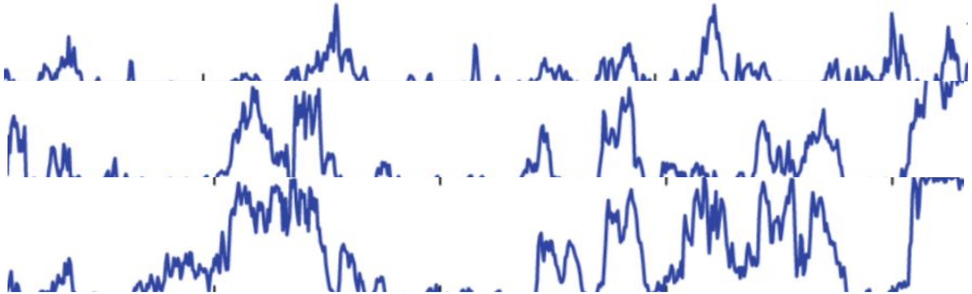
Large Language Model Guided Knowledge Distillation for Time Series Anomaly Detection

Chen Liu, Shibo He, Qihang Zhou, Shizhong Li, Wenchao Meng

Seminar KI | Prof. Schmidt-Schauß | SoSe 2025

Roman Christof



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

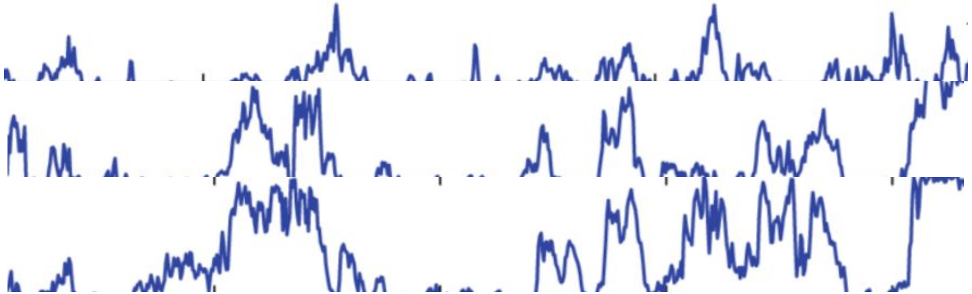
Large Language Model Guided Knowledge Distillation for Time Series Anomaly Detection

AnomalyLLM

Chen Liu, Shibo He, Qihang Zhou, Shizhong Li, Wenchao Meng

Seminar KI | Prof. Schmidt-Schauß | SoSe 2025

Roman Christof



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Agenda

1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

Agenda

1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

Was sind Zeitreihen und Anomalien?

Zeitreihe Blázquez-García et al. (2021)

- geordnete Folge reellwertiger Beobachtungen, die zu festgelegten Zeitpunkten erhoben werden

Anomalie Hawkins (1980)

- Beobachtung, die stark von den übrigen Daten abweicht und damit die Vermutung nahelegt von einem anderen Mechanismus erzeugt worden zu sein

Wo treten Zeitreihen auf?

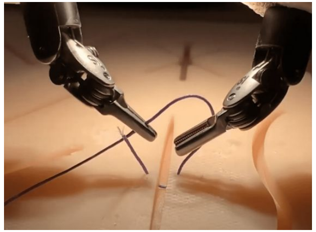
Boniol et al. (2024)

- relevant für eine Vielzahl an wissenschaftlichen Disziplinen und Industriezweige
    - Informatik und IT
    - Astronomie
    - Biologie
    - Ökonomie
    - Energie- & Ingenieurwissenschaften
    - Umweltwissenschaft
    - Medizin
    - Neurowissenschaft
    - Sozialwissenschaft
    - …

Beispiele für Zeitreihen

Boniol et al. (2024)

Medizin



[1]

Astrophysik

[2]

Energieproduktion

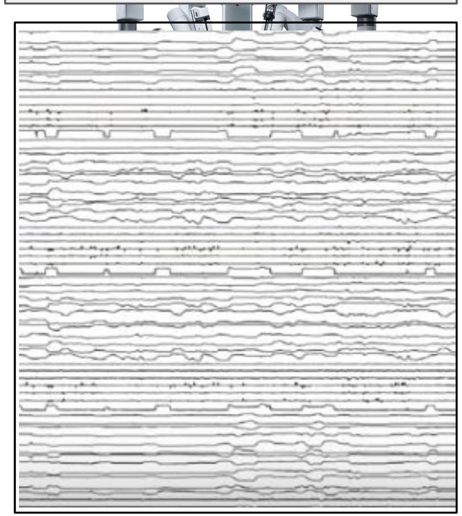
[3]

Beispiele für Zeitreihen

Boniol et al. (2024)

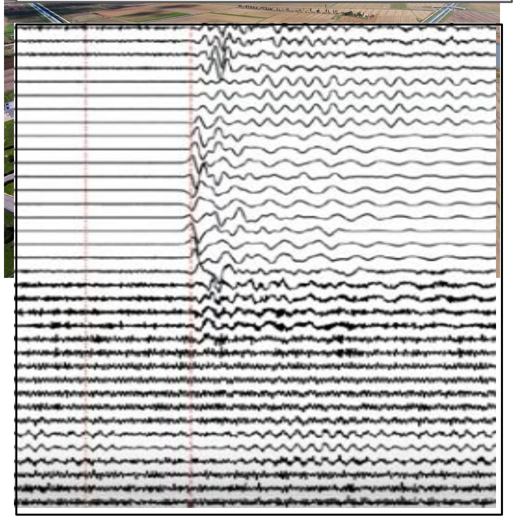
Medizin

Sensormessungen des Da-Vinci-Operations-roboters



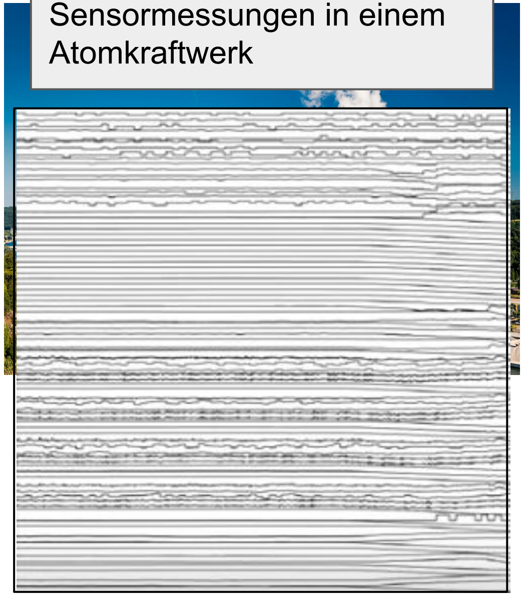
Astrophysik

Akustischer Sensoren eines Gravitationswellendetektor



Energieproduktion

Sensormessungen in einem Atomkraftwerk

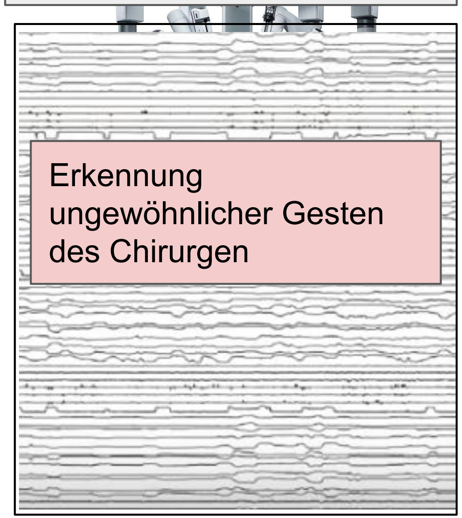


Beispiele für Zeitreihen

Boniol et al. (2024)

Medizin

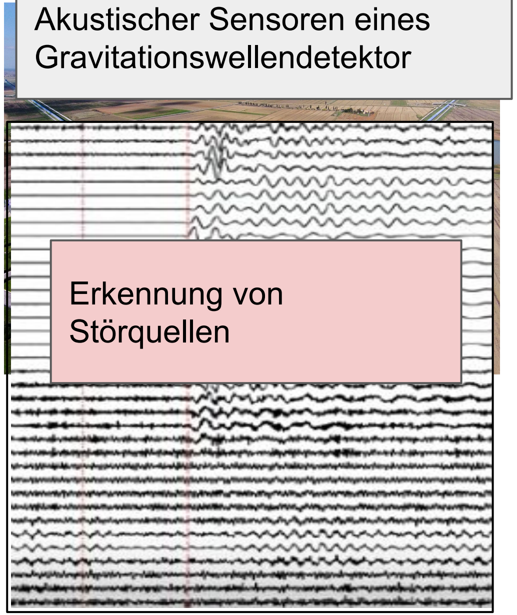
Sensormessungen des Da-Vinci-Operations-roboters



Erkennung ungewöhnlicher Gesten des Chirurgen

Astrophysik

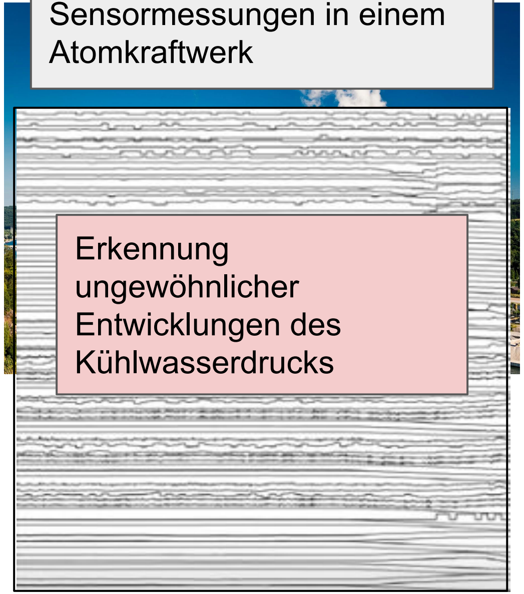
Akustischer Sensoren eines Gravitationswellendetektor



Erkennung von Störquellen

Energieproduktion

Sensormessungen in einem Atomkraftwerk



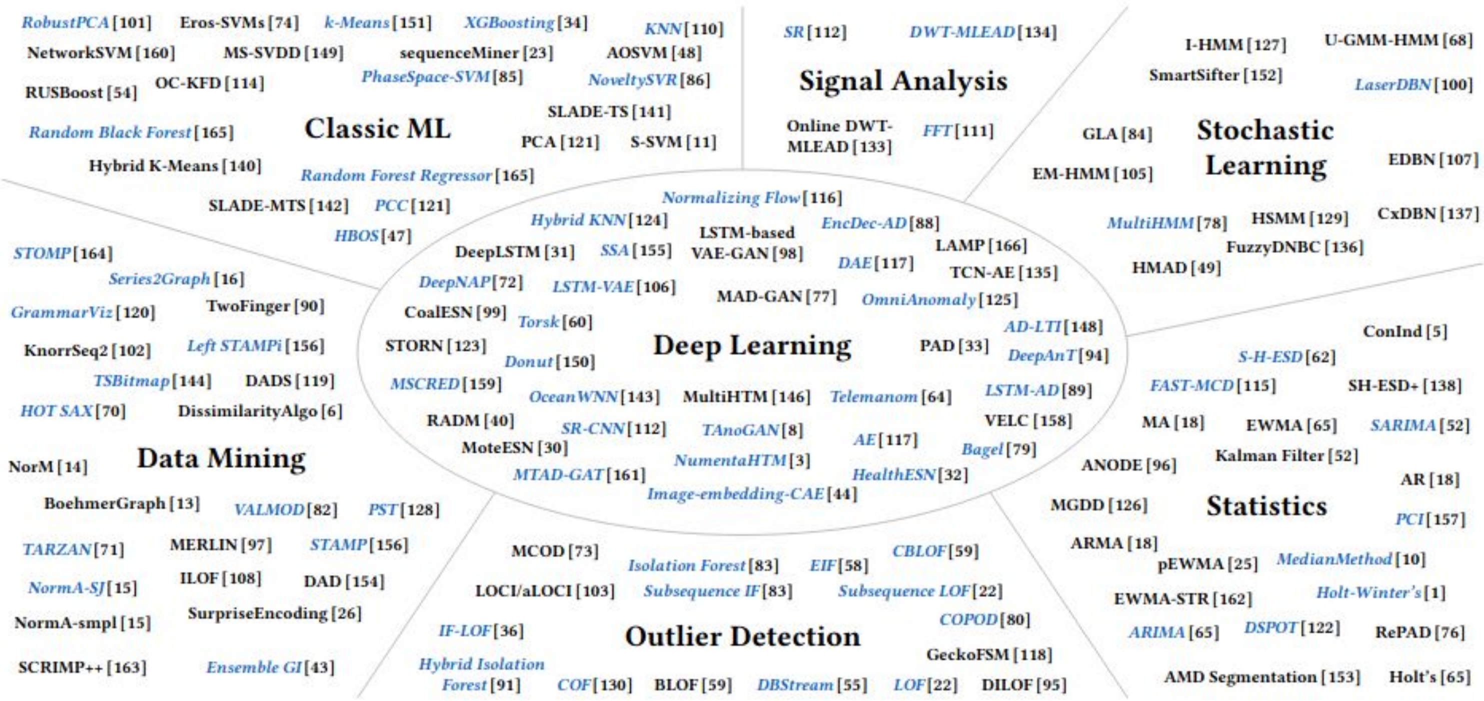
Erkennung ungewöhnlicher Entwicklungen des Kühlwasserdrucks

Agenda

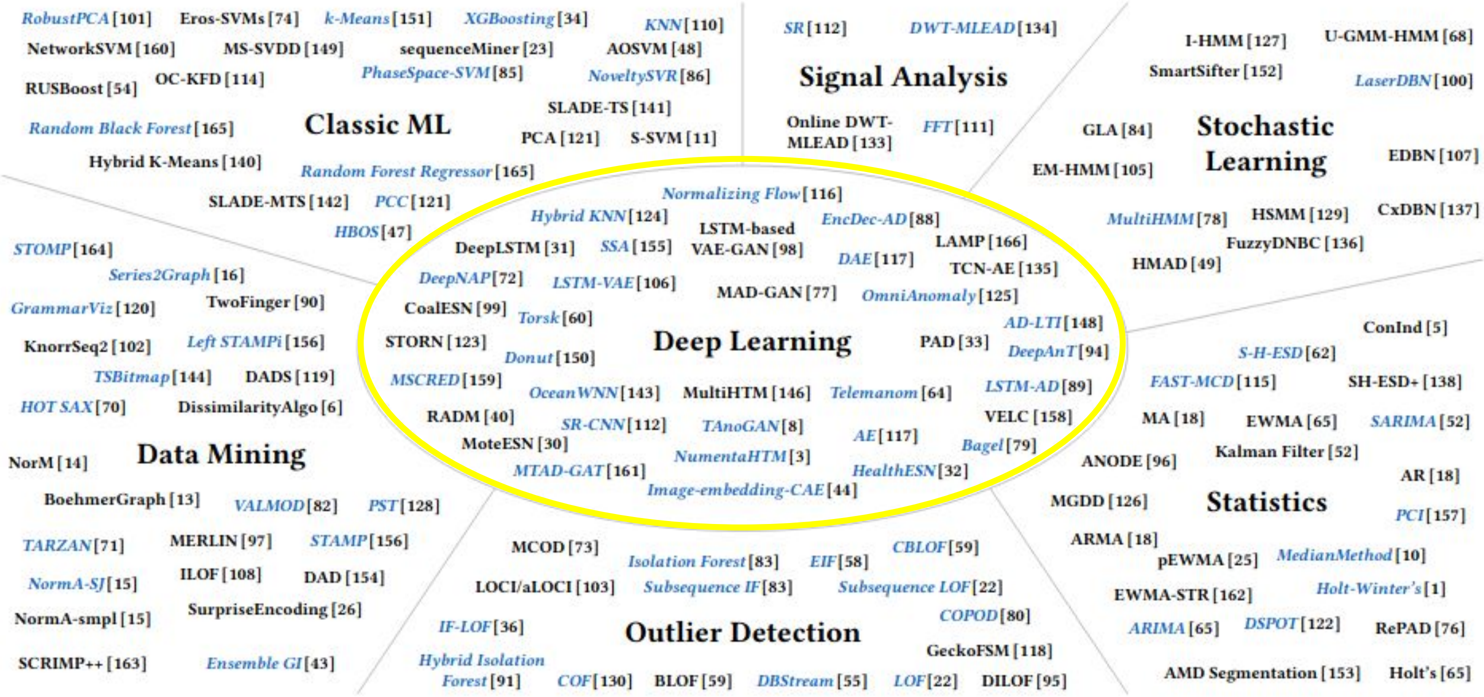
1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

Taxonomie der Methoden zur Anomalieerkennung

Schmidl et al. (2022)



Taxonomie der Methoden zur Anomalieerkennung Schmidl et al. (2022)



Self-supervised Verfahren

- generieren Pseudo-Labels aus den Daten selbst mittels Pretext-Tasks

  → Lernen von Repräsentationen ohne explizit annotierte Daten

Problem:

- scheitert bei Datenknappheit

Ziel von AnomalyLLM

Lernen robuster Repräsentationen von Zeitreihen in low-resource-Szenarien zur Detektion von Anomalien

→ Large Language Models

Fehlende Daten

- keine großen universellen Datensätze für Zeitreihen
    - Time Series: UCR, Monash Archive (~10 GB)
    - Computer Vision: ImageNET (130 GB), LAION-400M (10 TB)
    - NLP: Common Crawl (Petabytes)

Vorarbeiten - LLMs

Zhou et al. (2023)

GPT4TS

- LLMs erstmals zur Anomalieerkennung in Zeitreihen
- vortrainierte LLMs erzeugen universelle Repräsentationen – auch für Zeitreihen

Starke Generalisierungsfähigkeit des GPT-Modells führt zur Rekonstruktion anomaler Daten

→ viele *false negatives*

Vorarbeiten - Knowledge Distillation

Hinton et al. (2015), Salehi et al. (2021), Zhou et al. (2021)

- Knowledge Distillation: Übertragung von Wissen eines Teacher-Netzwerks auf eine kleineres Student-Netzwerk

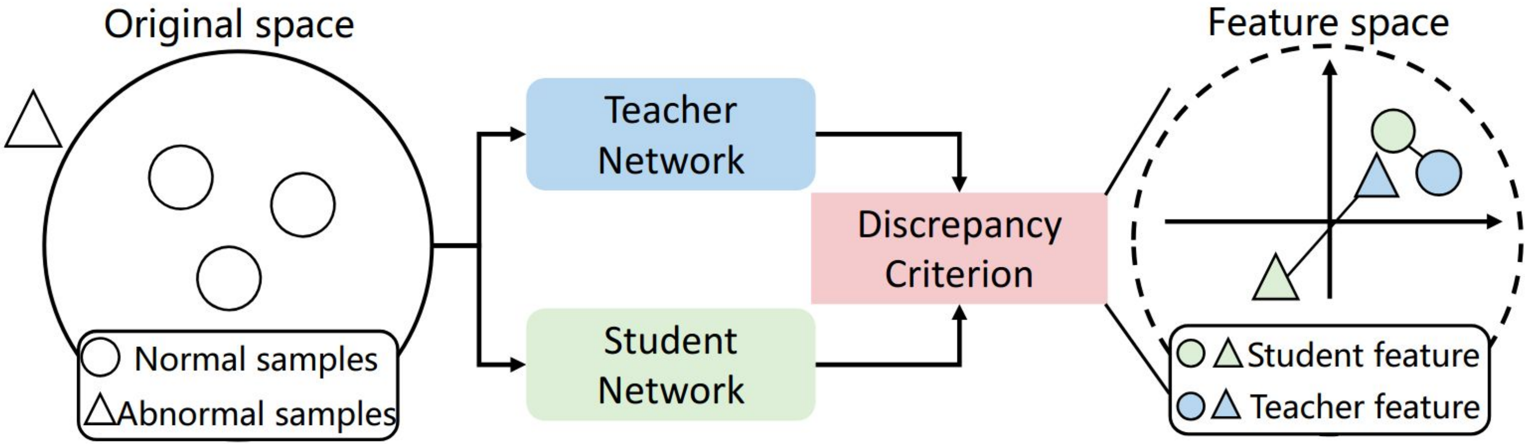
- Methode bereits in der Anomalieerkennung im Bereich der Bildverarbeitung angewandt

- AnomalyLLM: Knowledge Distillation erstmals für Anomalieerkennung bei Zeitreihen

Agenda

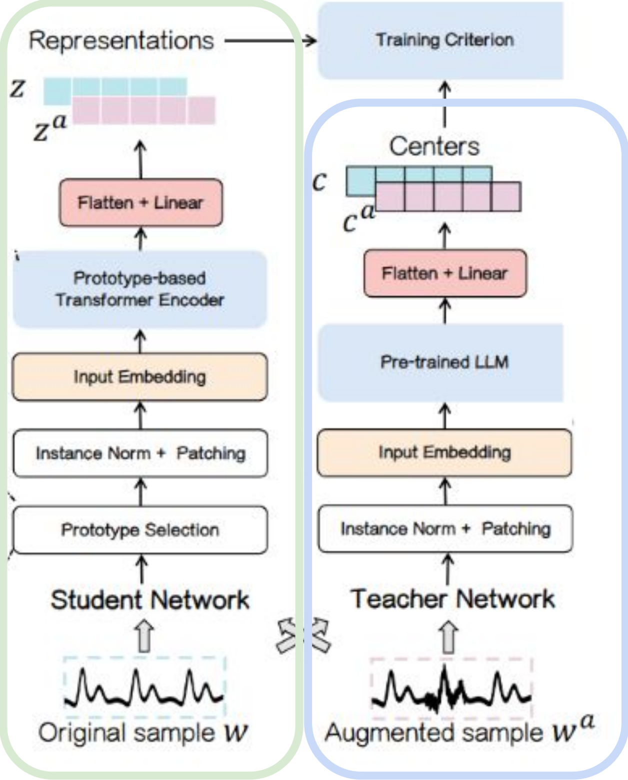
1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

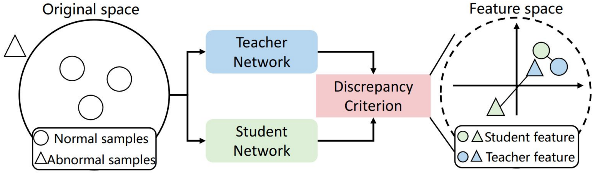
Idee: Teacher-Student-Ansatz



Liu et al. (2024)

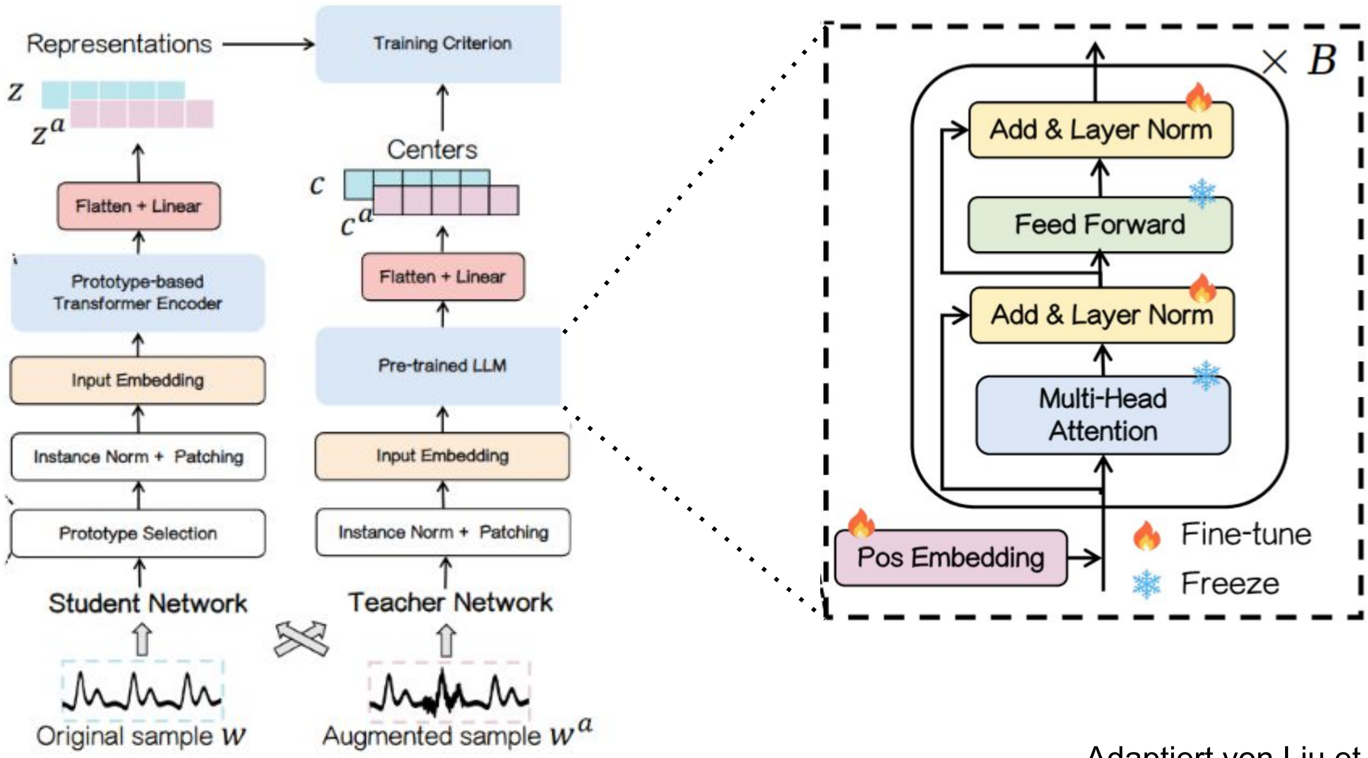
Teacher-Student-Ansatz





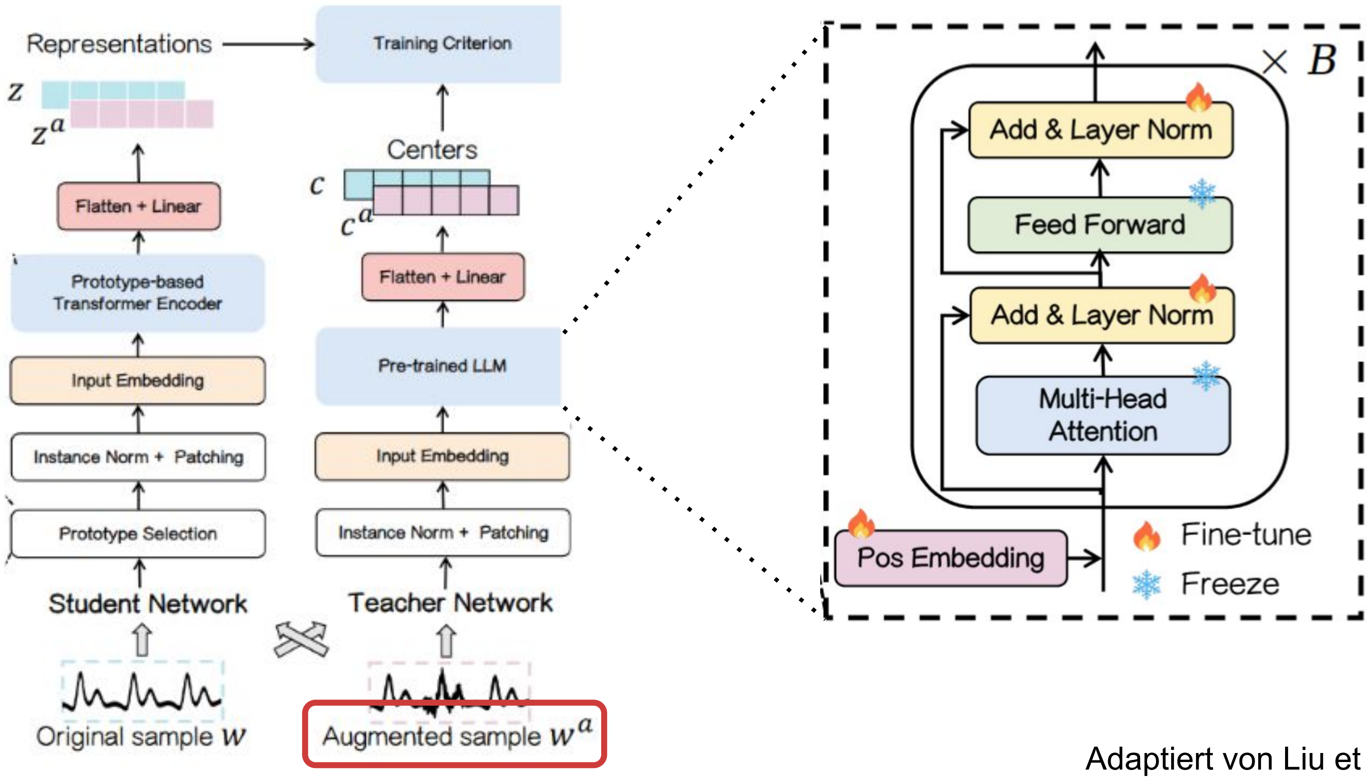
Adaptiert von Liu et al. (2024)

Teacher-Netzwerk - Adaptiertes LLM



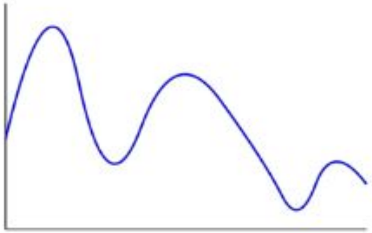
Adaptiert von Liu et al. (2024)

Teacher-Netzwerk - Augmentation

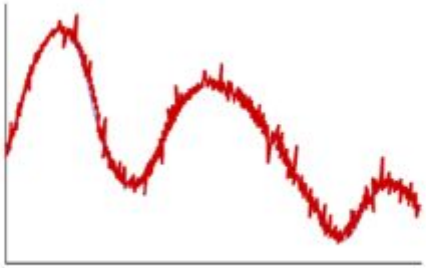


Adaptiert von Liu et al. (2024)

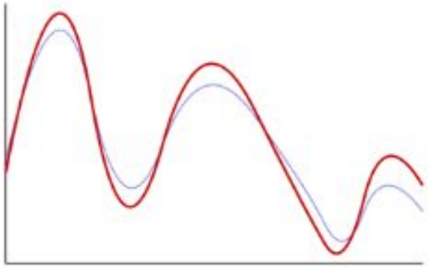
Teacher-Netzwerk - Augmentation



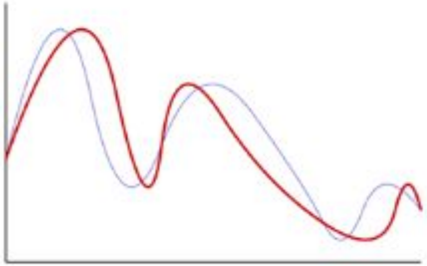
Ground Truth



Jittering



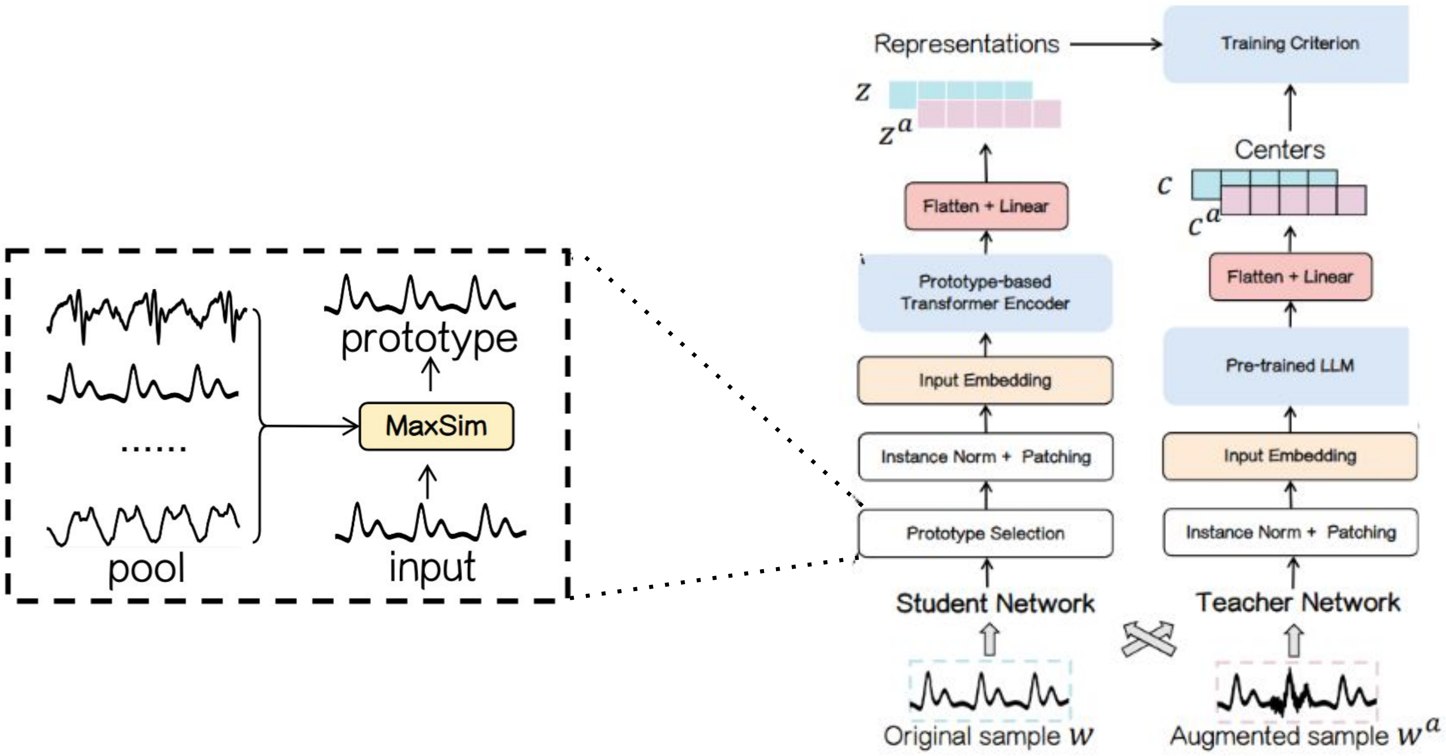
Scaling



Warping

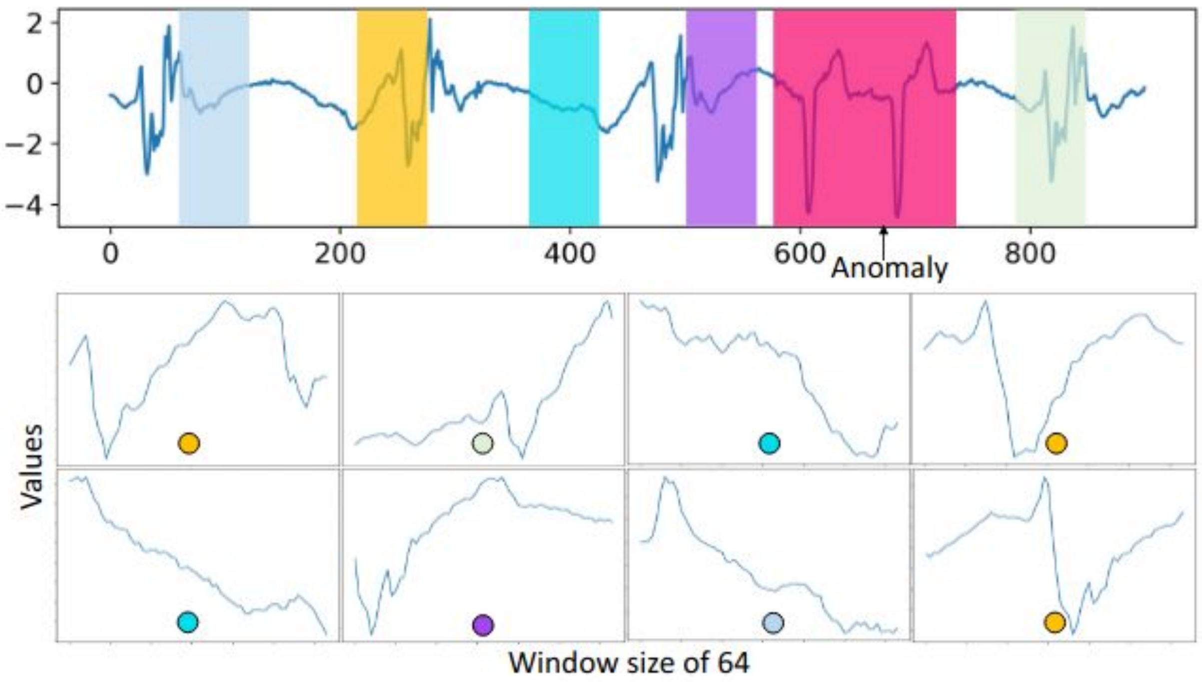
Adaptiert von Iglesias et al. (2023)

Student-Netzwerk



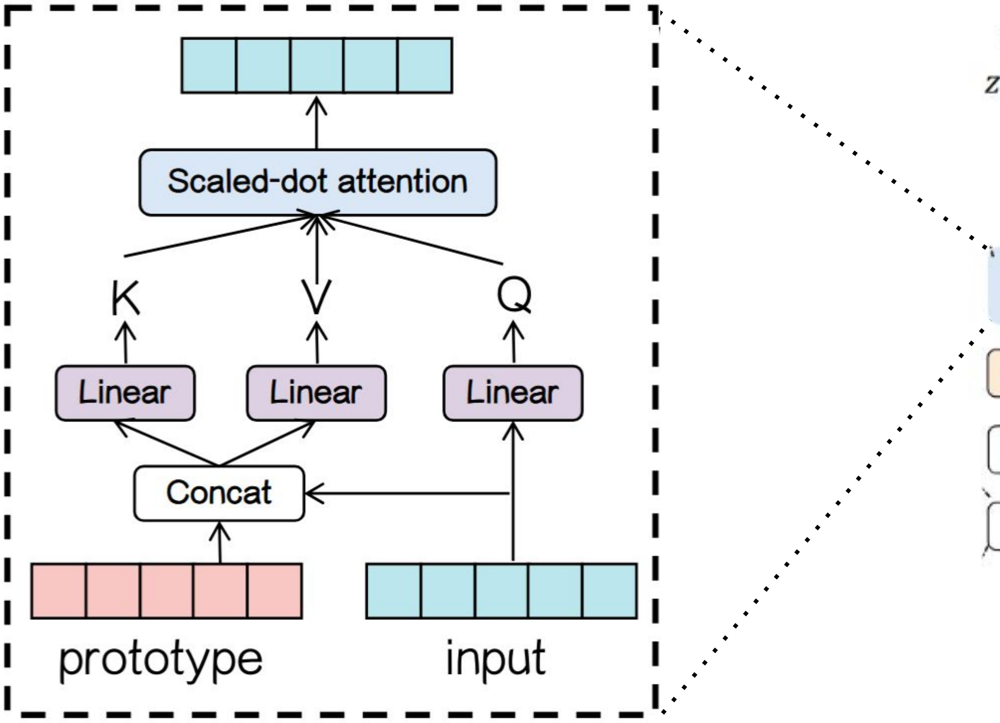
Adaptiert von Liu et al. (2024)

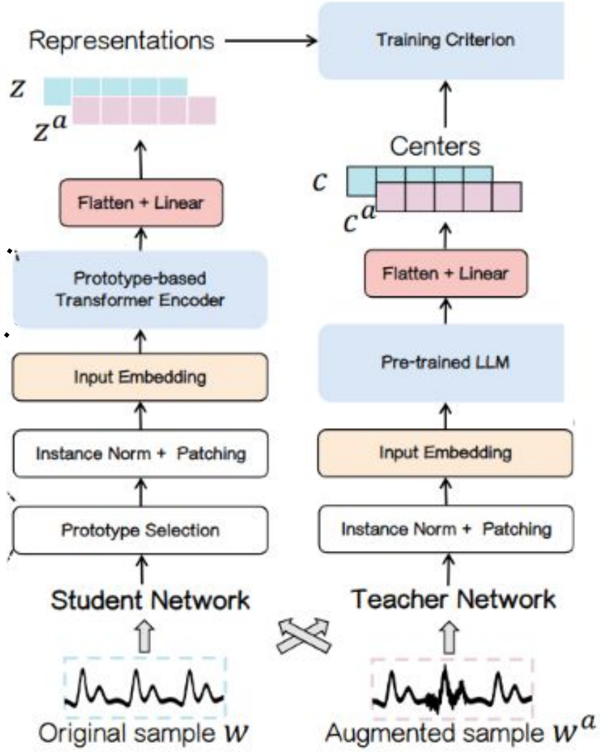
Prototypen



Liu et al. (2024)

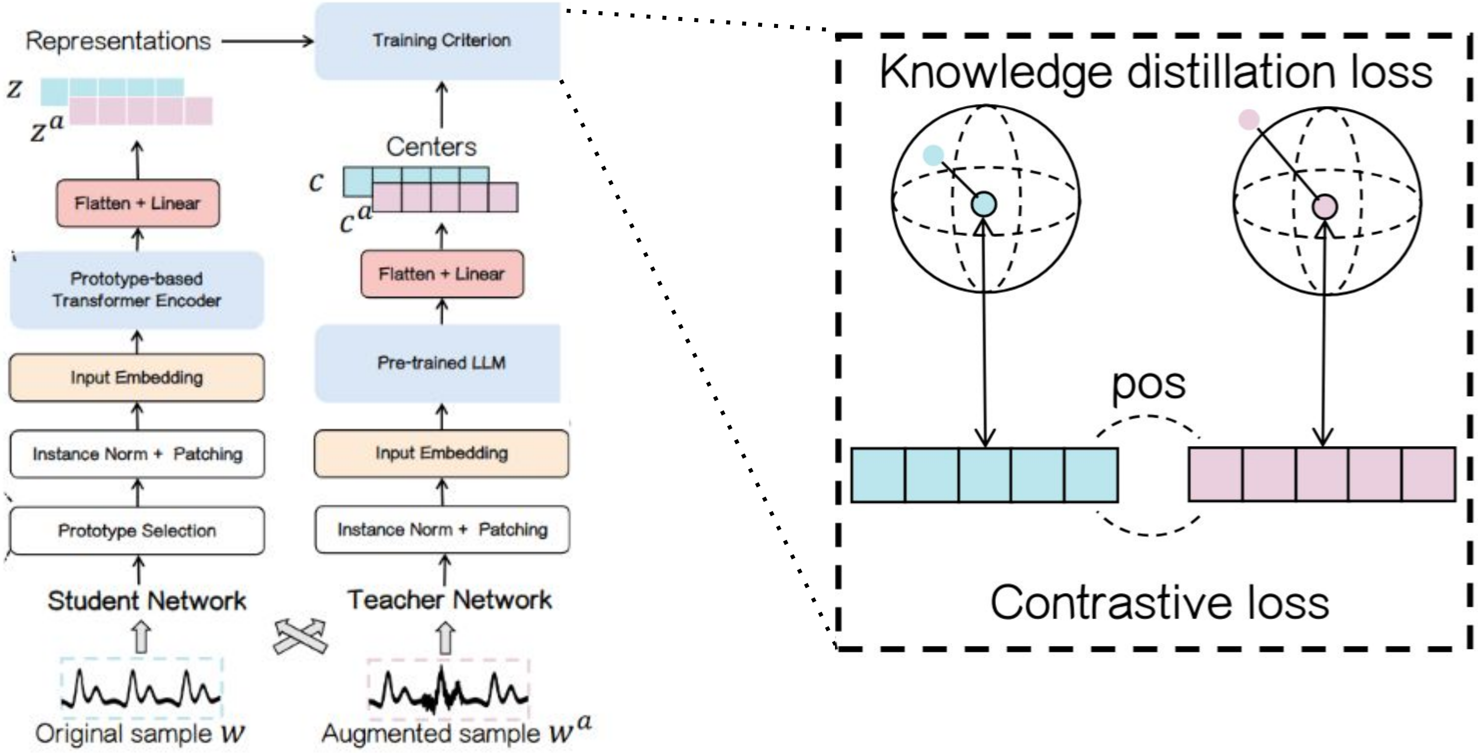
Teacher-Student-Ansatz





Adaptiert von Liu et al. (2024)

Teacher-Student-Ansatz



Adapted from Liu et al. (2024)

Knowledge Distillation Loss

Feature-Vektor des Student-Netwerks

Augmented

$$\mathcal{L}_{kd} = \frac{1}{N}\sum_{i=1}^{N} \|z_i - c_i\|_2^2 - \log(1 - \exp(-\|z_i^a - c_i^a\|_2^2))$$

i: Zeitfenster

Feature-Vektor des Teacher-Netzwerks

Knowledge Distillation Loss - Negative-sample-free Contrastive Loss

Feature-Vektor des Teacher-Netzwerks

Augmented

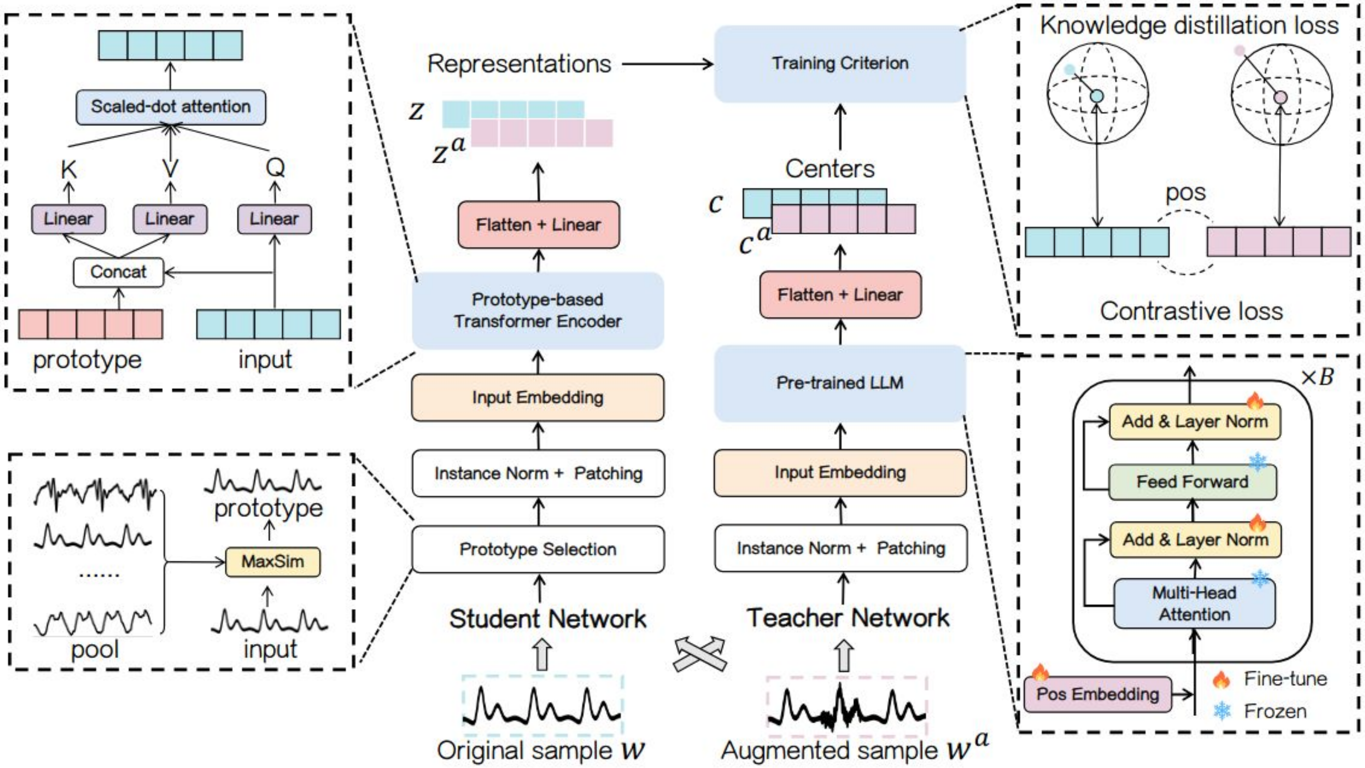
$$\mathcal{L}_{ce} = \frac{1}{N} \sum_{i=1}^{N} -\frac{c_i}{\|c_i\|_2} \cdot \frac{c_i^a}{\|c_i^a\|_2}$$

Knowledge Distillation Loss

Hyperparameter der Einfluss der Loss-Komponenten kontrolliert

$$\mathcal{L}_{total} = \mathcal{L}_{kd} + \lambda \mathcal{L}_{ce}$$

Übersicht



Liu et al. (2024)

Berechnung des Anomalie-Scores

Zeitfenster

Student-Repräsentation

$$A(\mathbf{w}) = \|\phi(\mathbf{w}) - \varphi(\mathbf{w})\|_2^2.$$

Teacher-Repräsentation

Liu et al. (2024)

Agenda

1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

UCR Datensatz (univariate)

	ABP		ACC		AirTem		ECG		EPG		Gait		NASA		Power		RESP		Total	
	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1	Acc	AF1
DeepSVDD	0.333	0.564	0.714	0.743	0.385	0.726	0.297	0.587	0.360	0.700	0.242	0.495	0.182	0.435	0.182	0.418	0.000	0.225	0.288	0.555
AnoTrans	0.357	0.645	0.286	0.580	0.538	0.759	0.297	0.655	0.400	0.755	0.364	0.683	0.636	0.819	0.455	0.660	0.117	0.667	0.348	0.679
Dcdetector	0.571	0.556	0.571	0.493	0.846	0.681	0.220	0.303	0.640	0.717	0.424	0.438	0.909	0.736	0.455	0.492	0.117	0.588	0.424	0.479
MEMTO	0.405	0.655	0.714	0.750	0.615	0.750	0.319	0.616	0.440	0.733	0.364	0.673	0.364	0.664	0.455	0.680	0.059	0.598	0.368	0.656
MTGflow	0.500	0.558	0.714	0.904	0.462	0.687	0.286	0.602	0.520	0.661	0.364	0.572	0.364	0.749	0.182	0.574	0.235	0.513	0.372	0.608
GPT4TS	0.476	0.681	0.429	0.504	0.462	0.686	0.330	0.608	0.360	0.759	0.212	0.461	0.364	0.849	0.182	0.598	0.353	0.544	0.348	0.623
TS-TCC	0.690	0.754	0.286	0.549	**1.000**	0.969	0.637	0.784	0.880	0.931	0.697	0.794	0.364	0.511	0.545	0.763	0.412	0.560	0.656	0.770
THOC	0.762	0.815	0.714	0.776	**1.000**	0.971	0.604	0.760	0.880	0.908	0.636	0.784	0.909	0.896	0.455	0.775	0.294	0.389	0.671	0.780
NCAD	0.680	0.794	0.846	0.849	0.714	0.758	0.593	0.735	0.760	0.789	0.848	0.858	0.818	0.861	0.545	0.723	0.353	0.613	0.663	0.767
COCA	0.714	0.740	0.428	0.545	**1.000**	0.946	0.637	0.767	0.640	0.785	0.545	0.699	0.818	0.841	0.364	0.632	0.235	0.562	0.620	0.742
Our	**0.857**	**0.920**	**1.000**	**0.956**	**1.000**	**0.974**	**0.758**	**0.787**	**0.920**	**0.933**	**0.878**	**0.871**	**1.000**	**0.961**	**0.818**	**0.886**	**0.471**	**0.736**	**0.820**	**0.858**

UCR Time Series Classification Archive

SMD, MSL, SMAP, PSM Datensatz (multivariate)

	SMD	MSL	SMAP	PSM
IForest	0.536	0.665	0.555	0.835
DAGMM	0.573	0.746	0.685	0.801
DeepSVDD	0.791	0.836	0.690	0.907
THOC	0.850	0.897	0.907	0.895
LSTMVAE	0.823	0.826	0.781	0.810
BeatGAN	0.781	0.875	0.696	0.920
DAEMON	**0.963**	0.910	0.953	0.928
OmniAnomaly	0.852	0.877	0.869	0.828
TranAD	0.961	0.892	0.949	0.967
AnomalyTrans	0.912	0.906	0.962	0.977
AnomalyBERT	0.925	0.914	0.585	0.950
DCdetector	0.872	**0.966**	**0.970**	0.979
MEMTO	0.935	0.944	0.966	0.983
Our	0.958	0.956	0.965	**0.997**

Table 4: F1 score on SMD, MSL, SMAP, and PSM.

Few-shot und Zero-shot Settings (F1-Score)

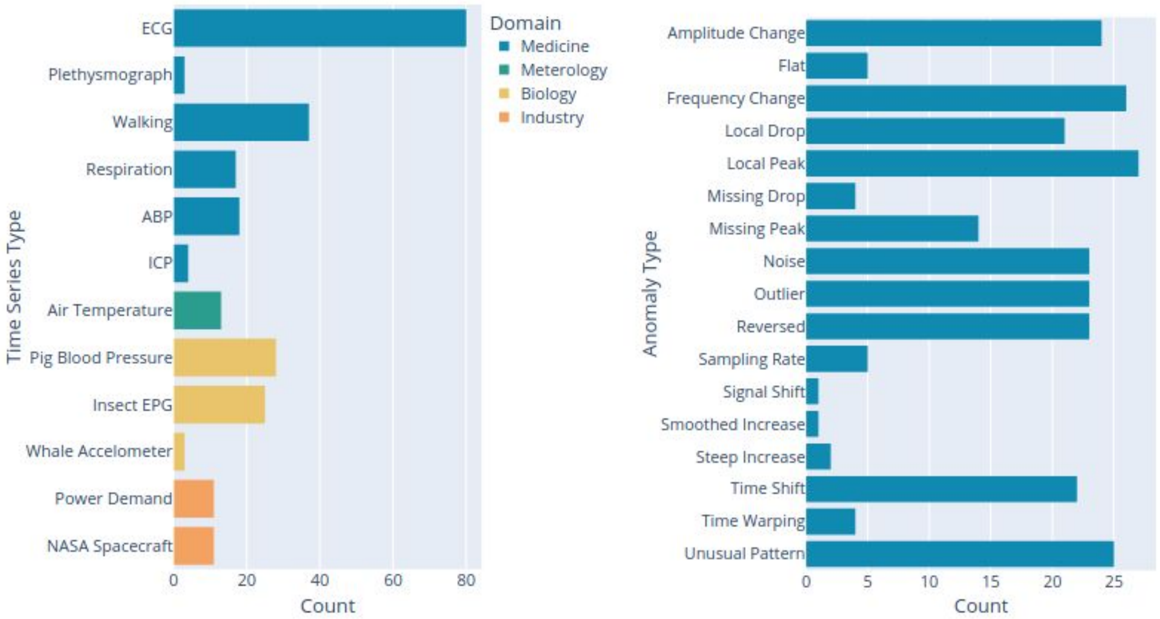
	100%	50%	20%	10%	zero-shot
TSTCC	0.763	0.577	0.488	0.213	0.139
THOC	0.775	0.686	0.570	0.473	0.269
COCA	0.632	0.584	0.562	0.478	0.150
Our	0.886	0.817	0.784	0.781	0.650

Power Demand Datensatz von UCR

Agenda

1. Einleitung
2. Grundlagen und Vorarbeiten
3. Methoden
4. Ergebnisse
5. Diskussion

Aufschlüsselung nach Anomalietypen fehlt



(a) Histogram of time series types included in the UCR Anomaly Archive. The color indicates the domain, the time series originates from.

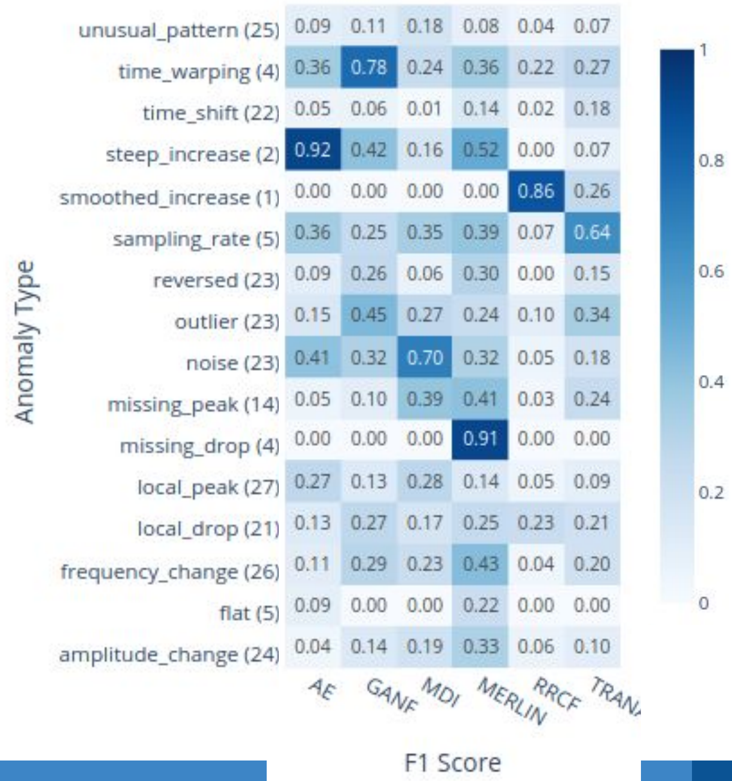
(b) Histogram of anomaly types present in the UCR Anomaly Archive.

Rewicki et al. (2023)

Aufschlüsselung nach Anomalietypen fehlt

- unterschiedliche Methoden haben Stärken bezüglich bestimmter Anomalietypen

	mechanism	class
RRCF	Isolation Forest	classical
MDI	Density Estimation	classical
MERLIN	Discord Discovery	classical
AE	Reconstruction	deep-learning
GANF	Density Estimation	deep-learning
TranAD	Reconstruction	deep-learning



Rewicki et al. (2023)

Experimente - Low-Resource Setting

- Evaluierung nur auf einem Teildatensatz von UCR (Power Demand)
  - Sollte auf dem ganzen Datensatz evaluiert werden, da dieses Szenario das Ziel des Papers ist
- Kein Vergleich zu klassischen Methoden, die weniger ‘datenhungrig’ sind

	100%	50%	20%	10%	zero-shot
TSTCC	0.763	0.577	0.488	0.213	0.139
THOC	0.775	0.686	0.570	0.473	0.269
COCA	0.632	0.584	0.562	0.478	0.150
Our	0.886	0.817	0.784	0.781	0.650

Power Demand Datensatz von UCR

Limitationen der Data Augmentation

- die drei Basis-Augmentierungen (Jitter, Scaling, Warping) repräsentieren nicht domänenspezifische Störungen
- mögliche Methoden zur Erzeugung komplexer und realitätsnaher synthetischer Daten:
  - variational autoencoders (VAE)
  - generative Adversarial Network (GAN)

Iglesias et al. (2023)

Limitationen der Data Augmentation

	ABP				Acceleration				Air Temperature				ECG				EPG			
	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1
Informer	0.738	0.810	0.805	0.807	1.000	0.929	0.921	0.925	0.692	0.757	0.752	0.754	0.450	0.694	0.693	0.693	0.560	0.830	0.826	0.828
Reformer	0.714	0.815	0.807	0.811	0.714	0.732	0.728	0.730	1.000	0.965	0.948	0.956	0.340	0.655	0.654	0.655	0.640	0.882	0.880	0.881
Autoformer	0.762	0.856	0.851	0.853	1.000	0.918	0.909	0.913	1.000	0.925	0.913	0.919	0.352	0.613	0.613	0.613	0.680	0.850	0.847	0.848
TCN	0.595	0.761	0.750	0.755	1.000	0.929	0.921	0.925	0.692	0.815	0.808	0.812	0.264	0.544	0.544	0.544	0.280	0.680	0.680	0.680
TimesNet	0.452	0.664	0.661	0.662	0.571	0.719	0.692	0.705	0.462	0.772	0.777	0.775	0.253	0.562	0.561	0.561	0.360	0.613	0.613	0.613
TST	0.667	0.768	0.756	0.762	0.571	0.733	0.733	0.733	0.923	0.925	0.902	0.913	0.451	0.685	0.683	0.684	0.440	0.742	0.738	0.740
Nonaug	0.595	0.735	0.730	0.732	0.571	0.762	0.761	0.761	0.846	0.778	0.773	0.775	0.230	0.575	0.575	0.575	0.360	0.652	0.653	0.652
w/o center	0.714	0.868	0.860	0.864	0.571	0.746	0.753	0.749	0.846	0.871	0.866	0.868	0.286	0.672	0.673	0.672	0.615	0.672	0.672	0.672
w/ feature	0.857	0.915	0.904	0.909	0.429	0.684	0.682	0.683	0.769	0.864	0.858	0.861	0.593	0.780	0.780	0.780	0.360	0.718	0.717	0.718
Ours	0.857	0.931	0.910	0.920	1.000	0.965	0.948	0.956	1.000	0.989	0.959	0.974	0.758	0.768	0.808	0.787	0.920	0.935	0.932	0.933

	Gait				NASA				PowerDemand				RESP				Avg			
	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1	Acc	AP	AR	AF1
Informer	0.485	0.750	0.747	0.748	0.455	0.567	0.567	0.567	0.364	0.561	0.560	0.561	0.353	0.564	0.563	0.563	0.532	0.724	0.720	0.722
Reformer	0.636	0.792	0.787	0.790	0.818	0.820	0.809	0.815	0.455	0.677	0.675	0.676	0.471	0.599	0.598	0.598	0.552	0.745	0.729	0.737
Autoformer	0.818	0.881	0.880	0.880	1.000	0.949	0.944	0.946	0.727	0.836	0.833	0.834	0.176	0.597	0.597	0.597	0.600	0.761	0.736	0.748
TCN	0.667	0.784	0.780	0.782	0.909	0.895	0.966	0.929	0.455	0.593	0.592	0.592	0.235	0.559	0.558	0.559	0.452	0.669	0.669	0.669
TimesNet	0.424	0.682	0.681	0.681	0.273	0.585	0.585	0.585	0.091	0.497	0.495	0.496	0.118	0.393	0.393	0.393	0.324	0.602	0.600	0.601
TST	0.576	0.716	0.710	0.713	0.455	0.596	0.595	0.596	0.455	0.623	0.622	0.623	0.176	0.557	0.557	0.557	0.512	0.707	0.703	0.705
Nonaug	0.636	0.804	0.800	0.802	0.727	0.814	0.812	0.813	0.455	0.700	0.700	0.700	0.353	0.748	0.748	0.748	0.440	0.683	0.670	0.676
w/o center	0.667	0.803	0.800	0.802	0.545	0.746	0.745	0.745	0.455	0.599	0.596	0.597	0.235	0.562	0.561	0.562	0.604	0.727	0.710	0.718
w/ feature	0.424	0.685	0.682	0.683	0.545	0.738	0.739	0.739	0.545	0.711	0.708	0.709	0.353	0.668	0.667	0.667	0.575	0.773	0.771	0.772
Ours	0.878	0.891	0.852	0.871	1.000	0.969	0.953	0.961	0.818	0.888	0.884	0.886	0.471	0.736	0.736	0.736	0.820	0.857	0.860	0.858

‘Nonaug’ = ohne Daten Augementation

Laufzeit und Latenz

- es fehlen substanzielle Angaben zur Laufzeit und Latenz

- LLM benötigt deutlich mehr Ressourcen und längere Laufzeit als herkömmliche Methoden
    - großes Problem für Online-Monitoring

Vorteile von AnomalieLLM bei begrenzten Daten

- LLM-basierte Wissensdistillation zeigt in datenarmen Szenarien Vorteile gegenüber klassischen unsupervised Methoden
- Ansatz zur Überwindung von Problemen der übermäßigen Generaliersierung von LLMs (GPT4TS)
- perspektivische Beispielanwendung: Medizinische Überwachung
  - Hohe physiologische Variabilität
  - Anomalien schwer zu definieren und zu labeln
  - Few-/Zero-Shot-Fähigkeit von AnomalyLLM ermöglicht Vorhersagen mit minimalen Daten

Bildnachweise

[1] https://www.medicalexpo.de/prod/intuitive-surgical/product-75060-613733.html, https://r3.ieee.org/wnc/technology-engineering-innovation-news/robotic-surgery-procedures-performed-by-surgical-robots/

[2] https://en.wikipedia.org/wiki/Virgo_interferometer

[3] https://montel.energy/commentary/nuclear-power-in-benelux-a-tale-of-two-worlds

Literatur

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. ACM Computing Surveys (CSUR), 54(3), April 2021.

Boniol, P., Liu, Q., Huang, M., Palpanas, T., & Paparrizos, J. (2024). Dive into Time-Series Anomaly Detection: A Decade Review. arXiv preprint, arXiv:2402.00551.

Hawkins, D. M. (1980). Identification of Outliers. Monographs on Statistics and Applied Probability, Springer Dordrecht.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint, arXiv:1503.02531.

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. Proceedings of the VLDB Endowment (PVLDB), 15(9), 1779–1797.

Zhou, T., Niu, P., Wang, X., Sun, L., & Jin, R. (2023). One fits all: Power general time series analysis by pretrained LM. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS '23), Curran Associates Inc.

Iglesias, G., Caparrini, M., Castro, A., & del Ser, J. (2023). Data augmentation techniques in time series domain: a survey and taxonomy. Neural Computing and Applications, 35(14), 10123–10145.

Rewicki, F., Denzler, J., & Niebling, J. (2023). Is it worth it? Comparing six deep and classical methods for unsupervised anomaly detection in time series. Applied Sciences, 13(3), 1778.

Liu, Chen & He, Shibo & Zhou, Qihang & Li, Shizhong & Meng, Wenchao. (2024). Large Language Model Guided Knowledge Distillation for Time Series Anomaly Detection. 2162-2170. 10.24963/ijcai.2024/239.

FRAGEN?

Dataset Statistics

Dataset	Train	Test	Anomaly Ratio
ABP	1036746	1841461	0.53%
Acceleration	38400	62337	2.45%
AirTemperature	52000	54392	2.86%
ECG	1795083	6047314	0.53%
EPG	119000	410415	1.29%
Gait	1157571	2784520	0.43%
NASA	38500	86296	2.35%
PowerDemand	197149	311629	0.81%
RESP	868000	2452953	0.19%

Table 1: Details of univariate datasets.

Dataset	Dimension	Train	Test	Anomaly Ratio
SMD	38	708377	708393	4.16%
MSL	55	58317	73729	10.50%
SMAP	25	135183	427617	12.79%
PSM	25	132481	87841	27.75%
GECCO	10	60000	60000	1.25%
SWAN	39	69260	69261	23.80%

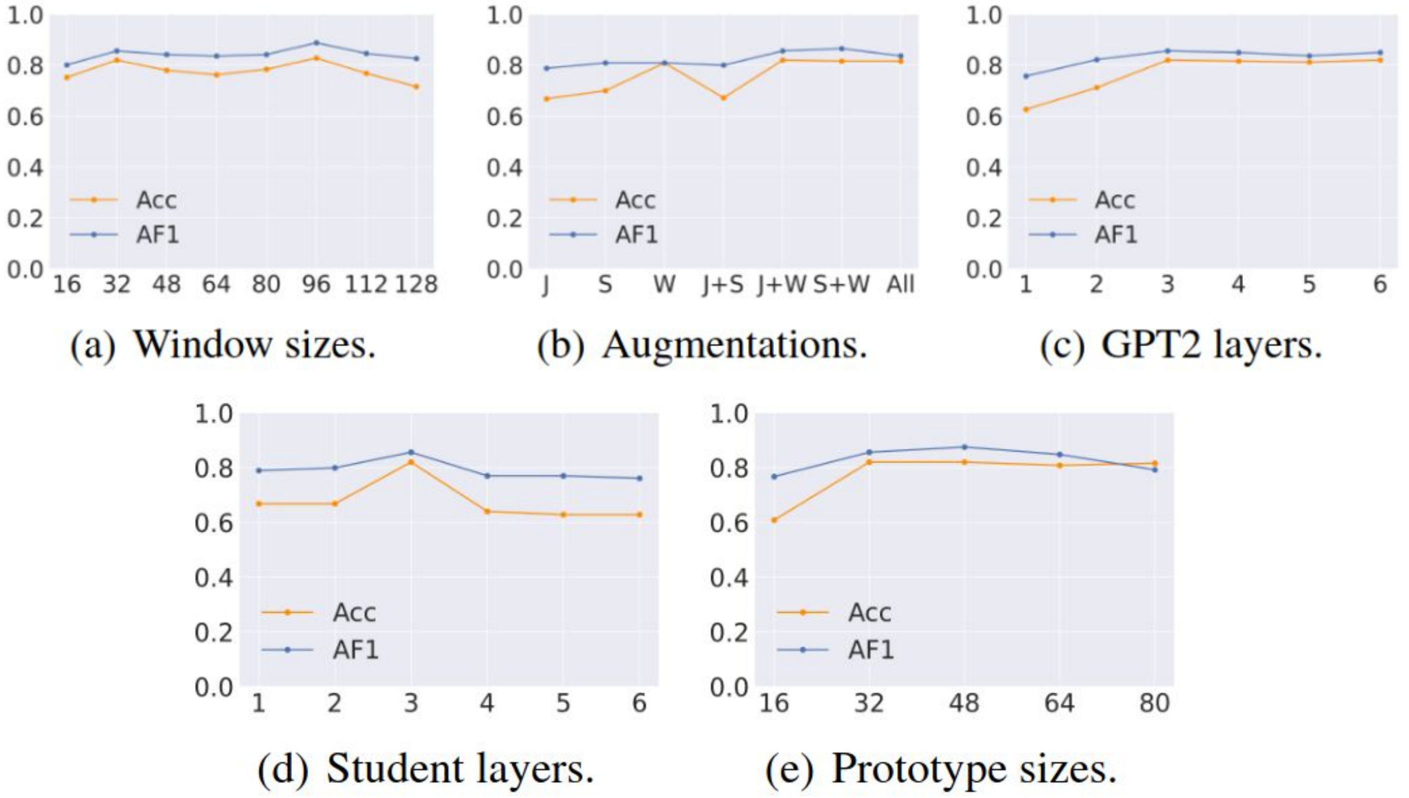
Table 2: Details of multivariate datasets.

Dataset Statistics

	WaQ			SWAN		
	P	R	F1	P	R	F1
OCSVM	0.021	0.341	0.040	0.193	0.001	0.001
MatrixProfile	0.046	0.185	0.074	0.167	0.175	0.171
GBRT	0.175	0.140	0.156	0.447	0.375	0.408
LSTM-RNN	0.343	0.275	0.305	0.527	0.221	0.312
Autoregression	0.392	0.314	0.349	0.421	0.354	0.385
IForest	0.439	0.353	0.391	0.569	0.598	0.583
AutoEncoder	0.424	0.340	0.377	0.497	0.522	0.509
AnomalyTrans	0.257	0.285	0.270	0.907	0.474	0.623
MTGFlow	0.333	0.125	0.182	**1.000**	0.494	0.662
DCdetector	0.383	0.597	0.466	0.955	0.596	0.734
Ours	**0.511**	**0.793**	**0.620**	0.873	**0.745**	**0.804**

Table 5: Overall results on the NIPS benchmark.

Parameter sensitivity studies of main hyperparameters in AnomalyLLM



Ablation Studies

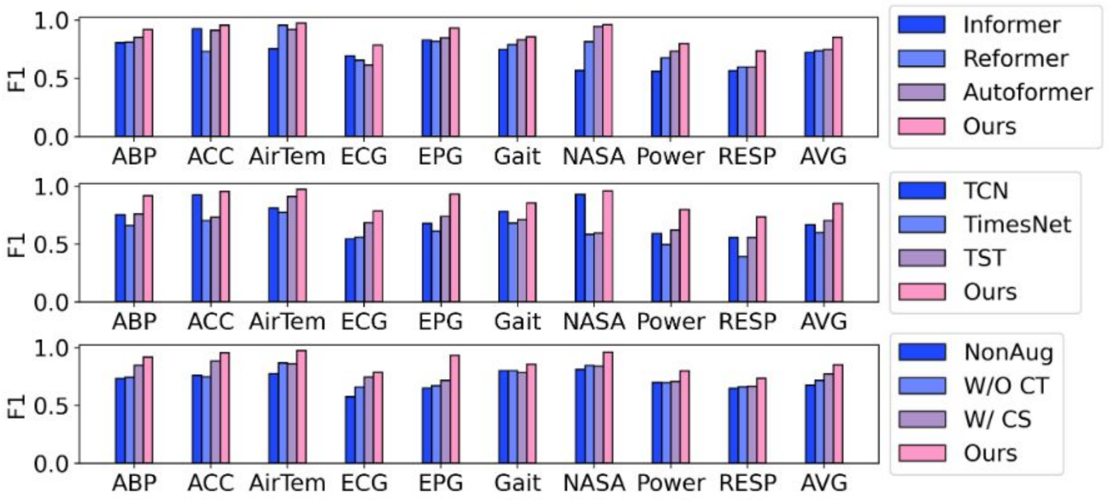


Figure 5: Ablation studies. **Top:** It depicts the performance of the model with different center predictors. **Middle:** It depicts the performance of the models with different projectors. **Bottom:** It depicts the performance of different training strategies.

Metriken

- Event-Based precision, recall, F1