

# TSMixer : An All-MLP Architecture for Time Series Fore-casting

<https://arxiv.org/pdf/2303.06053>

## 0. Introduction

- 시계열 예측은 전력, 교통, 금융, 날씨 등 다양한 산업 분야에서 핵심 문제임
- 최근 Transformer 기반 모델이 높은 성능을 보이지만 계산 비용과 구조 복잡도가 큼
- 긴 시계열을 처리할수록 메모리 사용량과 학습 비용이 크게 증가함
- 이미지 분야에서는 attention 없이 MLP 기반 구조가 경쟁력 있는 성능을 보임
- 본 논문은 시계열 예측에서도 단순한 MLP 구조로 충분한 성능을 낼 수 있는지 검증함
- 시간 정보와 변수 간 관계를 효율적으로 학습하는 All-MLP 구조를 제안함

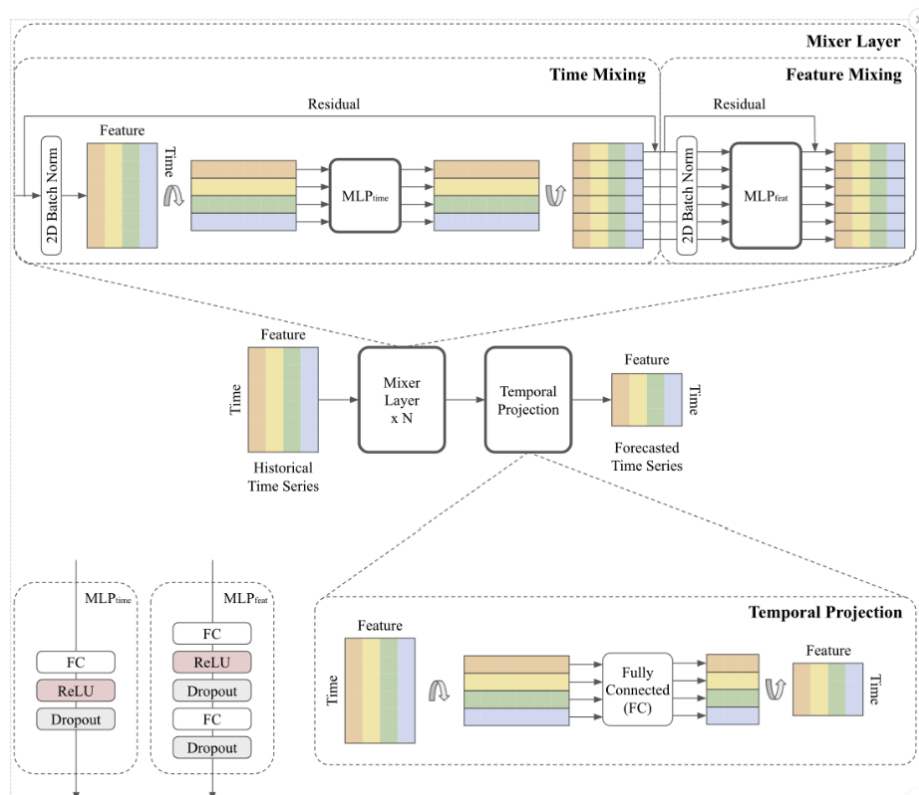
## 1. Overview

- 제안 모델은 **TSMixer**라는 All-MLP 기반 시계열 예측 구조임
- 핵심 아이디어는 시간 축과 변수 축 정보를 각각 독립적으로 혼합하는 구조임
- attention 대신 MLP mixing을 사용하여 계산 비용을 줄임
- 여러 공개 시계열 데이터셋에서 장기 예측 실험 수행
- 기존 Transformer 계열 모델과 성능 및 효율을 비교 평가함
- 목표는 단순한 구조로 실무 적용 가능한 예측 모델을 제시하는 것임

## 2. Challenges

- 시계열 데이터는 시간 흐름과 변수 간 상관관계를 동시에 고려해야 함
- 긴 입력 시퀀스에서는 기존 attention 기반 모델의 계산량이 급증함
- 복잡한 모델 구조는 실제 서비스 환경에 적용하기 어려움
- 변수 수가 많을 경우 모델이 관계를 효과적으로 학습하기 어려움
- 단순한 MLP 구조가 이런 문제를 충분히 해결할 수 있는지 검증이 필요함

### 3. Method



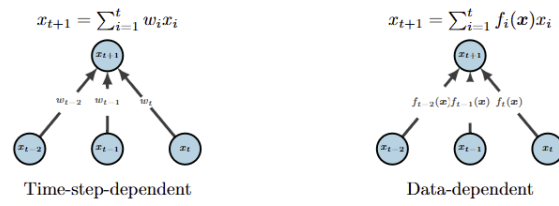


Figure 2: Illustrations of time-step-dependent and data-dependent models within a single forecasting time step.

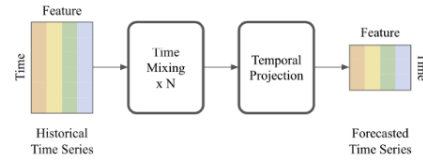
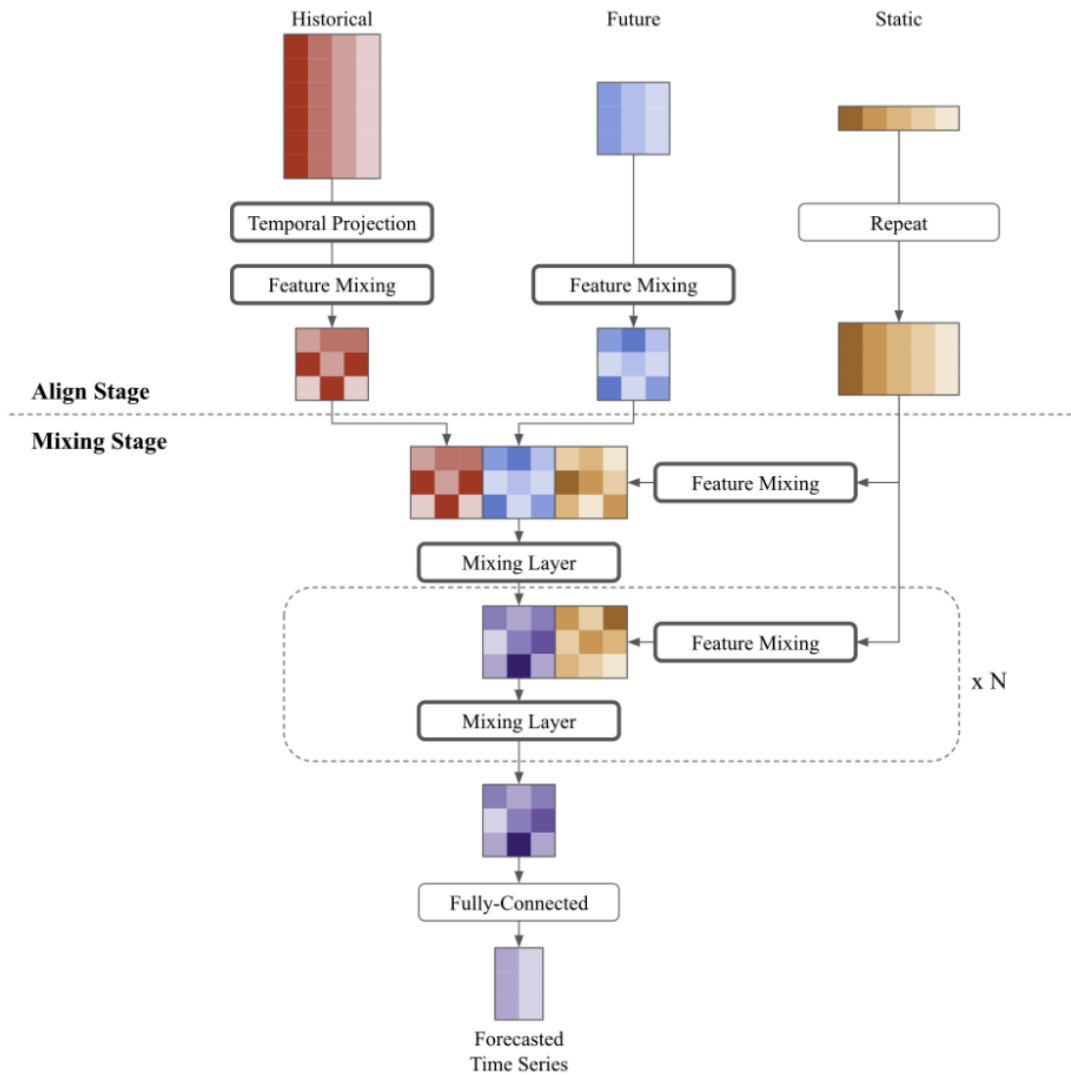


Figure 3: The architecture of TMix-Only. It is similar to TSMixer but only applies time-mixing.



- 입력 데이터는 일정 길이의 과거 시계열 구간을 잘라 모델에 입력함

- 모델은 여러 개의 Mixer Block을 반복해 구성됨
- 각 블록은 시간 축 정보와 변수 간 정보를 각각 MLP로 혼합함
- Time-Mixing 단계에서 시점 간 패턴을 학습함
- Feature-Mixing 단계에서 변수 간 상관관계를 학습함
- Residual 연결과 normalization을 적용해 학습 안정성을 확보함

## 4. Experiments

- Electricity, Traffic, Weather, Exchange 등 표준 시계열 데이터 사용
- 예측 길이는 96, 192, 336, 720 등 장기 예측 설정 포함
- Transformer 기반 모델과 동일 조건에서 성능 비교 수행
- 평가 지표는 MSE, MAE를 사용해 예측 오차를 측정함
- 다양한 예측 길이에서 모델 성능 안정성도 함께 분석함

## 5. Result

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: Evaluation results on the long-term forecasting datasets. The numbers of models marked with “\*” are obtained from Nie et al. (2023). The best numbers in each row are shown in **bold** and the second best numbers are underlined. We skip TMix-Only in comparisons as it performs similar to TSMixer. The last row shows the average percentage of MSE improvement of TSMixer over other methods.

		Multivariate Model										Univariate Model					
Models		TSMixer		TFT		FEDformer*		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	<b>0.361</b>	<b>0.392</b>	0.674	0.634	0.376	0.415	0.435	0.446	0.941	0.769	0.359	0.391	<b>0.368</b>	<b>0.392</b>	0.370	0.400
	192	<b>0.404</b>	<b>0.418</b>	0.858	0.704	0.423	0.446	0.456	0.457	1.007	0.786	0.402	0.415	<b>0.404</b>	<b>0.415</b>	0.413	0.429
	336	<b>0.420</b>	<b>0.431</b>	0.900	0.731	0.444	0.462	0.486	0.487	1.038	0.784	0.420	0.434	0.436	<b>0.439</b>	<b>0.422</b>	0.440
	720	<b>0.463</b>	<b>0.472</b>	0.745	0.666	0.469	0.492	0.515	0.517	1.144	0.857	0.453	0.467	0.481	0.495	<b>0.447</b>	<b>0.468</b>
ETTh2	96	<b>0.274</b>	<b>0.341</b>	0.409	0.505	0.332	0.374	0.332	0.368	1.549	0.952	0.275	0.342	0.297	0.363	<b>0.274</b>	<b>0.337</b>
	192	<b>0.339</b>	<b>0.385</b>	0.953	0.651	0.407	0.446	0.426	0.434	3.792	1.542	0.339	0.386	0.398	0.429	<b>0.341</b>	<b>0.382</b>
	336	<b>0.361</b>	<b>0.406</b>	1.006	0.709	0.400	0.447	0.477	0.479	4.215	1.642	0.366	0.413	0.500	0.491	<b>0.329</b>	<b>0.384</b>
	720	<b>0.445</b>	<b>0.470</b>	1.187	0.816	<b>0.412</b>	<b>0.469</b>	0.453	0.490	3.656	1.619	0.437	0.465	0.795	0.633	<b>0.379</b>	<b>0.422</b>
ETTm1	96	<b>0.285</b>	<b>0.339</b>	0.752	0.626	0.326	0.390	0.510	0.492	0.626	0.560	0.284	0.338	0.303	0.346	<b>0.293</b>	<b>0.346</b>
	192	<b>0.327</b>	<b>0.365</b>	0.752	0.649	0.365	0.415	0.514	0.495	0.725	0.619	0.324	0.362	0.335	<b>0.365</b>	<b>0.333</b>	0.370
	336	<b>0.356</b>	<b>0.382</b>	0.810	0.674	0.392	0.425	0.510	0.492	1.005	0.741	0.359	0.384	<b>0.365</b>	<b>0.384</b>	0.369	0.392
	720	<b>0.419</b>	<b>0.414</b>	0.849	0.695	0.446	0.458	0.527	0.493	1.133	0.845	0.419	0.414	<b>0.419</b>	<b>0.415</b>	<b>0.416</b>	0.420
ETTm2	96	<b>0.163</b>	<b>0.252</b>	0.386	0.472	0.180	0.271	0.205	0.293	0.355	0.462	0.162	0.249	0.170	0.266	<b>0.166</b>	<b>0.256</b>
	192	<b>0.216</b>	<b>0.290</b>	0.739	0.626	0.252	0.318	0.278	0.336	0.595	0.586	0.220	0.293	0.236	0.317	<b>0.223</b>	<b>0.296</b>
	336	<b>0.268</b>	<b>0.324</b>	0.477	0.494	0.324	0.364	0.343	0.379	1.270	0.871	0.269	0.326	0.308	0.369	<b>0.274</b>	<b>0.329</b>
	720	0.420	0.422	0.523	0.537	<b>0.410</b>	0.420	0.414	<b>0.419</b>	3.001	1.267	0.358	0.382	0.435	0.449	<b>0.362</b>	<b>0.385</b>
Weather	96	<b>0.145</b>	<b>0.198</b>	0.441	0.474	0.238	0.314	0.249	0.329	0.354	0.405	0.145	0.196	0.170	0.229	<b>0.149</b>	<b>0.198</b>
	192	<b>0.191</b>	<b>0.242</b>	0.699	0.599	0.275	0.329	0.325	0.370	0.419	0.434	0.190	0.240	0.213	0.268	<b>0.194</b>	<b>0.241</b>
	336	<b>0.242</b>	<b>0.280</b>	0.693	0.596	0.339	0.377	0.351	0.391	0.583	0.543	0.240	0.279	0.257	0.305	<b>0.245</b>	<b>0.282</b>
	720	0.320	<b>0.336</b>	1.038	0.753	0.389	0.409	0.415	0.426	0.916	0.705	0.325	0.339	<b>0.318</b>	0.356	<b>0.314</b>	<b>0.334</b>
Electricity	96	<b>0.131</b>	<b>0.229</b>	0.295	0.376	0.186	0.302	0.196	0.313	0.304	0.393	0.132	0.225	0.135	0.232	<b>0.129</b>	<b>0.222</b>
	192	0.151	<b>0.246</b>	0.327	0.397	0.197	0.311	0.211	0.324	0.327	0.417	0.152	0.243	<b>0.149</b>	<b>0.246</b>	<b>0.147</b>	<b>0.240</b>
	336	<b>0.161</b>	<b>0.261</b>	0.298	0.380	0.213	0.328	0.214	0.327	0.333	0.422	0.166	0.260	0.164	0.263	<b>0.163</b>	<b>0.259</b>
	720	<b>0.197</b>	<b>0.293</b>	0.338	0.412	0.233	0.344	0.236	0.342	0.351	0.427	0.200	0.291	0.199	0.297	<b>0.197</b>	<b>0.290</b>
Traffic	96	<b>0.376</b>	<b>0.264</b>	0.678	0.362	0.576	0.359	0.597	0.371	0.733	0.410	0.370	0.258	0.395	0.274	<b>0.360</b>	<b>0.249</b>
	192	<b>0.397</b>	<b>0.277</b>	0.664	0.355	0.610	0.380	0.607	0.382	0.777	0.435	0.390	0.268	0.406	0.279	<b>0.379</b>	<b>0.256</b>
	336	<b>0.413</b>	0.290	0.679	0.354	0.608	0.375	0.623	0.387	0.776	0.434	0.404	0.276	0.416	<b>0.286</b>	<b>0.392</b>	<b>0.264</b>
	720	<b>0.444</b>	<b>0.306</b>	0.610	0.326	0.621	0.375	0.639	0.395	0.827	0.466	0.443	0.297	0.454	0.308	<b>0.432</b>	<b>0.286</b>
TSMixer MSE Imp.				51.94%		16.69%		24.51%		62.40%		-0.66%		6.77%		-1.53%	

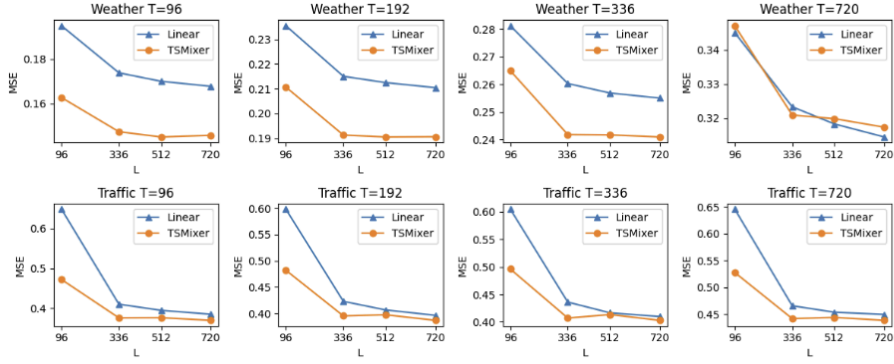


Figure 5: Performance comparison on varying lookback window size  $L$  of linear models and TSMixer.

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
Linear		$0.983 \pm 0.016$	$1.045 \pm 0.018$
PatchTST		$0.976 \pm 0.014$	$0.992 \pm 0.011$
<b>TMix-Only</b>		$0.960 \pm 0.041$	$1.000 \pm 0.027$
Autoformer	✓	$0.742 \pm 0.029$	$0.640 \pm 0.023$
FEDformer	✓	$0.804 \pm 0.039$	$0.674 \pm 0.014$
<b>TSMixer</b>	✓	<b><math>0.737 \pm 0.033</math></b>	<b><math>0.605 \pm 0.027</math></b>

Table 5: Evaluation on M5 with auxiliary information.

Models	Auxiliary feature		Test WRMSSE	Val WRMSSE
	Static	Future		
DeepAR	✓	✓	$0.789 \pm 0.025$	$0.611 \pm 0.007$
TFT	✓	✓	$0.670 \pm 0.020$	$0.579 \pm 0.011$
TSMixer-Ext	✓		$0.737 \pm 0.033$	$0.000 \pm 0.000$
			$0.657 \pm 0.046$	$0.000 \pm 0.000$
	✓	✓	$0.697 \pm 0.028$	$0.000 \pm 0.000$
		✓	<b><math>0.640 \pm 0.013</math></b>	<b><math>0.568 \pm 0.009</math></b>

Table 6: Computational cost on M5. All models are trained on a single NVIDIA Tesla V100 GPU. All models are implemented in PyTorch, except TFT, which is implemented in MXNet.

Models	Multivariate	Auxiliary feature	# of params	training time (s)	inference (step/s)
Linear			1K	2958.18	110
PatchTST			26.7K	886.101	120
<b>TMix-Only</b>			6.3K	4073.72	110
Autoformer	✓		471K	119087.64	42
FEDformer	✓		1.7M	11084.43	56
<b>TSMixer</b>	✓		189K	11077.95	96
DeepAR	✓	✓	1M	8743.55	105
TFT	✓	✓	2.9M	14426.79	22
<b>TSMixer-Ext</b>	✓	✓	244K	11615.87	108

- 대부분 데이터셋에서 기존 Transformer 모델과 유사하거나 더 좋은 성능 기록
- 장기 예측 구간에서 특히 안정적인 결과를 보임
- 구조가 단순해 학습 속도와 추론 속도가 개선됨

- 파라미터 수 대비 효율적인 성능을 확인함
- mixing 구조가 실제 예측 성능에 중요한 역할을 함을 확인함

## 6. Insight

- 시계열 예측에서 복잡한 attention 구조가 항상 필요한 것은 아님을 보여줌
- 단순한 MLP 구조도 충분한 성능을 낼 수 있음을 실험으로 증명함
- 계산 비용이 낮아 실무 적용 가능성이 높음
- 향후 효율적인 forecasting 모델 연구에 중요한 기준점이 될 수 있음
- 데이터 특성에 따라 구조 선택이 중요하다는 점을 시사함