0.221**	**0.293**	0.224	0.303	0.269	0.328
ETTm2 336	15.69%	0.320	0.373	**0.274**	**0.327**	0.281	0.342	0.325	0.366
ETTm2 720	12.58%	0.413	0.435	**0.368**	**0.384**	0.397	0.421	0.421	0.415

- Methods* are implemented by us; Other results are from FEDformer [31].

Outperforms previous SOTA ! Are Transfor ... 023

https://www.youtube.com/watch?v=J5Pl5a_mXfE

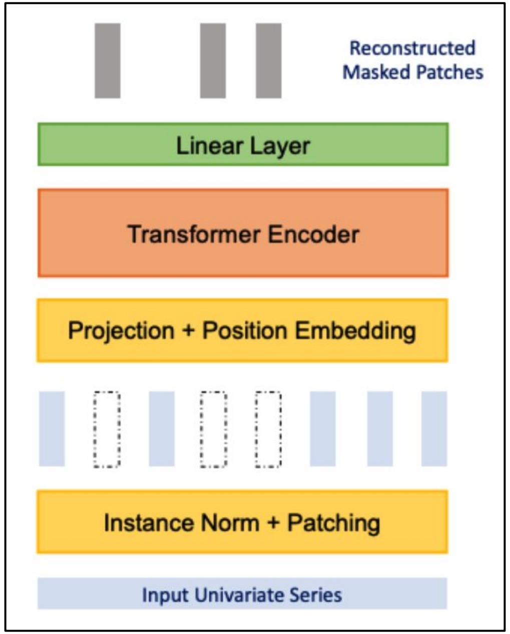
2. Recent Trends in LTSF

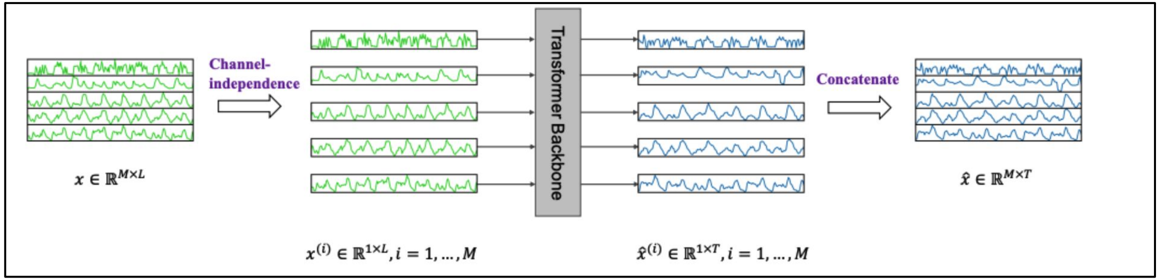
2023 : Simple / **SSL** algorithms

- Linear models (2023)
- **PatchTST (2023)**

A Time Series is Worth 64 Words : Long-term Forecasting with Transformers, ICLR 2023

PatchTST





Key Properties

(1) Channel Independence

(2) Patching time series (N time steps = 1 patch)

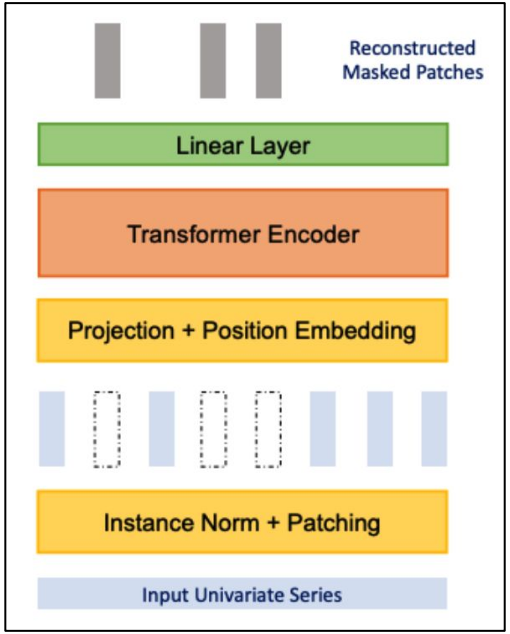
2023 : Simple / **SSL** algorithms

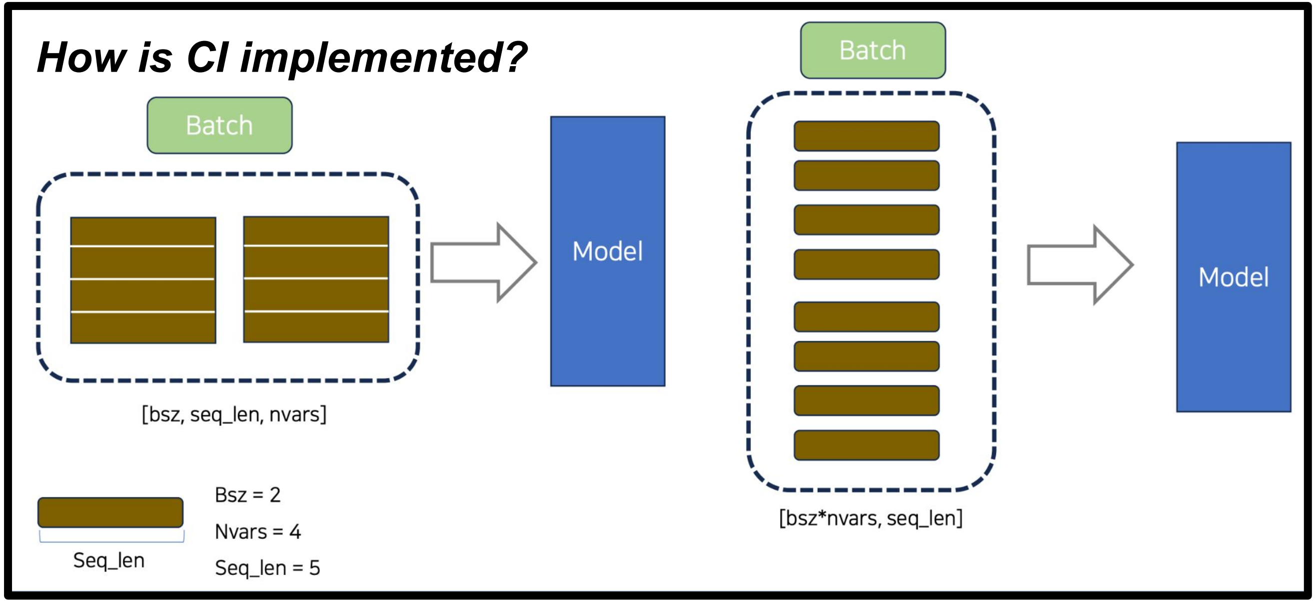
- Linear models (2023)
- **PatchTST (2023)**

2. Recent Trends in LTSF

A Time Series is Worth 64 Words : Long-term Forecasting with Transformers, ICLR 2023

PatchTST





2023 : Simple / SSL algorithms

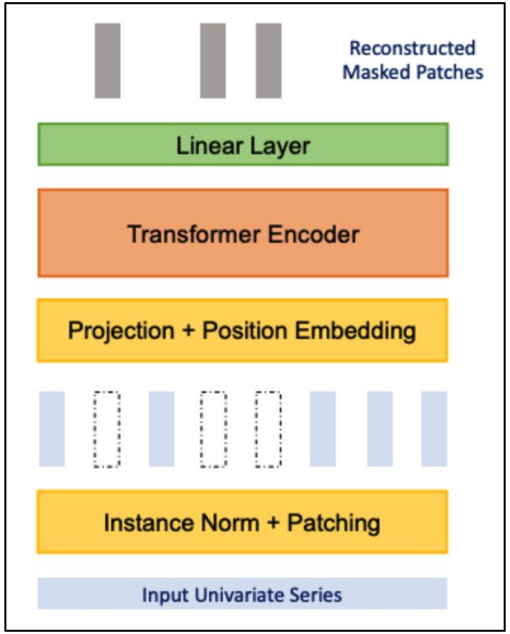
- Linear models (2023)
- PatchTST (2023)

2. Recent Trends in LTSF

A Time Series is Worth 64 Words : Long-term Forecasting with Tra

[SOTA summary (CI & CD)]

PatchTST > DLinear >>> Transformer-based



PatchTST

Models		PatchTST						DLinear		FEDformer	
		Fine-tuning		Lin. Prob.		Sup.					
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.144**	**0.193**	0.158	0.209	0.152	0.199	0.176	0.237	0.238	0.314
	192	**0.190**	**0.236**	0.203	0.249	0.197	0.243	0.220	0.282	0.275	0.329
	336	**0.244**	**0.280**	0.251	0.285	0.249	0.283	0.265	0.319	0.339	0.377
	720	**0.320**	**0.335**	0.321	0.336	**0.320**	**0.335**	0.323	0.362	0.389	0.409
Traffic	96	**0.352**	**0.244**	0.399	0.294	0.367	0.251	0.410	0.282	0.576	0.359
	192	**0.371**	**0.253**	0.412	0.298	0.385	0.259	0.423	0.287	0.610	0.380
	336	**0.381**	**0.257**	0.425	0.306	0.398	0.265	0.436	0.296	0.608	0.375
	720	**0.425**	**0.282**	0.460	0.323	0.434	0.287	0.466	0.315	0.621	0.375
Electricity	96	**0.126**	**0.221**	0.138	0.237	0.130	**0.222**	0.140	0.237	0.186	0.302
	192	**0.145**	**0.238**	0.156	0.252	0.148	0.240	0.153	0.249	0.197	0.311
	336	**0.164**	**0.256**	0.170	0.265	0.167	0.261	0.169	0.267	0.213	0.328
	720	**0.193**	**0.291**	0.208	0.297	0.202	**0.291**	0.203	0.301	0.233	0.344

3. TSMixer

3. TSMixer

Background

According to the recent trends …

can we say that **Channel Independence** **>** **Channel Dependence** **(?)**

TSMixer : ***"No! It is because of the DATASET BIAS!***

3. TSMixer

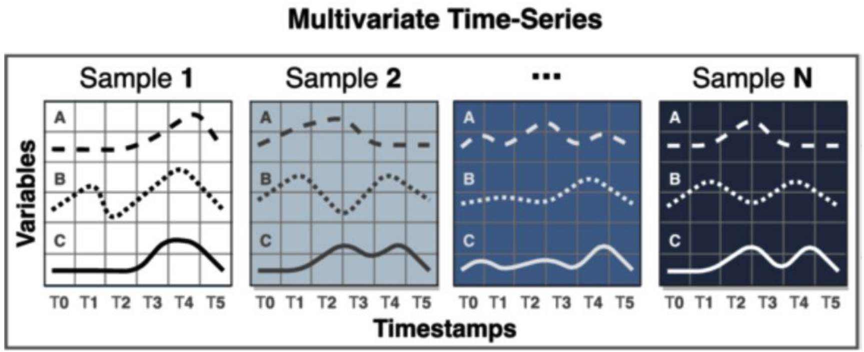
Abstract

- Investigate the **capabilities of linear models** for TS forecasting
- Present **TSMixer**
  - a novel architecture designed by **stacking MLPs**
  - based on mixing operations along both..
    - (1) **Time** dimension
    - (2) **Feature** dimension
  - Simple-to-implement

3. TSMixer

3 Major aspects of forecastability of TS

- **(1) Persistent temporal patterns**
    - trend & seasonal patterns
- **(2) Cross-variate information**
    - correlations between different variables
- **(3) Auxiliary features**
    - comprising static features and future information



3. TSMixer

- **(1) Persistent temporal patterns**
- **(2) Cross-variate information**
- **(3) Auxiliary features**

Categorization of recent TSF models

Table 1: Recent works in time series forecasting. Category I is univariate time series forecasting; Category II is multivariate time series forecasting, and Category III is time series forecasting with auxiliary information. In this work, we propose TSMixer for Category II. We also extend TSMixer to leverage auxiliary information including static and future time-varying features for Category III.

Category	Extrapolating temporal patterns	Consideration of cross-variate information (i.e. multivariateness)	Consideration of auxiliary features	Models
I	✔			ARIMA (Box et al., 1970) N-BEATS (Oreshkin et al., 2020) LTSF-Linear (Zeng et al., 2023) PatchTST (Nie et al., 2023)
II	✔	✔		Informer (Zhou et al., 2021) Autoformer (Wu et al., 2021) Pyraformer (Liu et al., 2022a) FEDformer (Zhou et al., 2022b) NS-Transformer (Liu et al., 2022b) FiLM (Zhou et al., 2022a) **TSMixer** (this work)
III	✔	✔	✔	MQRNN (Wen et al., 2017) DSSM (Rangapuram et al., 2018) DeepAR (Salinas et al., 2020) TFT (Lim et al., 2021) **TSMixer-Ext** (this work)

SOTA method only uses (1)

PROPOSED METHODS

3. TSMixer

2 Key Questions

- (1) Does **cross-variate information** truly provide a **benefit** for TS forecasting?
- (2) When **cross-variate information** is **not beneficial**, can multivariate models still perform **as well as univariate models**?

3. TSMixer

Linear Models for TS Forecasting

- a) Theoretical Insights on the **capacity of linear models**
  - have been overlooked due to its simplicity
- b) **Comparison with DL methods**

Notation

$\boldsymbol{X} \in \mathbb{R}^{L \times C_x}$: input

$\boldsymbol{Y} \in \mathbb{R}^{T \times C_y}$: target

Focus on the case where $(C_y \leq C_x)$

Linear model params : $\boldsymbol{A} \in \mathbb{R}^{T \times L}, \boldsymbol{b} \in \mathbb{R}^{T \times 1}$

- $\boxed{\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{X} \oplus \boldsymbol{b} \in \mathbb{R}^{T \times C_x}}$.
  - $\oplus$: column-wise addition

3. TSMixer

Linear Models for TS Forecasting

- a) Theoretical Insights on the **capacity of linear models**

Most impactful real-world applications have either ..

- (1) smoothness
- (2) periodicity

Assumption 1) Time series is periodic (Holt, 2004; Zhang & Qi, 2005).

(A) arbitrary **periodic function** $x(t) = x(t-P)$, where $P < L$ is the period

perfect solution :

- $\boldsymbol{A}_{ij} = \begin{cases} 1, & \text{if } j = L - P + (i \bmod P) \\ 0, & \text{otherwise} \end{cases}, \boldsymbol{b}_i = 0.$.

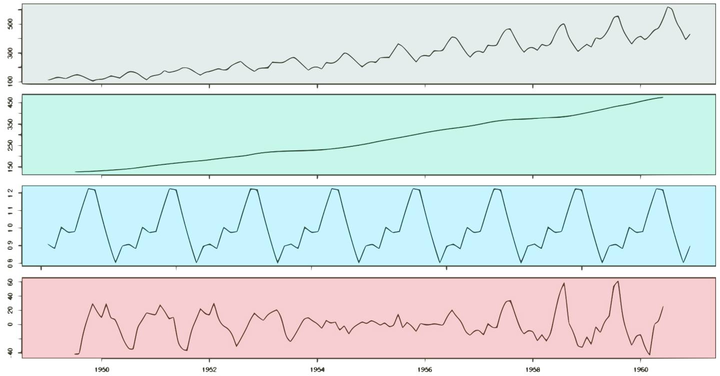
(B) affine-transformed periodic sequences, $x(t) = a \cdot x(t-P) + c$ where $a, c \in \mathbb{R}$ are constants

perfect solution :

- $\boldsymbol{A}_{ij} = \begin{cases} a, & \text{if } j = L - P + (i \bmod P) \\ 0, & \text{otherwise} \end{cases}, \boldsymbol{b}_i = c.$.

3. TSMixer

https://sthalles.github.io/assets/time-series-decomposition/cover.png



Linear Models for TS Forecasting

- a) Theoretical Insights on the capacity of linear models

Most impactful real-world applications have either ..

- (1) smoothness
- (2) periodicity

Assumption 2) Time series can be decomposed into a periodic sequence and a sequence with smooth trend

- proof in Appendix A

3. TSMixer

Linear Models for TS Forecasting

- b) Comparison with **DL methods**

Deeper insights into ***why previous DL models tend to overfit the data***

- **Linear models** = “**time-step-dependent**”
    - weights of the mapping are fixed for each time step
- **Recurrent / Attention models** = "**data-dependent**"
    - weights over the input sequence are outputs of a "data-dependent" function

3. TSMixer

Linear Models for TS Forecasting

- b) Comparison with **DL methods**

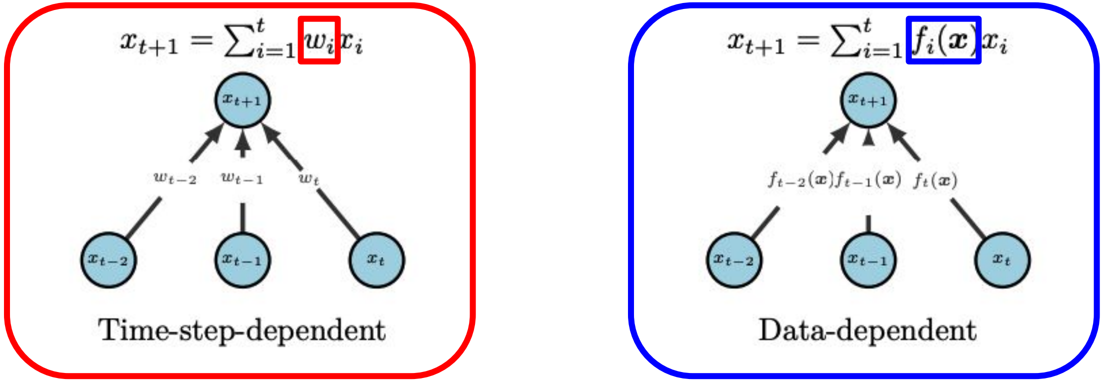


Figure 2: Illustrations of time-step-dependent and data-dependent models within a single forecasting time step.

3. TSMixer

Linear Models for TS Forecasting

- b) Comparison with **DL methods**

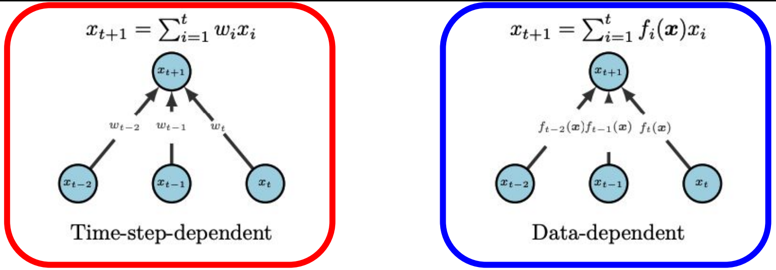


Figure 2: Illustrations of time-step-dependent and data-dependent models within a single forecasting time step.

"Time-step"-dependent linear models

- simple & highly effective in modeling temporal patterns

"Data"-dependent models

- high representational capacity

  (= achieving time-step independence is challenging)
- usually overfit on the data

  (= instead of solely considering the positions)

3. TSMixer

TSMixer Architecture

Propose a natural enhancement by ***stacking linear models with non-linearities***

Use common DL techniques

- (1) normalization
- (2) residual connections

However, this **DOES NOT consider cross-variate information**

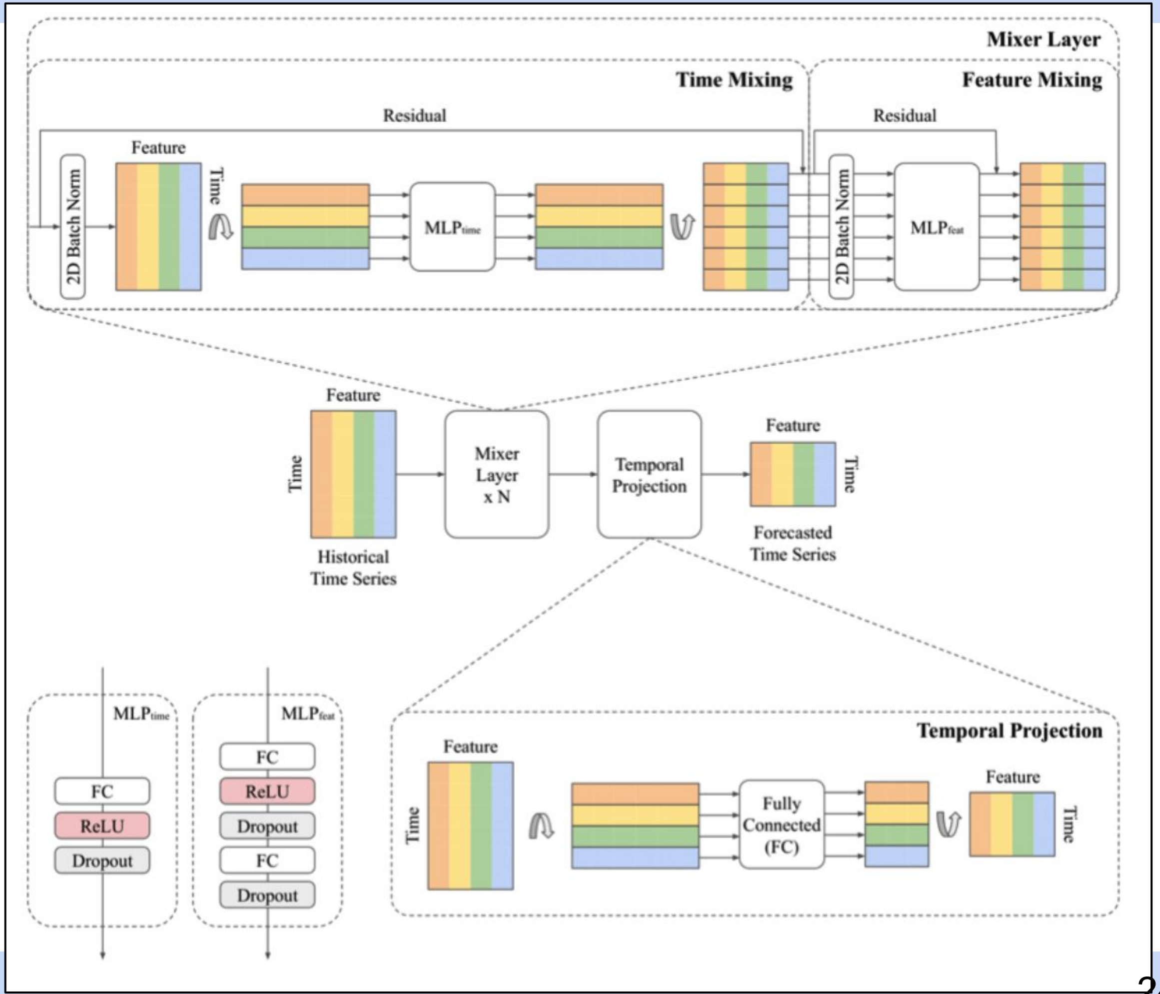
For ***cross-variate information…*** propose the application of MLPs in

- the **time-domain**
- the **feature-domain**

in an alternating manner.

3. TSMixer

TSMixer Architecture



3. TSMixer

TSMixer Architecture

Time-domain MLPs

- shared across all of the features

Feature-domain MLPs

- shared across all of the time steps.

just by transposing!

3. TSMixer

TSMixer Architecture

Time-domain MLPs

- shared across all of the features

Feature-domain MLPs

- shared across all of the time steps.

Interleaving design between these 2 operations

- efficiently utilizes both **TEMPORAL** dependencies & **CROSS-VARIATE** information

  (while limiting computational complexity and model size)
- allows to use a long lookback window
  - parameter growth in only $O(L + C)$, not $O(LC)$ if FC-MLPs were used

3. TSMixer

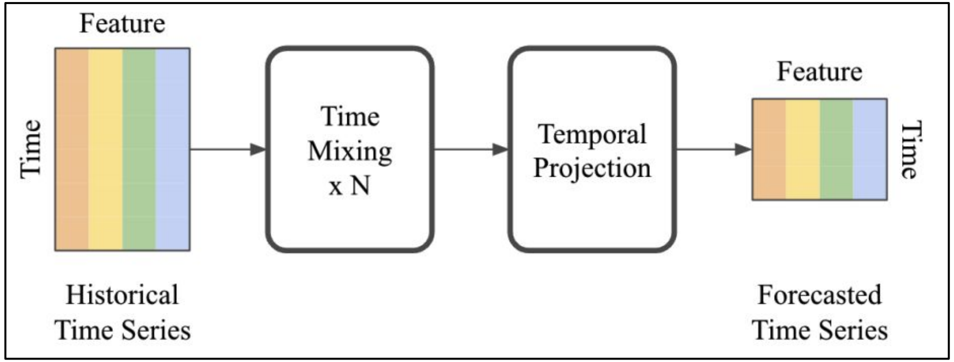
Time-domain MLPs

- shared across all of the features

TSMixer Architecture

TMix-Only : also consider a simplified variant of TSMixer

- only employs time-mixing
- consists of a residual MLP shared across each variate

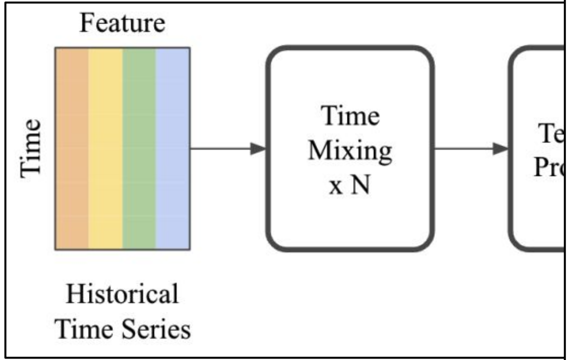


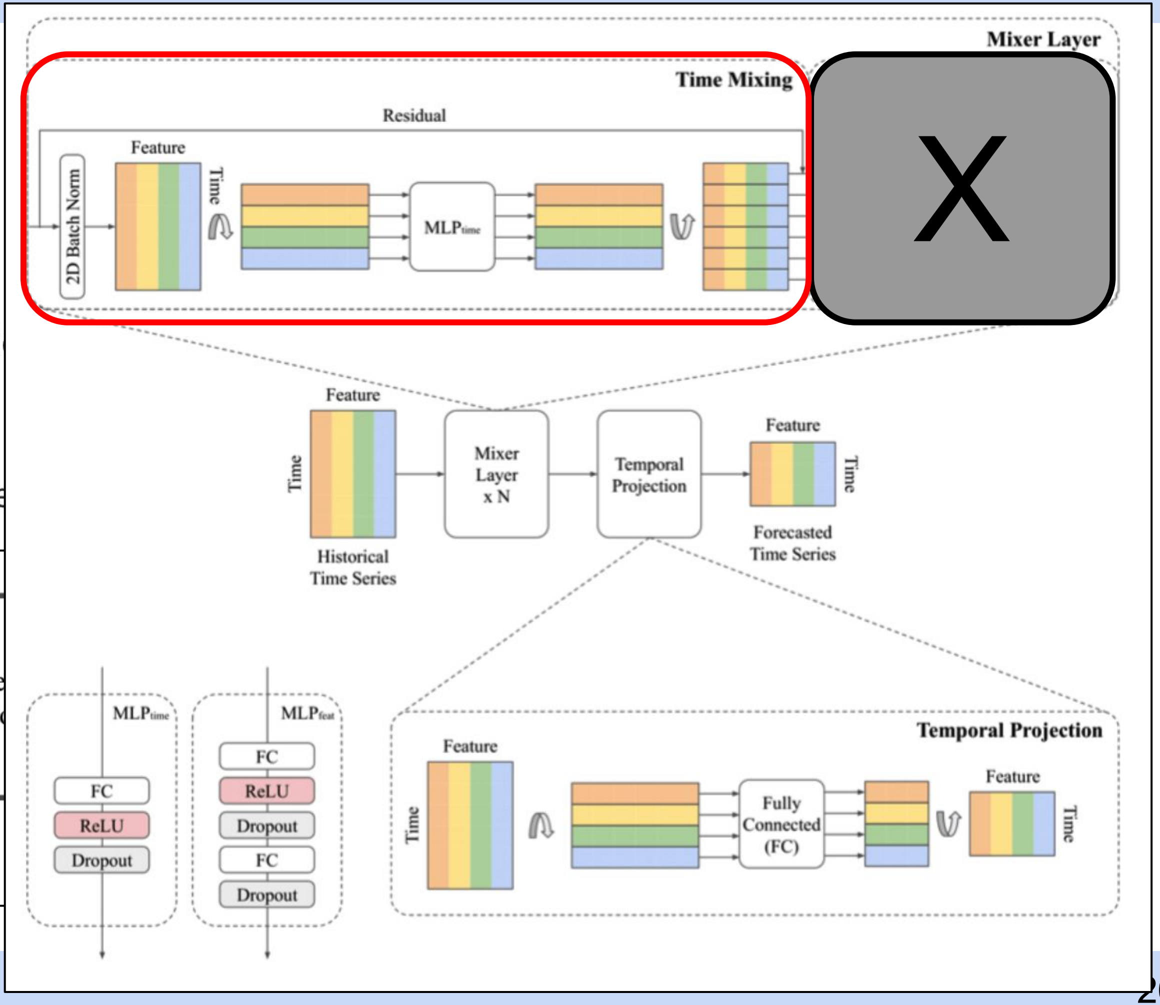
3. TSMixer

TSMixer Architecture

TMix-Only : also consider a simplifi

- only employs time-mixing
- consists of a residual MLP share





3. TSMixer

TSMixer Architecture

Notation : (N,T,C)

- N : number of TS
- T : length of TS
- C : channel(dimension) of TS

ex) 200 sensors, tracking 5 variates for 300 min.

- N = 200
- T = 300
- C = 5

```
def res_block(inputs, norm_type, activation, dropout, ff_dim):
  """Residual block of TSMixer."""

  norm = (
      layers.LayerNormalization
      if norm_type == 'L'
      else layers.BatchNormalization
  )

  # Temporal Linear
  x = norm(axis=[-2, -1])(inputs)
  x = tf.transpose(x, perm=[0, 2, 1])  # [Batch, Channel, Input Length]
  x = layers.Dense(x.shape[-1], activation=activation)(x)
  x = tf.transpose(x, perm=[0, 2, 1])  # [Batch, Input Length, Channel]
  x = layers.Dropout(dropout)(x)
  res = x + inputs

  # Feature Linear
  x = norm(axis=[-2, -1])(res)
  x = layers.Dense(ff_dim, activation=activation)(
      x
  )  # [Batch, Input Length, FF_Dim]
  x = layers.Dropout(dropout)(x)
  x = layers.Dense(inputs.shape[-1])(x)  # [Batch, Input Length, Channel]
  x = layers.Dropout(dropout)(x)
  return x + res
```

3. TSMixer

TSMixer Architecture

(N, T, C)

Time-domain weight

(N, C, T)

(N, C, T) x **(T,T)** = (N, C, T)

(N, T, C)

Feature-domain weight

(N, T, C) x **(C,D)** = (N, T, D)

(N, T, D) x **(D,C)** = (N, T, C)

```
def res_block(inputs, norm_type, activation, dropout, ff_dim):
  """Residual block of TSMixer."""

  norm = (
      layers.LayerNormalization
      if norm_type == 'L'
      else layers.BatchNormalization
  )

  # Temporal Linear
  x = norm(axis=[-2, -1])(inputs)
  x = tf.transpose(x, perm=[0, 2, 1])  # [Batch, Channel, Input Length]
  x = layers.Dense(x.shape[-1], activation=activation)(x)
  x = tf.transpose(x, perm=[0, 2, 1])  # [Batch, Input Length, Channel]
  x = layers.Dropout(dropout)(x)
  res = x + inputs

  # Feature Linear
  x = norm(axis=[-2, -1])(res)
  x = layers.Dense(ff_dim, activation=activation)(
      x
  )  # [Batch, Input Length, FF_Dim]
  x = layers.Dropout(dropout)(x)
  x = layers.Dense(inputs.shape[-1])(x)  # [Batch, Input Length, Channel]
  x = layers.Dropout(dropout)(x)
  return x + res
```

4. TSMixer-Ext

4. TSMixer-Ext

3 Major aspects of forecastability of TS

- **(1) Persistent temporal patterns**
- **(2) Cross-variate information**
- **(3) Auxiliary features**

Extended TSMixer with **Auxiliary Information**

Real-world scenarios : access to ..

- (1) static features : $\boldsymbol{S} \in \mathbb{R}^{1 \times C_s}$
- (2) future time-varying features : $\boldsymbol{Z} \in \mathbb{R}^{T \times C_z}$

$\rightarrow$ extended to multiple TS, represented by $\boldsymbol{X}^{(i)}{}_{i=1}^{N}$,

- N : number of TS

4. TSMixer-Ext

3 Major aspects of forecastability of TS

- (1) Persistent temporal patterns
- (2) Cross-variate information
- (3) Auxiliary features

Extended TSMixer with **Auxiliary Information**

Long-term forecasting

- (In general) Only consider the historical features & targets on all variables (i.e. $C_x = C_y > 1, \boxed{C_s = C_z = 0}$).
- (In this paper) Also consider the case where auxiliary information is available (i.e. $\boxed{C_s > 0, C_z > 0}$).

4. TSMixer-Ext

Architecture

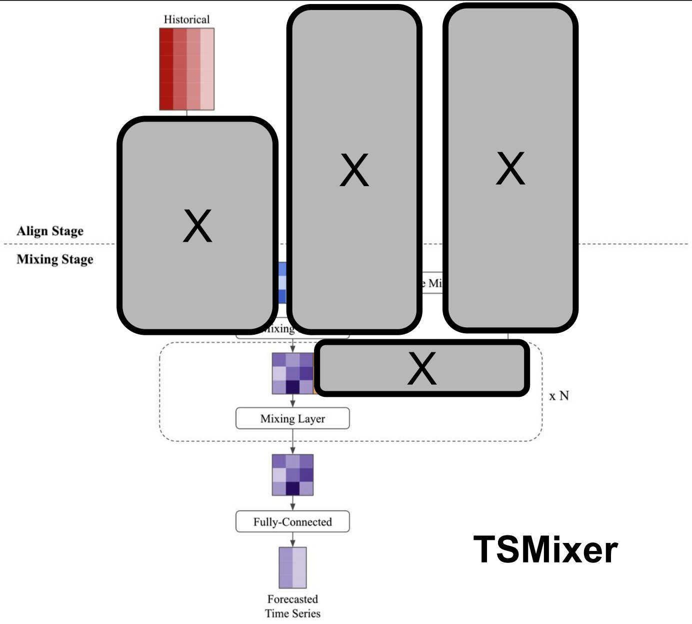


Figure 4: TSMixer with auxiliary information. The columns of the inputs are features and the rows are time steps. We first align the sequence lengths of different types of inputs to concatenate them. Then we apply mixing layers to model their temporal patterns and cross-variate information jointly.

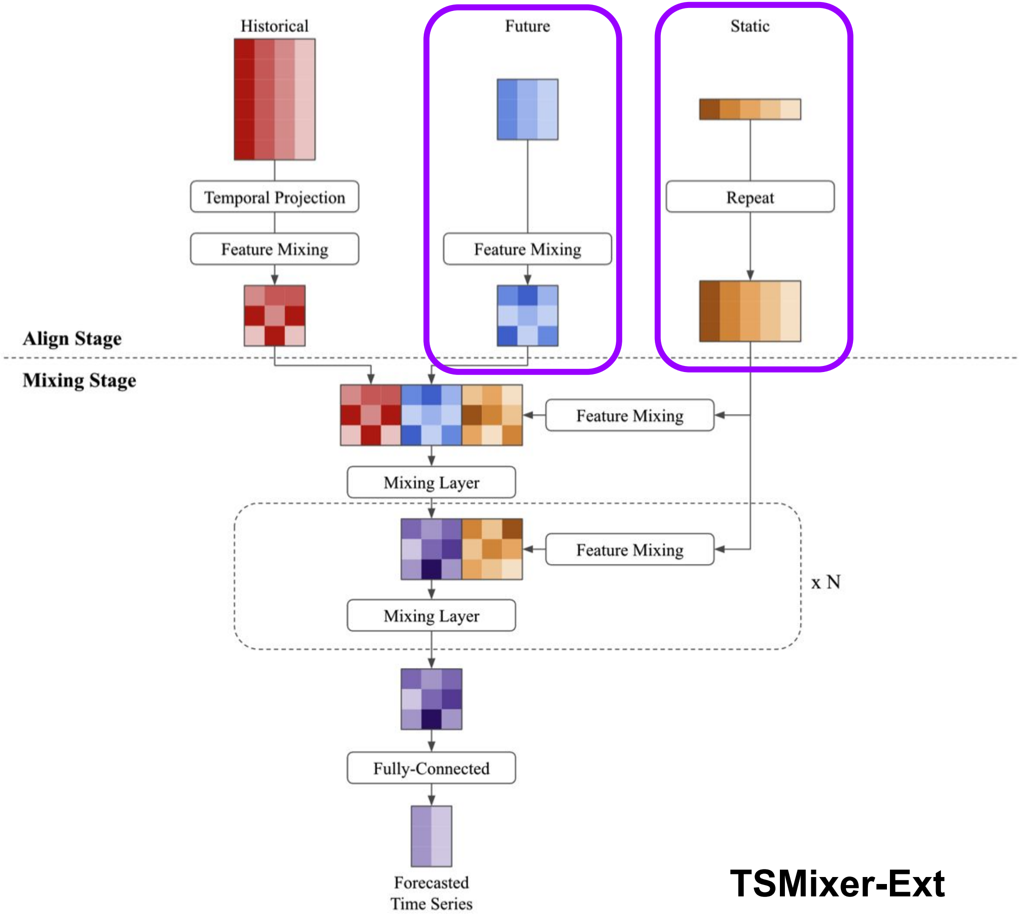


Figure 4: TSMixer with auxiliary information. The columns of the inputs are features and the rows are time steps. We first align the sequence lengths of different types of inputs to concatenate them. Then we apply mixing layers to model their temporal patterns and cross-variate information jointly.

4. TSMixer-Ext

Architecture

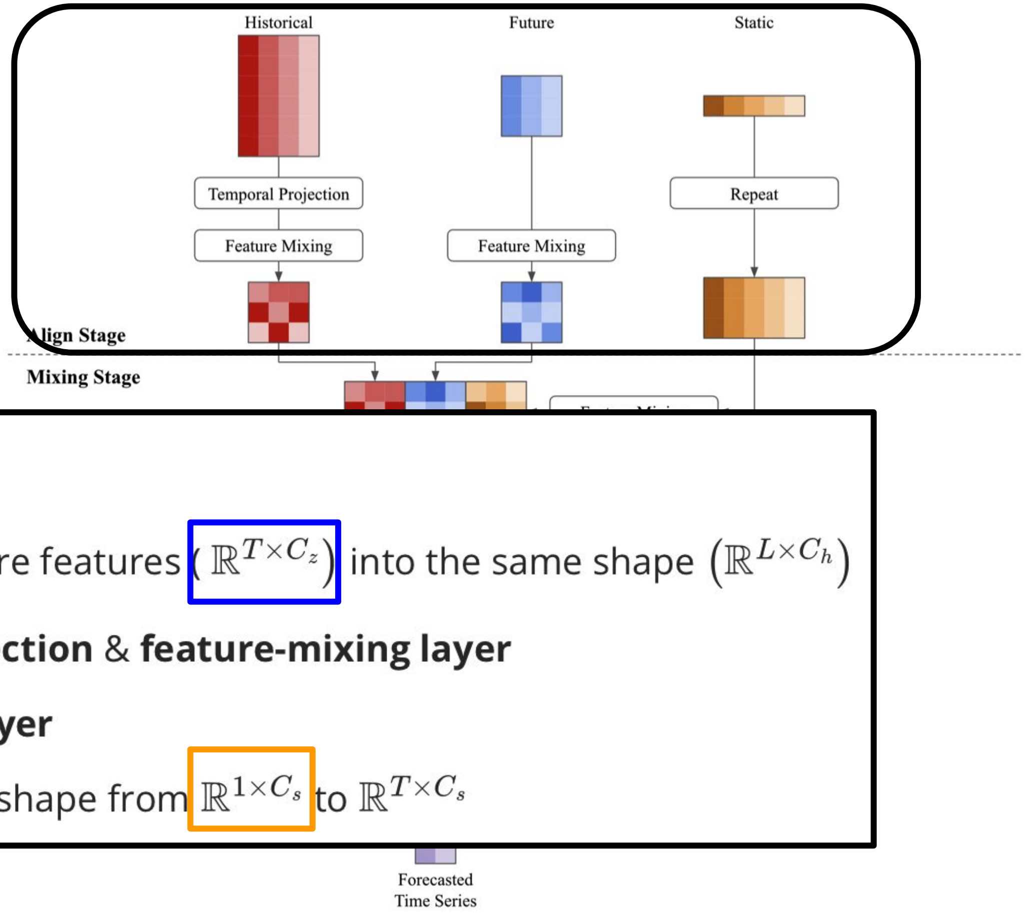


Figure 4: TSMixer with auxiliary information. The columns of the inputs are features and the rows are time steps. We first align the sequence lengths of different types of inputs to concatenate them. Then we apply mixing layers to model their temporal patterns and cross-variate information jointly.

a) Align

aligns historical features $(\mathbb{R}^{L\times C_x})$ and future features $(\mathbb{R}^{T\times C_z})$ into the same shape $(\mathbb{R}^{L\times C_h})$

- [Historical input] apply **temporal projection** & **feature-mixing layer**
- [Future input] apply **feature-mixing layer**
- [Static input] repeat to transform their shape from $\mathbb{R}^{1\times C_s}$ to $\mathbb{R}^{T\times C_s}$

4. TSMixer-Ext

Architecture

b) Mixing

Mixing layer

- time-mixing & feature-mixing operations
- leverages temporal patterns and cross-variate information from all features collectively.

FC layer

- to generate outputs for each time step.
- slightly modify mixing layers to better handle M5 dataset (described in Appendix B)

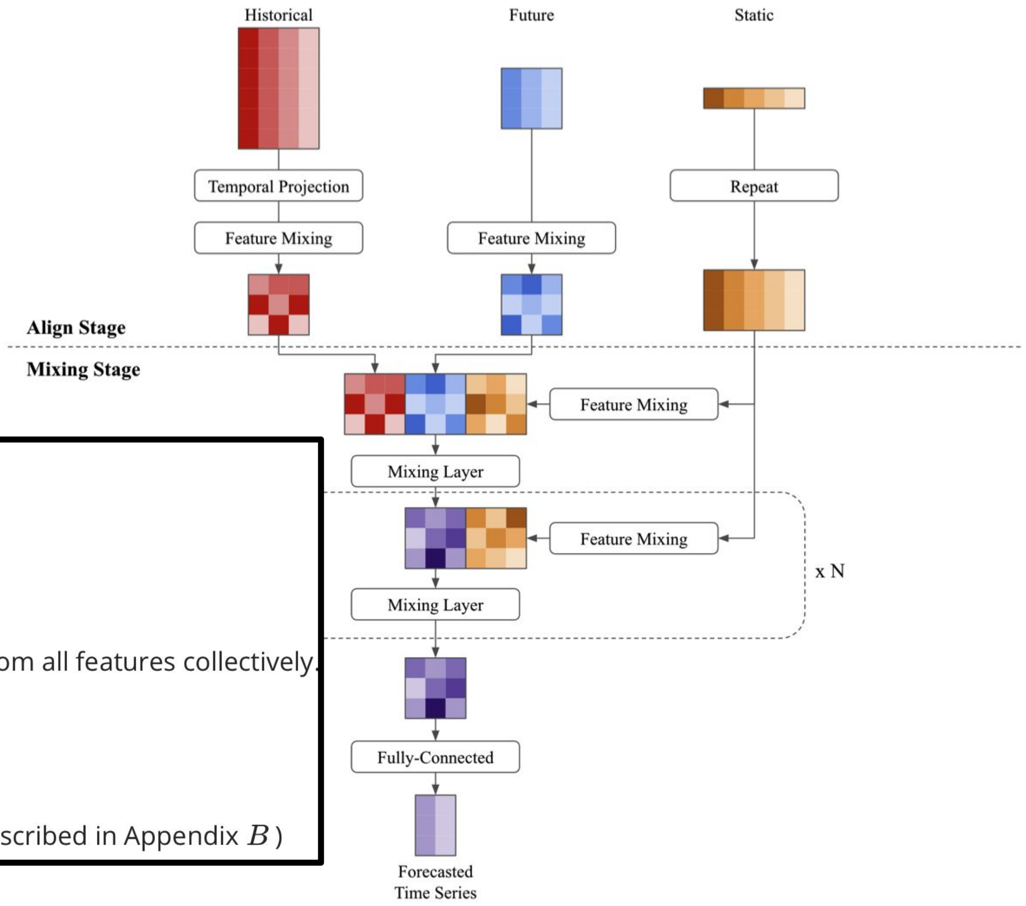


Figure 4: TSMixer with auxiliary information. The columns of the inputs are features and the rows are time steps. We first align the sequence lengths of different types of inputs to concatenate them. Then we apply mixing layers to model their temporal patterns and cross-variate information jointly.

5. Experiments

5. Experiments

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

5. Experiments

LTSF datasets

Real-world Retail dataset

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

Auxiliary Features

single MULTIVARIATE time series

multiple MULTIVARIATE time series

5. Experiments

Real-world Retail dataset

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series ... ysical meaning in the dataset (e.g. electr...

	...city	Traffic	M5
#		1	30,490
#	.	862	1
Ti	04	17,544	1,942
Gr	ur	1 hour	1day
Hi		0	14
Fu		0	13
St		0	6
Da (T			1886/28/28 (day)

Auxiliary Features

- static features: store locations …
- time-varying features: campaign information ..

=> **more challenging benchmark** to explore the potential benefits of **cross-variate information** and **auxiliary features** (compared to LTSF task)

Auxiliary Features

multiple MULTIVARIATE time series

5. Experiments

(1) LTSF task

Table 3: Eval
are obtained f
CD
CI
numbers are underlined.

Models		Multivariate Models: TSMixer		TFT		FEDformer*		Autoformer*		Informer*		Univariate Models: TMix-Only		Linear		PatchTST*	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.361	0.392	0.674	0.634	0.376	0.415	0.435	0.446	0.941	0.769	**0.359**	**0.391**	0.368	0.392	0.370	0.400
	192	0.404	0.418	0.858	0.704	0.423	0.446	0.456	0.457	1.007	0.786	**0.402**	0.415	0.404	**0.415**	0.413	0.429
	336	**0.420**	**0.431**	0.900	0.731	0.444	0.462	0.486	0.487	1.038	0.784	0.420	0.434	0.436	0.439	0.422	0.440
	720	0.463	0.472	0.745	0.666	0.469	0.492	0.515	0.517	1.144	0.857	0.453	**0.467**	0.481	0.495	**0.447**	0.468
ETTh2	96	0.274	0.341	0.409	0.505	0.332	0.374	0.332	0.368	1.549	0.952	0.275	0.342	0.297	0.363	**0.274**	**0.337**
	192	0.339	0.385	0.953	0.651	0.407	0.446	0.426	0.434	3.792	1.542	**0.339**	0.386	0.398	0.429	0.341	**0.382**
	336	0.361	0.406	1.006	0.709	0.400	0.447	0.477	0.479	4.215	1.642	0.366	0.413	0.500	0.491	**0.329**	**0.384**
	720	0.445	0.470	1.187	0.816	0.412	0.469	0.453	0.490	3.656	1.619	0.437	0.465	0.795	0.633	**0.379**	**0.422**
ETTm1	96	0.285	0.339	0.752	0.626	0.326	0.390	0.510	0.492	0.626	0.560	**0.284**	**0.338**	0.303	0.346	0.293	0.346
	192	0.327	0.365	0.752	0.649	0.365	0.415	0.514	0.495	0.725	0.619	**0.324**	**0.362**	0.335	0.365	0.333	0.370
	336	**0.356**	**0.382**	0.810	0.674	0.392	0.425	0.510	0.492	1.005	0.741	0.359	0.384	0.365	0.384	0.369	0.392
	720	0.419	**0.414**	0.849	0.695	0.446	0.458	0.527	0.493	1.133	0.845	0.419	0.414	0.419	0.415	**0.416**	0.420
ETTm2	96	0.163	0.252	0.386	0.472	0.180	0.271	0.205	0.293	0.355	0.462	**0.162**	**0.249**	0.170	0.266	0.166	0.256
	192	**0.216**	**0.290**	0.739	0.626	0.252	0.318	0.278	0.336	0.595	0.586	0.220	0.293	0.236	0.317	0.223	0.296
	336	**0.268**	**0.324**	0.477	0.494	0.324	0.364	0.343	0.379	1.270	0.871	0.269	0.326	0.308	0.369	0.274	0.329
	720	0.420	0.422	0.523	0.537	0.410	0.420	0.414	0.419	3.001	1.267	**0.358**	**0.382**	0.435	0.449	0.362	0.385
Weather	96	0.145	0.198	0.441	0.474	0.238	0.314	0.249	0.329	0.354	0.405	**0.145**	**0.196**	0.170	0.229	0.149	0.198
	192	0.191	0.242	0.699	0.599	0.275	0.329	0.325	0.370	0.419	0.434	**0.190**	**0.240**	0.213	0.268	0.194	0.241
	336	0.242	0.280	0.693	0.596	0.339	0.377	0.351	0.391	0.583	0.543	**0.240**	**0.279**	0.257	0.305	0.245	0.282
	720	0.320	0.336	1.038	0.753	0.389	0.409	0.415	0.426	0.916	0.705	0.325	0.339	0.318	0.356	**0.314**	**0.334**
Electricity	96	0.131	0.229	0.295	0.376	0.186	0.302	0.196	0.313	0.304	0.393	0.132	0.225	0.135	0.232	**0.129**	**0.222**
	192	0.151	0.246	0.327	0.397	0.197	0.311	0.211	0.324	0.327	0.417	0.152	0.243	0.149	0.246	**0.147**	**0.240**
	336	**0.161**	0.261	0.298	0.380	0.213	0.328	0.214	0.327	0.333	0.422	0.166	0.260	0.164	0.263	0.163	**0.259**
	720	0.197	0.293	0.338	0.412	0.233	0.344	0.236	0.342	0.351	0.427	0.200	0.291	0.199	0.297	**0.197**	**0.290**
Traffic	96	0.376	0.264	0.678	0.362	0.576	0.359	0.597	0.371	0.733	0.410	0.370	0.258	0.395	0.274	**0.360**	**0.249**
	192	0.397	0.277	0.664	0.355	0.610	0.380	0.607	0.382	0.777	0.435	0.390	0.268	0.406	0.279	**0.379**	**0.256**
	336	0.413	0.290	0.679	0.354	0.608	0.375	0.623	0.387	0.776	0.434	0.404	0.276	0.416	0.286	**0.392**	**0.264**
	720	0.444	0.306	0.610	0.326	0.621	0.375	0.639	0.395	0.827	0.466	0.443	0.297	0.454	0.308	**0.432**	**0.286**

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

5. Experiments

(1) LTSF task

Table 3: Eval... are obtained f... numbers are underlined.

CD

CI

Models		Multivariate Models										Univariate Models					
		TSMixer		TFT		EDformer*		Autoformer*		Informer*		TMix-Only		Linear		PatchTST*	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.361	0.392	.674	0.634	.376	0.415	0.435	0.446	0.941	0.76	**0.359**	**0.391**	.368	0.392	.370	0.400
	192	0.404	0.418	.858	0.704	.423	0.446	0.456	0.457	1.007	0.78	**0.402**	0.415	.404	**0.415**	.413	0.429
	336	**0.420**	**0.431**	.900	0.731	.444	0.462	0.486	0.487	1.038	0.78	0.420	0.434	.436	0.439	.422	0.440
	720	0.463	0.472	.745	0.666	.469	0.492	0.515	0.517	1.144	0.85	0.453	**0.467**	.481	0.495	**.447**	0.468
ETTh2	96	0.274	0.341	.409	0.505	.332	0.374	0.332	0.368	1.549	0.95	0.275	0.342	.297	0.363	**.274**	**0.337**
	192	0.339	0.385	.953	0.651	.407	0.446	0.426	0.434	3.792	1.54	**0.339**	0.386	.398	0.429	.341	**0.382**
	336	0.361	0.406	.006	0.709	.400	0.447	0.477	0.479	4.215	1.64	0.366	0.413	.500	0.491	**.329**	**0.384**
	720	0.445	0.470	.187	0.816	.412	0.469	0.453	0.490	3.656	1.61	0.437	0.465	.795	0.633	**.379**	**0.422**
ETTm1	96	0.285	0.339	.752	0.626	.326	0.390	0.510	0.492	0.626	0.56	**0.284**	**0.338**	.303	0.346	.293	0.346
	192	0.327	0.365	.752	0.649	.365	0.415	0.514	0.495	0.725	0.61	**0.324**	**0.362**	.335	0.365	.333	0.370
	336	**0.356**	**0.382**	.810	0.674	.392	0.425	0.510	0.492	1.005	0.74	0.359	0.384	.365	0.384	.369	0.392
	720	0.419	**0.414**	.849	0.695	.446	0.458	0.527	0.493	1.133	0.84	0.419	0.414	.419	0.415	**.416**	0.420
ETTm2	96	0.163	0.252	.386	0.472	.180	0.271	0.205	0.293	0.355	0.46	**0.162**	**0.249**	.170	0.266	.166	0.256
	192	**0.216**	**0.290**	.739	0.626	.252	0.318	0.278	0.336	0.595	0.58	0.220	0.293	.236	0.317	.223	0.296
	336	**0.268**	**0.324**	.477	0.494	.324	0.364	0.343	0.379	1.270	0.87	0.269	0.326	.308	0.369	.274	0.329
	720	0.420	0.422	.523	0.537	.410	0.420	0.414	0.419	3.001	1.26	**0.358**	**0.382**	.435	0.449	.362	0.385
Weather	96	0.145	0.198	.441	0.474	.238	0.314	0.249	0.329	0.354	0.40	**0.145**	**0.196**	.170	0.229	.149	0.198
	192	0.191	0.242	.699	0.599	.275	0.329	0.325	0.370	0.419	0.43	**0.190**	**0.240**	.213	0.268	.194	0.241
	336	0.242	0.280	.693	0.596	.339	0.377	0.351	0.391	0.583	0.54	**0.240**	**0.279**	.257	0.305	.245	0.282
	720	0.320	0.336	.038	0.753	.389	0.409	0.415	0.426	0.916	0.70	0.325	0.339	.318	0.356	**.314**	**0.334**
Electricity	96	0.131	0.229	.295	0.376	.186	0.302	0.196	0.313	0.304	0.39	0.132	0.225	.135	0.232	**.129**	**0.222**
	192	0.151	0.246	.327	0.397	.197	0.311	0.211	0.324	0.327	0.41	0.152	0.243	.149	0.246	**.147**	**0.240**
	336	**0.161**	0.261	.298	0.380	.213	0.328	0.214	0.327	0.333	0.42	0.166	0.260	.164	0.263	.163	**0.259**
	720	0.197	0.293	.338	0.412	.233	0.344	0.236	0.342	0.351	0.42	0.200	0.291	.199	0.297	**.197**	**0.290**
Traffic	96	0.376	0.264	.678	0.362	.576	0.359	0.597	0.371	0.733	0.41	0.370	0.258	.395	0.274	**.360**	**0.249**
	192	0.397	0.277	.664	0.355	.610	0.380	0.607	0.382	0.777	0.43	0.390	0.268	.406	0.279	**.379**	**0.256**
	336	0.413	0.290	.679	0.354	.608	0.375	0.623	0.387	0.776	0.43	0.404	0.276	.416	0.286	**.392**	**0.264**
	720	0.444	0.306	.610	0.326	.621	0.375	0.639	0.395	0.827	0.46	0.443	0.297	.454	0.308	**.432**	**0.286**

1 2 3 4 5 6

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

	CD vs CI	SL vs. SSL	Architecture
1	CD	SL	MLP
2		SSL	Transformer
3		SL	
4	CI	SL	MLP
5			Linear
6		SSL	Transformer

SL : supervised learning & SSL : self-supervised learning

GOOD & BAD

5. Experiments

(1) LTSF task

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

Table 3: Eval... are obtained ... numbers are underlined.

CD

CI

Models	Multivariate Models								Univariate Models					
Metric														
ETTh1														
ETTh2														
ETTm1														
ETTm2														
Weather														
Electricity														
Traffic 336	0.413	0.290	.679	0.354	.608	0.375	0.623	0.387	0.776	0.43	0.404	0.276	.416	0.286	.392	0.264
Traffic 720	0.444	0.306	.610	0.326	.621	0.375	0.639	0.395	0.827	0.46	0.443	0.297	.454	0.308	.432	0.286

1 2 3 4 5 6

	CD vs CI	SL vs SSL	Architecture
			MLP
			...nsformer
			MLP
			...inear
		SSL	Transformer

CI > CD seems to be better for **LTSF task**!

(& TSOnly > TSMixer)

Then, can we conclude that CI > CD ??

5. Experiments

(2) M5 competition

(w.o Auxiliary Information)

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

Table 4: Evaluation on M5 without auxiliary information. We report the mean and standard deviation of WRMSSE across 5 different random seeds. TMix-Only is a univariate variant of TSMixer where only time-mixing is applied. The multivariate models outperforms univariate models with a significant gap.

	Models	Multivariate	Test WRMSSE	Val WRMSSE
CI	Linear		0.983±0.016	1.045±0.018
	PatchTST		0.976±0.014	0.992±0.011
	TMix-Only		0.960±0.041	1.000±0.027
CD	FEDformer	✔	0.804±0.039	0.674±0.014
	TSMixer	✔	**0.737±0.033**	0.605±0.027

5. Experiments

(2) M5 competition

(w.o Auxiliary Information)

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

Table 4: Evaluation on M5 without auxiliary information. We report the mean and standard deviation of WRMSSE across 5 different random seeds. TMix-Only is a univariate variant of TSMixer where only time-mixing is applied. The multivariate models outperforms univariate models with a significant gap.

	Models	Multivariate	Test WRMSSE	Val WRMSSE
CI	Linear		0.983±0.016	1.045±0.018
	PatchTST		0.976±0.014	0.992±0.011
	TMix-Only		0.960±0.041	1.000±0.027
CD	FEDformer	✔	0.804±0.039	0.674±0.014
	TSMixer	✔	**0.737±0.033**	0.605±0.027

Summary :

- LTSF : CI > CD
- M5 : CD > CI

Previous works : CI > CD

TSMixer : ***It depends on the dataset!***

Table 2: Statistics of all datasets. Note that Electricity and Traffic can be considered as multivariate time series or multiple univariate time series since all variates share the same physical meaning in the dataset (e.g. electricity consumption at different locations).

	ETTh1/h2	ETTm1/m2	Weather	Electricity	Traffic	M5
# of time series (M)	1	1	1	1	1	30,490
# of variants (C)	7	7	21	321	862	1
Time steps	17,420	699,680	52,696	26,304	17,544	1,942
Granularity	1 hour	15 minutes	10 minutes	1 hour	1 hour	1day
Historical feature (C_x)	0	0	0	0	0	14
Future feature (C_z)	0	0	0	0	0	13
Static feature (C_s)	0	0	0	0	0	6
Data partition (Train/Validation/Test)	12/4/4 (month)		7:2:1			1886/28/28 (day)

5. Experiments

(2) M5 competition

Effect of Auxiliary Information

Models	Auxiliary feature		Test WRMSSE	Val WRMSSE
	Static	Future		
DeepAR	✔	✔	0.789±0.025	0.611±0.007
TFT	✔	✔	0.670±0.020	0.579±0.011
TSMixer-Ext			0.737±0.033	0.000±0.000
	✔		0.657±0.046	0.000±0.000
		✔	0.697±0.028	0.000±0.000
	✔	✔	**0.640**±0.013	0.568±0.009

Conclusion

1. (Architecture) **Transformer** may not be necessary! (**Linear / MLP** might be enough for TS modeling)

2. (Framework) **Channel Independence** is not always the best! It depends on the dataset!

3. Desirable properties of TSF model

= "**automatically** detects the need for capturing the **cross-variate relationships**"

4. Limitations of TSMixer

- Which model to choose, TSMixer vs. TMix-Only?
- Hyperparameter tuning

  (Different experimental settings for each prediction length & dataset (96, 192, 336, 720))

Thank You!