Article

Time Series Dataset Survey for Forecasting with Deep Learning

Yannik Hahn *, **Tristan Langer**, **Richard Meyes** and **Tobias Meisen**

Institute for Technologies and Management of Digital Transformation (TMDT), Rainer-Gruenter-Straße 21, 42119 Wuppertal, Germany
* Correspondence: yhahn@uni-wuppertal.de

Abstract: Deep learning models have revolutionized research fields like computer vision and natural language processing by outperforming traditional models in multiple tasks. However, the field of time series analysis, especially time series forecasting, has not seen a similar revolution, despite forecasting being one of the most prominent tasks of predictive data analytics. One crucial problem for time series forecasting is the lack of large, domain-independent benchmark datasets and a competitive research environment, e.g., annual large-scale challenges, that would spur the development of new models, as was the case for CV and NLP. Furthermore, the focus of time series forecasting research is primarily domain-driven, resulting in many highly individual and domain-specific datasets. Consequently, the progress in the entire field is slowed down due to a lack of comparability across models trained on a single benchmark dataset and on a variety of different forecasting challenges. In this paper, we first explore this problem in more detail and derive the need for a comprehensive, domain-unspecific overview of the state-of-the-art of commonly used datasets for prediction tasks. In doing so, we provide an overview of these datasets and improve comparability in time series forecasting by introducing a method to find similar datasets which can be utilized to test a newly developed model. Ultimately, our survey paves the way towards developing a single widely used and accepted benchmark dataset for time series data, built on the various frequently used datasets surveyed in this paper.

Keywords: time series; forecasting; dataset; deep learning; survey

1. Introduction

Digital transformation and the internet of things (IoT) have led to an increase in interconnected devices used in a variety of different industrial fields. These devices, typically individual sensors acquiring time-encoded data, provide vast potential for automated and data-driven analysis methods from the field of machine learning and deep learning [1]. The temporal component of the data is essential to identify different patterns and to draw conclusions from the data by finding anomalies, classifying different behavior, or predicting the future temporal course of the time series, and, thus, categorizing the field of time series analytics into three primary tasks: classification, anomaly detection, and forecasting.

In this paper, we exclusively focused on one of the three main tasks, the forecasting task, which is an increasingly prominent task to be tackled with deep learning models, as shown in Figure 1. Forecasting is an exciting and relevant problem that combines the need to understand the information in a time series with predicting the most likely future given that information. Moreover, forecasting plays an essential role in supporting the process of decision-making or managing resources [2].

Despite the high relevance of research on deep learning applications for time series data, the field has not seen a revolution similar to those of computer vision (CV) and natural language processing (NLP). While deep learning has shown remarkable results in learning complex tasks in CV and NLP, traditional machine learning and specialized stochastic models are still perfectly viable and even outperform deep learning models in time series tasks in some cases [3]. However, such models have disadvantages, including the fact that

their creation requires in-depth domain knowledge and their usage often results in high computational costs [3].
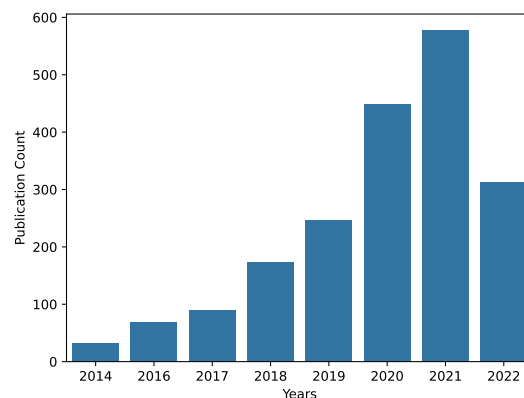

Figure 1. Illustration of the number of publications per year in the context of time series forecasting and deep learning, extracted on 22 June, 2022, on "Web of Science" [4] using the query described in Section 3.

Unlike in the field of time series analysis, the fields of CV and NLP have a number of different benchmark datasets that are frequently used to develop new models and compare cutting-edge developments with prior state-of-the-art models. Therefore, it is common practice for these advancements to be evaluated against these benchmark datasets. Two of the most prominent and widely-used datasets in CV are the MNIST [5] and ImageNet [6] datasets for training small and large models, respectively. Specifically, the ImageNet dataset has been frequently used to pre-train large state-of-the-art models that are used for transfer learning, utilizing the pre-trained features for new tasks and individual datasets. Exploiting the pre-trained features allows researchers to achieve state-of-the-art performances even on small datasets after fine-tuning the pre-trained networks. Consequentially, benchmark datasets played an essential role in the advancements of these fields.

Unfortunately, such commonly used benchmark datasets do not exist for all time series tasks, while the frequently used ones are challenged with respect to their data quality. For example, for the common task of time series classification, the UCR/UEA [7] archive provides a number of datasets, but they contain anomalies and discrepancies that can bias classification results [8]. Similarly, in the case of the common task of anomaly detection, commonly used datasets exist [9,10]. However, Wu and Keogh [11] showed that these datasets suffer from fundamental flaws. Finally, the task of time series forecasting does not benefit whatsoever from frequently used datasets, let alone benchmark datasets.

The lack of such benchmark datasets for time series analysis is likely due to the unbound character of this particular data type. Images, for example, are naturally bound by the RGB space and the size of the image, i.e., a finite amount of pixels, whereby each one is defined by three values in the closed interval of $[0, 255]$. Time series are not as well characterized, and values can be, in principal, unbound in the range of $[-\infty, \infty]$. However, practical applications and the laws of physics somewhat limit the range in which time series data is typically acquired. In turn, the vast variety of domains and sensor sources for time series data makes the ranges highly variable. Additionally, the temporal character of the data adds another layer of complexity as time series data are acquired across a large range of temporal resolutions, ranging from nanoseconds to days, months, and years. These circumstances make the development of a cross-domain and cross-task benchmark dataset for time series data highly challenging and the development has yet to emerge until this day.

As a consequence of this lack of a widely recognized benchmark dataset, existing work mainly focuses on comparing different models or within a specific domain on individual domain datasets, as we show in Section 2. Consequently, this paper aimed to provide a comprehensive overview of the current state of the art regarding openly available datasets

for time series forecasting tasks and to address their effect on the research field of deep learning for time series forecasting. The contributions of this work are fourfold:

1. We provide a cross-domain overview of existing publicly available time series forecasting datasets that have been used in research.
2. Furthermore, we analyze these datasets regarding their domain and provide file and data structure, as well as general statistical characteristics, and compare them quantitatively with each other by computing their similarity.
3. We provide an overview of all public time series forecasting datasets identified in this publication and facilitate easy access with a list of links to all these datasets.
4. Finally, we facilitate comparability in the time series forecasting research area by calculating a grouping of datasets using the aforementioned similarity measures.

2. Related Survey Publications

Before we present the results of our own comprehensive survey, we briefly review related work and surveys connected to the task of time series forecasting. For this purpose, we selected publications that reviewed several papers and analyzed how these publications describe and compare the used datasets.

In order to compare these surveys in more than just qualitative terms, we examined each for the characteristics described below, which are summarized in Table 1. First, we verified the accessibility by reviewing if a dataset was cited (Column 1, Table 1) and if any effort had been made to allow easy access through a direct link to the data (Column 2, Table 1). Furthermore, we investigated if multiple datasets were utilized (Column 3, Table 1), if these datasets were from different domains (Column 4, Table 1), and if they were compared to each other in text form or in a table (Column 5, Table 1). Next, we checked if at least two statistical values from each dataset were presented, including size, the number of dimensions, forecasting window, or time interval, shown in the columns "dataset statistics". Finally, we checked if the datasets were further analyzed in the last column. This could include a comparison by some distance metrics or identifying some characteristics.

Table 1. This table compares how the extracted surveys identified in this paper analyzed time series datasets. The tick indicates if the corresponding publication met the condition of the column.

Paper	Datasets with References	Easy access to the datasets	Multiple Datasets	Cross-domain	Comparison of Datasets	Dataset Statistics	Dataset Analysis
Ahmed et al. [12]							
Aslam et al. [13]	✓		✓		✓	✓	
Chandra et al. [14]	✓		✓	✓			
Chen et al. [15]							
Dikshit et al. [16]							
Ghalehkhondabi et al. [17]							
Lara-Benitez et al. [18]	✓		✓	✓	✓	✓	
Liu et al. [19]							
Mosavi et al. [20]			✓				
Sengupta et al. [21]	✓	✓	✓	✓			
Somu et al. [22]							
Sun and Scanlon [23]							
Wang et al. [24]							
Wei et al. [25]			✓				
Weiss et al. [26]			✓				
Zambrano et al. [27]	✓		✓				

The existing work mainly focuses on deep learning architectures, which were compared by all of the publications in Table 1. This table shows that only three out of 16 publications used datasets from multiple domains, which shows the highly domain-specific view of these publications. Furthermore, no survey satisfied all points. Especially, the analysis of datasets was not conducted by any publication, and the point of providing easy access to the cited datasets was only met by Mosavi et al. [20]. To conclude, the surveys from Table 1 show a deficit of publications investigating and analyzing datasets. Especially, a combination of a cross-domain overview, an analysis of the datasets, and easy accessibility was missing. If some statistical values were presented, they were mostly restricted to the length, a forecast horizon, or a time interval. Notwithstanding, more information is needed to make profound decisions about what datasets to use.

Lara-Benitez et al. [18] stands out by comparing multiple datasets from different domains. In addition, they list the used datasets, give a short description, and show some sample plots. This is the only publication in Table 1 that met six out of eight points. However, this publication primarily focused on comparing the model architectures. Their research did not focus on the datasets used. In contrast, this work focuses on the datasets used in time series forecasting research.

3. Methodology

3.1. Paper Screening

The primary objective was to identify all relevant datasets with their domains used in papers published in the context of time series forecasting with deep learning. Focusing on recent research in the field of deep learning ensured that the datasets used were still valid and well-known in the community, as deep learning is relatively new to the field of time series foresting.

This combination resulted in a large number of relevant papers. However, two domains could be excluded because of their specific peculiarities meaning they could be considered fields of their own. One of these domains is COVID-19, which has large amounts of publicly available data. However, world data on COVID-19 is already collected and merged by the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The other domain is the field of finance and stock market data which is widely recognized as a field of its own due to its global economic meaning. We excluded this domain due to its extensive amount of publicly available data given by the permanently active stock market and corresponding survey literature [28].

The methodology used in this work consists of three steps and is depicted in Figure 2:

1. As a first step, we performed a screening of papers found in the "Web of Science" [4] to identify publicly available time series forecasting datasets which are used in research. To ensure the impact of the datasets in research, we limited our choice to papers that had been cited at least ten times. This reduced the number of papers from over 1000 to 207 and ensured that relevant papers with common datasets were not excluded. Another goal was to identify datasets already used for deep learning. We achieved this by adding the constraint of "deep learning" to the search query. As a result, the following Web of Science query was used:

   "ALL = (Time Series) AND ALL = (Deep Learning) AND ALL = (Forecasting) NOT ALL = (Finance) NOT ALL = (COVID)" (Date of query 19 October 2021)

   We focused on examining publicly available datasets and identifying the domains of the selected papers.

2. To address the fact that newly published papers had not had the chance yo acquire ten citations to date, papers from the last year, 2021, had a restriction of having at least five citations and a maximum of ten citations as other papers were already included by step one. Therefore, we used the same Web of Science query as before and extracted 43 new publications on 17 December 2021.

3. To widen the search for datasets used in academic publications, we further utilized the website "papers with code" [29]. Then, "papers with code" was used to ensure the inclusion of recent publications from conferences that were not listed on "Web of Science". The "papers with code" website has a collection of publicly available datasets with the associated papers that have published their code and results on a dataset. Furthermore, the website ranks publications per dataset by the number of stars of their corresponding Github repositories. We filtered the datasets by the categories of "time series" and "forecasting" to collect the datasets with their corresponding publications. We selected the top 10 ranked publications if a dataset had more than ten publications, resulting in 43 additional publications with eight datasets. Then, we used these publications for additional screening of datasets to find public datasets not identified by the website "papers with code".
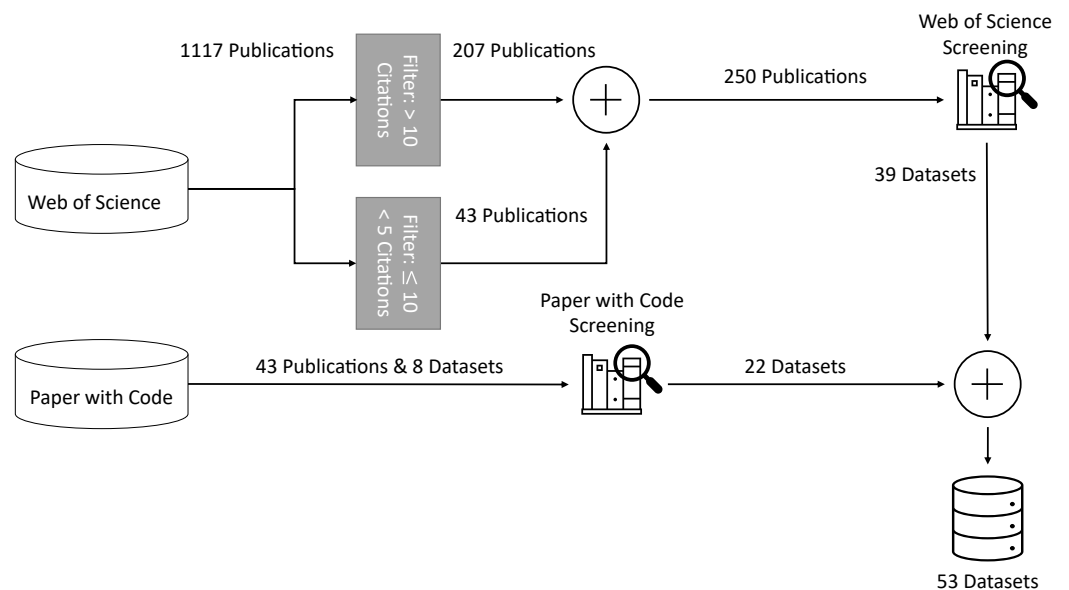

Figure 2. Visualization of the public dataset search process with the number of publications used for the review.

After combining the before-mentioned steps, we utilized 293 publications. Based on these publications, we identified 53 different public datasets. Moreover, we identified 39 datasets while screening the papers from "Web of Science" and 22 datasets with the website "papers with code". Eight of the 22 datasets overlap with the datasets from the website "papers with code".

3.2. Fundamentals of Statistical Time Series Characteristics

In this section, we briefly introduce three statistical measurements that we utilized to identify stationarity, seasonality, and whether a time series dataset consists of repeating values.

Identifying if a time series is stationary shows if the statistical properties change over time and that there is no trend or shift in the time series. This impacts the methods used to analyze the datasets or forecast values from the dataset. Moreover, it shows that the mean, variance, and covariance are constant and independent of time. Consequently, we decided to use the augmented Dickey–Filler (ADF) test [30].

The ADF test is a widely used unit–root test, which is an extended version of the Dickey–Fuller test. It is derived from an autoregressive (AR(k)) model and shows if a time series is stationary [31]. The ADF test removes the structural effects from the Dickey–Fuller test and involves the following regression:

$$\Delta x_i = \mu + \gamma t + \alpha_1 x_{t-1} + \sum_{j=2}^{k} \alpha_j \Delta x_{t-j+1} + u_t \tag{1}$$

with $t = 1, .., n$, Δ as the difference operator, u_t being white noise, and $\alpha_1 = 0$ representing the null hypothesis of the unit root test [32]. A p-value lower than 0.05 confirms the null hypothesis and indicates that a time series is stationary [33].

Most tests for seasonality need domain knowledge, a graphical analysis, or certain assumptions [34]. Therefore, we decided to use auto-correlation (AC) to identify to what degree a time series correlated with itself or a lagged version of itself. The resulting value indicated if a time series had many repeating patterns, which indicates seasonality. Moreover, the resulting value $f_{agg} = (R(1), .., R(m))$ with $m = max(n, l_{max})$ is the aggregated mean with a maximum lag l_{max} of 40 and the sequence length n [35].

$$R(l) = \frac{1}{(n-l)\sigma^2} \sum_{i=1}^{n-l} (X_t - \mu)(X_{t+l} - \mu) \tag{2}$$

Furthermore, we computed the percentage of reoccurring values (PRV) to show to what percentage a dataset consisted of repeating values. The PRV values show the percentage of how many values are not unique in a time series. This metric has the disadvantage that it depends on the dataset size because the probability of repeating values grows with the size.

$$PRV(X) = \frac{\sum_{i=0}^{m} 1[|u_i| > 0]}{m} \qquad u \in unique(X) = \{u_1, .., u_m\} \tag{3}$$

4. Time Series Domains

The high variability in research and application domains of time series analysis makes analyzing domains challenging. For example, suppose two groups of authors of two different papers came out related to a different field of research. In that case, two papers might refer to the same domain but use different wording because of specializations and domain expertise. For example, one can refer to the domain wind speed forecasting as "power load forecasting for renewable energy sources". In this case, the goal is to forecast the energy output. First, however, the wind speed or other energy-generating factors are forecast to achieve that goal.

To ensure a wide cross-domain overview, we used the publications identified in "Web of Science" to analyze domains according to the following rules. If a paper referred to multiple datasets, the domain of every dataset was counted. On the condition that a domain occurred less than three times, we checked whether the domain could be subsumed in a different domain. If not, the domain was not visualized in Figure 3 but listed in Table 2. On the contrary, if one domain occurred more than 30 times and the domain could clearly be separated into different sub-domains, they were divided into these sub-domains. For example, the weather domain could have included the domain wind speed. However, even without the publications of the wind speed domain, the weather domain had more than 30 publications. As a result, the wind speed domain was excluded and treated as a separate domain.

Figure 3 presents the different domains and their frequency in the reviewed literature. The largest domain is the electricity domain. This domain includes publications that forecast the energy output of renewable energy sources, other energy sources, and the consumption of small groups or single households. The second-largest domain is the weather domain. This domain includes multiple weather forecasting publications that highly overlap with publications of the electricity domain's renewable energy sources. Furthermore, the wind domain could be seen as a sub-domain of the weather domain. Nevertheless, many papers forecast wind speed without considering other weather-related measurements. Correspondingly, we separated the wind speed domain, the third most often screened domain. Next, the air quality domain includes publications that mostly forecast the air quality of cities. Some work published in this context used weather-related features to improve their forecasts. Therefore, the publications show a slight overlap with the weather domain.

Despite the proposed web of science query, which should have ignored the finance domain, the domain was identified as the third most common domain. Nevertheless, as previously set, we ignored publications within this domain entirely. The occurrence of finance publications was caused by publications using more specific keywords than finance.
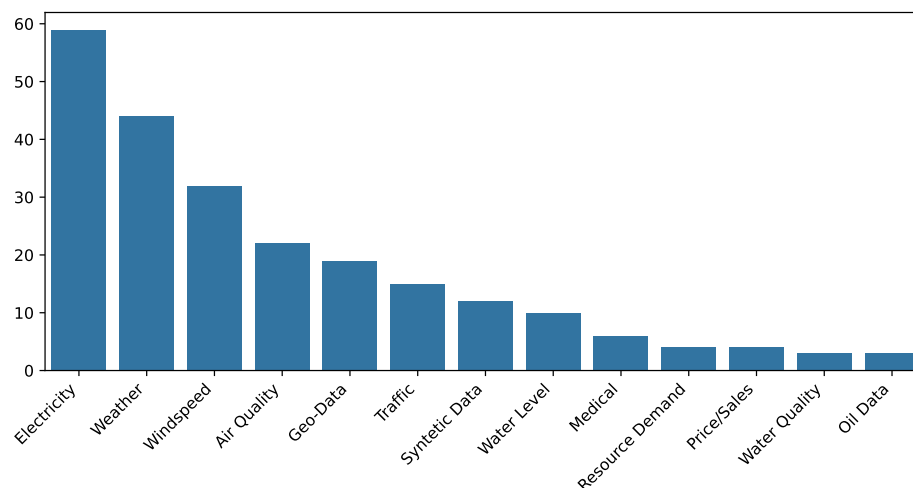

Figure 3. Domain distribution of the datasets which were identified while screening the "Web of Science" publication corpus.

The high variability in domains indicates that the time series field is mostly domain-driven and time series occur in many different domains, which is shown in Table 2. Contrary to expectations, some domains did not appear frequently. For example, we expected a high relevance of the machine sensor domain driven through industry 4.0. Nevertheless, machine sensor publications were only seen less than three times. Then again, this phenomenon was likely due to the general aversion of companies to publish their proprietary data in order to protect their intellectual property. Another reason could be that time-encoded machine sensor data is mostly used to classify different states or errors and not directly to forecast.

Table 2. A collection of domains that occurred less than three times in the paper screening.

Astro	Machine Sensor	Mortality rate	Fertility rate
Physics Simulation	Supply Chain	Tourism	NLP
Crime	Garbage/Waste Prediction	Gas Consumption	Mobile Network
Network Security	Trend Forecasting	Yield Prediction	Machine Sensor
Chemicals	Cloud Load	AD Exchange	Bike-sharing
Non-linear Problems	Web Traffic		

Furthermore, environmentally-related domains, including weather, wind speed, air quality, and geospatial data, accounted for 35% of all domains. Moreover, the top five domains accounted for around two-thirds of all papers.

5. Screening of Public Datasets

Tables 3 and 4 shows a non-domain-specific overview of public time series forecasting datasets collected by the before-described methods. Moreover, the datasets in the tables were selected with the following conditions:

1. Our first condition was that the data must be publicly accessible and not hidden behind a particular sign-in, or only available on request, to give an overview of general publicly available datasets.
2. The dataset must be directly downloadable as files to ensure reproducibility. Datasets that can only be accessed through a web view or dashboard, where multiple parameters need to be selected, were not included.

3. Datasets would not be considered if the data was only available in a specific country or the website was not in English, to ensure consistent access to the datasets.

Datasets with IDs 0 to 38 were found during the paper screening of the 250 identified papers, and all datasets with IDs higher than 38 were found using the website papers with code. The datasets with the IDs 3, 6, 23, 24, 25, 26, 29, and 38 were identified through the "Web of Science" and "papers with code" screening. These datasets are presented in combination with their relevant information on domain, structure, and how the dataset is provided. A direct link to each dataset can be found in Table 3 by using the ID of the dataset. The missing values in Table 4 were caused by the different formats in which the datasets were published. If a dataset had multiple files with different formats and dimensions, it was impossible to obtain these values. Furthermore, if an equation or algorithm generated the dataset synthetically, the attributes of the table depended on the generation and were, therefore, not relevant in this context.

We analyzed the data extracted from Table 4 by investigating the number of citations per dataset to determine if a commonly used benchmark dataset existed. A benchmark dataset is one that should have been cited by a high percentage of the publications found in Table 4, have a size that can be used for deep learning, and include multiple domains. In addition, we looked for datasets already accepted by the research community.

However, only four datasets were cited multiple times if we only considered the datasets and papers from the "Web of Science" paper screening. Moreover, the maximum number of citations from the "Web of Science" paper screening held the dataset with ID 2 with four citations. In contrast, we observed that the "papers with code" screening resulted in a maximum of 16 citations in one dataset. All datasets from Table 4 were cited on average 2.60 times.

Furthermore, the electricity dataset with ID 16 was the most cited dataset in Table 4 with 16 citations resulting from the "papers with code" screening, where one paper was already identified through the 'Web of Science' screening. The dataset is primarily used for multivariate forecasting. This indicates that the dataset with ID 16 is a common dataset for multivariate forecasting. However, the dataset only covers the electricity domain, and other datasets in the same publications are not as frequently used and vary. Consequently, it was not a benchmark dataset as defined in our work.

Table 3. This Table presents all links to the shown datasets from Table 4. The web links can be used to retrieve the before-shown datasets.

ID	Direct Link
0, 1	https://github.com/chennnnnyize/Renewables_Scenario_Gen_GAN/ (accessed on 1 February 2023)
2	https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data (accessed on 1 February 2023)
3	https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption (accessed on 1 February 2023)
4	https://archive.ics.uci.edu/ml/datasets/Air+quality (accessed on 1 February 2023)
5	https://www.microsoft.com/en-us/research/publication/forecasting-fine-grained-air-quality-based-on-big-data/?from=http%3A%2F%2Fresearch.microsoft.com%2Fapps%2Fpubs%2F%3Fid%3D246398 (accessed on 1 February 2023)
6	https://archive.ics.uci.edu/ml/datasets/PEMS-SF (accessed on 1 February 2023)
7	https://www.opendataphilly.org/dataset/crime-incidents (accessed on 1 February 2023)
8	https://archive.ics.uci.edu/ml/datasets/sml2010 (accessed on 1 February 2023)
9	http://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection (accessed on 1 February 2023)
10, 11	https://github.com/robjhyndman/demography (accessed on 1 February 2023)
12, 13	https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset (accessed on 1 February 2023)
14, 15	https://www.emc.ncep.noaa.gov/mmb/nldas/LDAS8th/forcing/forcing.shtml (accessed on 1 February 2023)
16	http://www.cs.fit.edu/~pkc/nasa/data/ (accessed on 1 February 2023)
17, 18, 19, 20, 21	https://github.com/numenta/NAB (accessed on 1 February 2023)
22	https://github.com/maziarraissi/DeepHPMs (accessed on 1 February 2023)

Table 3. *Cont.*

ID	Direct Link
23	https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
24, 25, 26	https://github.com/laiguokun/multivariate-time-series-data (accessed on 1 February 2023)
27	https://zenodo.org/record/2826939#.Ya9-JdDMI60 (accessed on 1 February 2023)
28	https://data.mendeley.com/datasets/bhjgdhgzjr/1 (accessed on 1 February 2023)
29, 30	https://www.emc.ncep.noaa.gov/mmb/nldas/LDAS8th/forcing/forcing.shtml (accessed on 1 February 2023)
31	https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports--Historical-2003/tmnf-yvry (accessed on 1 February 2023)
32	https://zenodo.org/record/1306527#.YXKIxxpBw60 (accessed on 1 February 2023)
33	https://irafm.osu.cz/cif/main.php?c=Static&page=download (accessed on 1 February 2023)
34	https://forvis.github.io/datasets/m3-data/ (accessed on 1 February 2023)
35	http://www.neural-forecasting-competition.com/downloads/NNGC1/datasets/download.htm (accessed on 1 February 2023)
36	https://www.kaggle.com/competitions/tourism2/data?select=tourism2_revision2.csv (accessed on 1 February 2023)
37	https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/data?select=train_1.csv.zip (accessed on 1 February 2023)
38	https://github.com/Mcompetitions/M4-methods (accessed on 1 February 2023)
39	https://github.com/zhouhaoyi/ETDataset (accessed on 1 February 2023)
40	https://git.opendfki.de/koochali/forgan/-/tree/master/datasets/lorenz (accessed on 1 February 2023)
41	https://www.nrel.gov/grid/solar-power-data.html (accessed on 1 February 2023)
42	https://github.com/yandex-research/shifts (accessed on 1 February 2023)
43	https://kilthub.cmu.edu/articles/dataset/Data_Collected_with_Package_Delivery_Quadcopter_Drone/12683453/1 (accessed on 1 February 2023)
44	https://theairlab.org/trajair/#download (accessed on 1 February 2023)
45	https://www.kaggle.com/sohier/30-years-of-european-wind-generation (accessed on 1 February 2023)
46	https://www.kaggle.com/datasets/city-of-seattle/seattle-burke-gilman-trail (accessed on 1 February 2023)
47	https://data.mendeley.com/datasets/byx7sztj59/1 (accessed on 1 February 2023)
48, 49, 50, 51	https://drive.google.com/drive/folders/1ZOYpTUa82_jCcxIdTmyr0LXQfvaM9vIy (accessed on 1 February 2023)
52	https://drive.google.com/drive/folders/1ohGYWWohJlOlb2gsGTeEq3Wii2egnEPR (accessed on 1 February 2023)

Table 4. A general overview of the public datasets found through the paper screening of "Web of Science" and the "papers with code" as defined in Section 3.1. The coding of the column "Data Structure" column is defined in Table 5 with the underlying structure (*FileDatastructure/DatasetDescription/Timestamp*).

ID	Domain	Data Structure	File Format	# Data Points	# Dimensions	Time Interval	Paper
0	Windspeed	(−/−/−)	csv	105,119	51	5 min	[36]
1	Electricity	(−/−/−)	csv	105,119	31	5 min	[36]
2	Air Quality	(+/+/+)	csv	43,824	12	1 h	[37–40]
3	Electricity	(+/+/+)	csv	2,075,259	8	1 min	[37,41–43]
4	Air Quality	(+/+/+)	xlsx	9471	16	1 h	[37,44]
5	Air Quality	(+/+/+)	csv	2,891,393	7	1 h	[38]
6	Traffic	(+/o/−)	txt	3,997,413	11	1 h	[2,37,45,46]
7	Crime	(+/+/+)	csv	2,678,959	15	irregular	[47]
8	Weather	(+/+/+)	txt	2764	24	15 min	[48]
9	Ozone Level	(+/o/+)	csv	2536	74	1 h	[44]
10	Fertility	(+/+/+)	rda	574	4	1 yr	[49]
11	Mortality	(+/+/+)	csv	21,201	8	1 yr	[49]
12	Weather, Bike-Sharing	(+/+/+)	csv	731	15	1 d	[50]
13	Weather, Bike-Sharing	(+/+/+)	csv	17,379	16	1 h	[50]
14	Electricity, Weather	(+/+/+)	xlsx	713	3	1 d	[48]
15	Weather	(+/+/+)	xlsx	15,072	12	1 h	[48]
16	Machine Sensor	(−/o/−)	txt	-	-	100 ms	[51]

Table 4. *Cont.*

ID	Domain	Data Structure	File Format	# Data Points	# Dimensions	Time Interval	Paper
17	AD Exchange Rate	(+/o/+)	csv	9610	3	1 h	[51]
18	Multiple	(+/o/+)	csv	69,561	3	5 min	[51]
19	Traffic	(+/o/+)	csv	15,664	3	5 min	[51]
20	Cloud Load	(+/o/+)	csv	67,740	3	5 min	[51]
21	Tweet Count	(+/o/+)	csv	158,631	3	5 min	[51]
22	Synthetic	(+/+/−)	mat	-	-		[52]
23	Electricity	(+/−/−)	txt	140,256	370	15 min	[45,53,53–67]
24	Exchange Rate	(+/−/−)	txt	7587	7	1 d	[53,54,57,58,64–67]
25	Traffic	(+/−/−)	txt	17,543	861	1 h	[53,57,58,64–67]
26	Solar	(+/−/−)	txt	52,559	136	10 min	[53,63–65]
27	Weather	(+/+/+)	csv	-	-	1 min	[68]
28	Water Level	(+/+/+)	xlsx	36,160	4	1 d	[69]
29	Air Quality	(+/+/+)	csv	420,768	19	15 min	[62,70]
30	Air Quality	(+/+/+)	csv	79,559	11	15 min	[50,71]
31	Crime	(+/+/+)	csv	2,129,525	34	1 min	[47]
32	Chemicals	(+/+/+)	xlsx	120,630	7	1 min	[72]
33	Multiple	(+/−/−)	txt	71	110	1 M.	[18]
34	Multiple	(+/+/+)	txt	167,562	3	1 yr, 1 q, 1 m	[18,73–76]
35	Traffic	(+/+/−)	xls	-	-	1 d	[18,73]
36	Tourism	(+/−/−)	csv	309	794	1 m 1 q	[18]
37	Web Traffic	(+/+/+)	csv	290,126	804	1 d	[18,73]
38	Multiple	(+/o/+)	csv	414	960	1 yr, 1 q, 1 m, 1 w, 1 d, 1 h	[2,18,45,46,73–78]
39	Machine Sensor	(+/+/+)	csv	34,840	9	1 h, 1 m	[54–60,79–81]
40	Synthetic	(−/−/−)	pickle	-	-		[82]
41	Electricity	(+/+/+)	csv	4,055,880	6	5 min, 1 h	[2,45,53,54,54,63,77]
42	Weather	(+/+/+)	csv	633,494,597	125	1 yr	[83,84]
43	Electricity	(+/−/+)	csv	257,896	27	1 h	[85,86]
44	Trajectory	(+/+/+)	txt	8,241,680	14	1 s	[87]
45	Wind	(+/+/−)	csv	262,968	254	hourly	[2,45]
46	Bike-Usage	(+/+/+)	csv	52,584	5	hourly	[77]
47	Electricity	(+/+/+)	csv	48,048	16	hourly	[80]
48	Illness	(+/+/+)	csv	966	7	weekly	[57,58,66,67]
49	Sales	(+/+/+)	csv	1,058,297	9	daily	[43]
50	Weather	(+/+/+)	csv	52,696	21	10 min	[57,58,66,67,81]
51	Traffic	(+/−/−)	mat	57,636	48	hourly	[43]
52	Weather	(+/+/+)	csv	35,064	12	hourly	[60,64]

Table 5. Description of the column "Data Structure" with (*FileDatastructure/DatasetDescription/ Timestamp*) from Table 4.

	File Data Structure	Dataset Description	Timestamp
+	One file or multiple files with a clear structure and documentation.	The dataset contains a description of every field, which could lead to an understanding of all fields.	It is a timestamp, date or any date-related column defined.
o		The dataset contains different placeholders in the data, which are explained later in a description.	
−	Multiple files in different directories without any obvious order and relation between each other.	There is no field description or an incomplete one.	There is no timestamp, date or any date-related column defined.

On the other hand, the dataset with ID 38 (M4 dataset) was used in a competition and was the second most cited dataset compared to other datasets in this work. Furthermore, the dataset combined multiple domains and consisted of 100,000 randomly sampled time series from the ForeDeCk database [88]. As a result, we found that the dataset with ID 38 was the closest dataset to a benchmark dataset. Moreover, there already exists a follow-up competition with the M5 dataset [89], which included the forecasting of 42,840 unit sales of the retail company Walmart. The M4 dataset is not a commonly used dataset,

as we identified the M4 dataset only once during our "Web of Science" paper screening, although there were more than ten citations on the "papers with code" website.

Next, we analyzed the number of public datasets in different domains to identify the essential domains for time series forecasting with publicly available data. We found no dominating domain, as no domain appeared more than eight times in the table. Moreover, "weather" was the most common domain, occurring eight times, followed closely by the electricity domain with seven occurrences and the air quality domain with five occurrences. Additionally, Table 4 shows that domains with public datasets varied and were not restricted to a few domains. Even domains identified less than three times, seen in Table 2, have public datasets.

Furthermore, we investigated the number of dimensions $d \in \mathbb{N}$, data points n with $x_i \in X = \{x_1, .., x_n\}$ where $x_i \in \mathbb{R}^d$, and time intervals to investigate if a typical pattern existed. Moreover, the number of dimensions shows that time series data is naturally multivariate. All of the datasets had at least three dimensions, where at least one dimension was temporal. Moreover, the other dimension could include locations, categorical variables, a forecasted value, and variables that present additional information. The number of data points from the datasets varied from $7 \cdot 10^2$ to $6 \cdot 10^8$, had an approximate average of $1.4 \cdot 10^7$, and an approximate median of $65.3 \cdot 10^4$. Moreover, the size or number of data points of the dataset depended on the time interval. If the time interval was daily or yearly, the dataset was most likely smaller because the data had to be recorded over a much larger time period. The two datasets with an annual time interval had an average number of 10,886 data points, while the average number of data points from the datasets with a one-minute time interval was 1,441,804.

Table 4 shows that the column "Data Structure" varies, as only 57% of the values from the data structure column exclusively had "+" values. Furthermore, the timestamp was missing in 28% of the datasets. This can be problematic if some time features, like holidays or other seasonal effects, are relevant to a forecasting task. Finally, considering the different file formats in which the datasets were published, the CSV format was the most used format with 62%, and the second most used format was the text format, which included data in a table structure similar to the CSV format. As a result, we could conclude that there is currently no established standard way of publishing a time series dataset.

Furthermore, we identified multiple challenges while screening the selected papers for datasets, like non-unique names of datasets from different domains, different parts of a publication where the datasets were described and cited, a false web link, or no link to the dataset. Consequently, it would be helpful for reproducibility to refer directly to the dataset used. Furthermore, the data's availability could increase the subject's visibility and the work done. The missing publicly available data could be caused by restrictions made by the data owner if the data came from a company because the company may not want to give their competitors insights. Another reason could be privacy concerns if user data is involved. If an author only wanted to show that deep learning has the potential to improve forecasting in his research field, it would not be his first concern to make the data publicly available. On the other hand, authors who wanted to introduce a new deep learning model should have comparability in mind and use a publicly available dataset or make their datasets publicly available.

Nevertheless, sometimes, publications only describe what the data is and which features are present in the data. Some other papers, for example [90,91], describe where they extracted the data and which parameter they used to select their training and testing data. These papers would be reproducible in principle if the data sources were still available. However, in some cases, the linked data source can be unavailable in a particular country or not available anymore in general. Hence, it is impossible to access the data with the original URL. An example is a company or institute with a changed web domain or website structure. Another option to cite a dataset is to cite a publication that introduced the dataset. However, this could also be problematic if the cited paper does not have a direct link to the referred to dataset.

To summarize, Table 4 shows that there are some large datasets publicly available that are already used in the context of deep learning. Nevertheless, these datasets are widely spread between the different domains. Due to the high percentage of public datasets that come out of domains with less than three publications and many publications only using one dataset from one domain, we concluded that the time series forecasting research field is primarily domain-driven. Additionally, we identified that publishing datasets are not common in time series forecasting and that there are multiple reasons why authors are not publishing their datasets. Finally, we determined that even though some datasets are more frequently used, a commonly used benchmark dataset does not exist.

6. Comparison of Selected Datasets

In this section, we aimed to extract general patterns or statistics of the dataset to identify commonalities and differences between the datasets. We chose to rely exclusively on statistics and distances that could be computed automatically without any manual effort. To give a better overview of the available datasets and the underlying structure of the data, we analyzed datasets by multiple statistical methods. Only a subset from Table 4 was used because the datasets and their description had to meet the following requirements:

1. The forecast value must be clearly defined in a paper or a dataset description.
2. The defined forecasting value should not be aggregated over a period of time or locations.
3. For comparability, the target must be a univariate time series.

For all further analysis, the determined dimension of the forecasting value was used for comparison with other datasets, which can be found in the "Forecasting Value" column of Table 6. Even though one dimension does not represent the full datasets and their characteristics, the target dimension is the most important dimension in a forecasting task, and restriction is needed to ensure comparability.

6.1. Comparison of Selected Datasets with MPdist

We decided to use the MPdist measurement to compute the distance from all datasets to each other. The MPdist measurement can be used for different time series lengths and is robust against spikes, dropouts, and wandering baseline [92]. MPdist shows if two time series share similar subsequences under the euclidean distance [92]. We utilized MPdist to visualize the normalized distance between all datasets in Figure 4. The distances resulted from first computing the maximum subsequent window using Matrix Profile [93] and then using the window to compute the MPdist normalized by the window size. Figure 4 illustrates the sorted distances of the datasets. We sorted the heatmap in both dimensions so that the datasets with the lowest distances were in the bottom left corner, and the highest distances were on the top right side. MPdist is not a symmetrical distance measurement, so it results in different distances depending on the direction of the distance computations. Moreover, if we computed the distances between two datasets X_0 and X_1, we computed $d_{MPdist}(X_0, X_1, window_{X_0})$ and $d_{MPdist}(X_1, X_0, window_{X_1})$.

Furthermore, the black colors of Figure 4 show that the datasets were similar under the MPdist, and the light orange colors show that the datasets did not share many subsequences.

Figure 4 shows that datasets 17, 28, and 4 were the datasets with the lowest distance to all other datasets, which is presumably correlated to common patterns in these datasets that can be found in multiple of the other datasets with their corresponding window size. Nevertheless, these datasets were not the ones to which most other datasets were the closest, meaning that despite their common patterns, the other datasets did not have patterns with their corresponding window size that could fit into these datasets. Figure 5 presents samples from all datasets. These samples show that datasets 4, 17, and 28 had frequent short peaks, which could be fitted well into the other datasets.
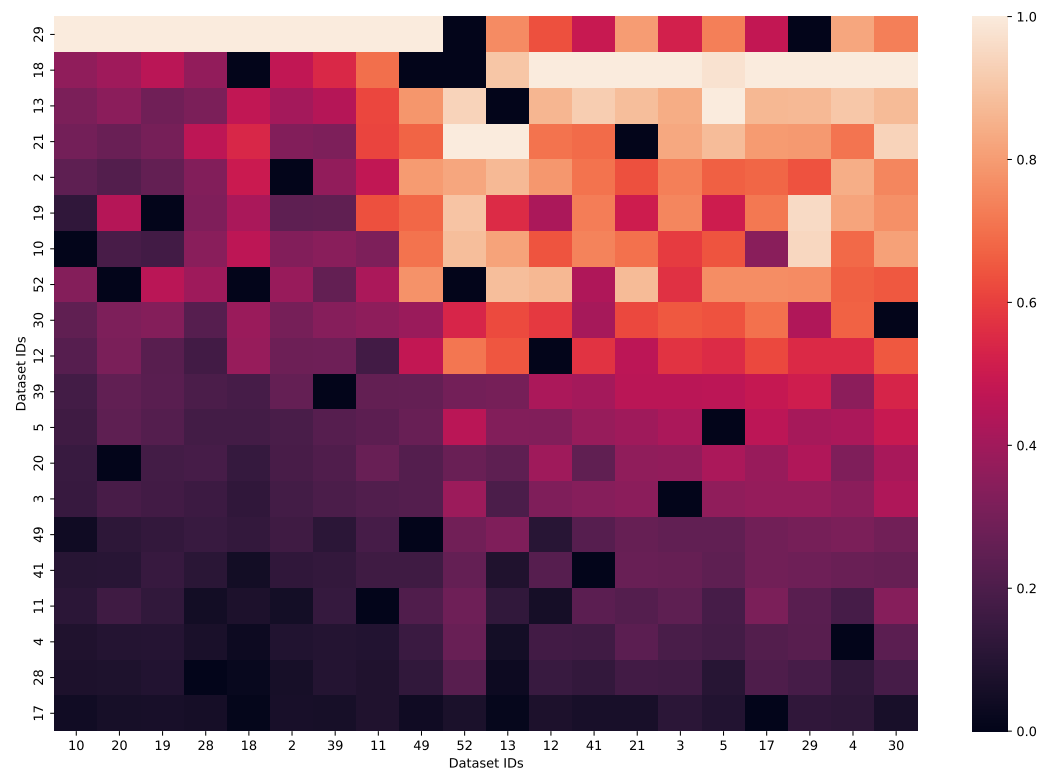
Figure 4. Heatmap of the sorted MP-distances of all datasets compared to each other.

Datasets 10, 20, and 19 were the datasets to which most other datasets were the closest. This could be caused by patterns of these datasets that are common in other datasets regardless of a certain window size. The dataset with ID 10 was an outlier due to its small size. Moreover, combining the small dataset size with large window sizes of other datasets could lead to some bias in the results and should be carefully considered.

Furthermore, datasets with IDs 18 and 29 had high distances to most other datasets. The dataset with ID 18 contained data from multiple domains, and combined with the one computed window size could lead to high distances. The dataset with the ID 29 had longer irregular peaks, which might not be fitted into different datasets well. Moreover, this dataset was the dataset to which the other datasets had the third highest distance. This indicated that these irregular peaks did not match with the other datasets in both directions.

The MPdist measurement enabled us to compare datasets directly to one another. Nevertheless, we needed more than the distances to draw general insights or to group the datasets. Therefore, we utilized the MPdist distance matrix for clustering in Section 7 to draw more general insights.
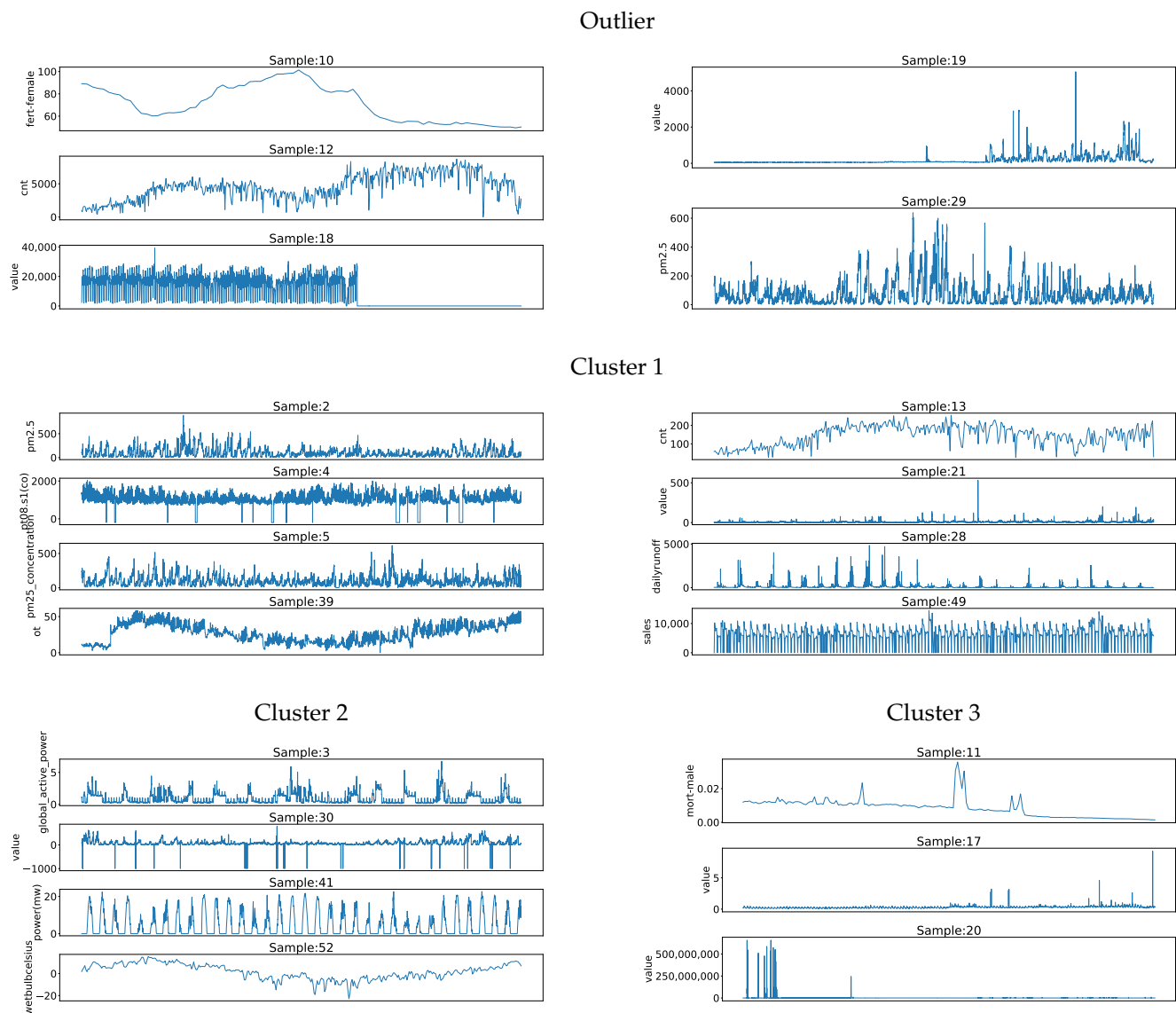
Figure 5. Samples from all dataset from Table 6 with a maximum length of 100 k.

6.2. Comparison of Selected Datasets with Statistical Characteristics

The before-shown MPdist measurements only compare the datasets directly against each other, and, when adding a new dataset, all distances to and from the new dataset must be computed. This makes a fast comparison to new datasets impossible. Therefore, we decided to use additional statistical values to enable other researchers to compare their datasets and find similar datasets without having to rely on extensive computation. Unfortunately, there are no commonly used advanced statistics that are applicable and could describe a complete time series dataset. As a result, we focused on three different statistics representing characteristics like trend, seasonality, and repeating values, which are described in Section 3.2.

To compute the statistical values ADF, AC, and PRV, the dataset column with the forecasting values was used to calculate the metrics with the ts-fresh library [94]. If a dataset exceeded the number of 1M data points, a representative contiguous sample of the size with 1M data points was used for the computation.

Table 6 shows that around 90% of the datasets were stationary. Only the datasets with IDs 10, 12, and 18 were not stationary. Furthermore, the datasets with the IDs 10 and 12 were the smallest datasets with less than 1000 data points. Datasets with a low

number of data points could include more likely trends if the datasets were not recorded for an extended period. Then again, the dataset with ID 18 was larger and had nearly 70k data points. Next, the auto-correlation values showed a high variance in the datasets. The AC values started at 0.02 and went up to 0.86 with a mean of 0.41 and a standard deviation of 0.29. The AC values were mostly equally distributed, except for a small focus of values near zero. Finally, the PRV showed two distributions with one outlier. One distribution included high PV values over 0.75, and the other included low values under 0.1. Furthermore, the outlier with ID 19 did not lie in these two distributions, with a PRV value of 0.55. This indicated characteristic differences in the studied datasets, which could be grouped or categorized.

Table 6. Overview of statistical features of public dataset time series forecasting datasets.

ID	Time Interval	Domain	# Data Points	# Dimensions	Forecasting Value	ADF	AC	PRV
2	1 h	Air Quality	43,824	12	pm2.5	0.0000	0.4332	0.8795
3	1 min	Electricity	2,075,259	8	global_active_power	0.0000	0.7028	0.9088
4	1 h	Air Quality	9471	16	pt08.s1(co)	0.0000	0.4333	0.8714
5	1 h	Air Quality	2,891,393	7	pm25_concentration	0.0000	0.3299	0.9154
10	1 yr	Fertility	574	4	fert-female	0.2438	0.3113	0.2000
11	1 yr	Mortality	21,201	8	mort-female	0.0179	0.0851	0.0896
12	1 d	Weather, Bike-Sharing	731	15	cnt	0.3427	0.6827	0.0503
13	1 h	Weather, Bike-Sharing	17,379	16	cnt	0.0000	0.0963	0.8872
17	1 h	AD Exchange Rate	9610	3	value	0.0032	0.0085	0.0000
18	5 min	Multiple	69,561	3	value	0.0000	0.9195	0.0000
19	5 min	Traffic	15,664	3	value	0.0000	0.0852	0.5500
20	5 min	Cloud Load	67,740	3	value	0.0000	0.0496	0.0584
21	5 min	Tweet Count	158,631	3	value	0.0000	0.1992	0.7619
28	1 d	Water Level	17,543	861	dailyrunoff	0.0000	0.1435	0.8597
29	15 min	Air Quality	52,559	136	pm2.5	0.0000	0.8577	0.7647
30	15 min	Air Quality	79,559	11	value	0.0000	0.5948	0.8852
39	1 h	Machine Sensor	34,840	9	ot	0.0052	0.5594	0.8544
41	5 min, 1 h	Electricity	4,634,040	6	power(mw)	0.0000	0.6134	0.9773
49	1 d	Sales	1,058,297	9	sales	0.0000	0.2742	0.8488
52	1 h	Weather	35,064	12	wetbulbcelsius	0.0000	0.7422	0.9557

7. Categorize the Datasets

As we have discussed, time series forecasting of datasets out of different domains exists. However, there are multiple domains in which only one or no public time series forecasting dataset exists. This makes it difficult to compare model performance as the similarity of the used datasets to train and evaluate these models cannot be investigated. Therefore, we introduced a method to find datasets with similar characteristics to enable researchers to publish comparable results if a used dataset cannot be published due to, for example, privacy concerns or company restrictions. Then, these found datasets can be used in addition to the original dataset to publish results that are reproducible for other researchers.

We identified clusters from the already computed ADF, AC, PRV, and MPdist values. We utilized the density-based DB-scan [95] clustering algorithm with the hyperparameters $\epsilon = 0.15$ and $min_{samples} = 3$, determined by an initial visual-assisted hyperparameter search. Moreover, we used AC, ADF, PRV, and MPdist, weighed equally, as input for the DB-scan algorithm. Using MPdist as an additional distance for the clustering algorithm had the advantage of enforcing similar patterns under the euclidian distance in the later-derived categories. Table 7 presents an overview of the values from Table 6, which are divided into four different clusters with corresponding outliers.

The clustering resulted in five outliers which can be seen in the first part of Table 7. Furthermore, the first cluster was the largest cluster, including eight datasets. Moreover, the second cluster included four datasets, and the third cluster had three datasets. Interestingly, three out of four datasets from the air quality domain could be found in cluster one, suggesting some degree of similarity between these datasets given their domain. However, it was also the largest cluster, and only the combination of weather and bike-sharing and electricity appeared in addition to the air quality domain more than once. Furthermore, the second cluster included the electricity domain twice, and the third cluster consisted only of different domains.

Table 7. Overview of the identified clusters and the corresponding statistical features of the dataset.

ID	Domain	Time Interval	# Data Points	# Dimensions	ADF	AC	PRV
			Outliers				
10	Fertility	1 yr	574	4	0.2438	0.3113	0.2000
12	Weather, Bike-Sharing	1 d	731	15	0.3427	0.6827	0.0503
18	Multiple	5 min	69,561	3	0.0000	0.9195	0.0000
19	Traffic	5 min	15,664	3	0.0000	0.0852	0.5500
29	Air Quality	15 min	420,768	19	0.0000	0.8577	0.7647
			Cluster 1				
2	Air Quality	1 h	43,824	12	0.0000	0.4332	0.8795
4	Air Quality	1 h	9471	16	0.0000	0.4333	0.8714
5	Air Quality	1 h	2,891,393	7	0.0000	0.3299	0.9154
13	Weather, Bike-Sharing	1 h	17,379	16	0.0000	0.0963	0.8872
21	Tweet Count	5 min	158,631	3	0.0000	0.1992	0.7619
28	Water Level	1 d	36,160	4	0.0000	0.1435	0.8597
39	Machine Sensor	1 h, 1 m	34,840	9	0.0052	0.5594	0.8544
49	Sales	1 d	1058297	9	0.0000	0.2742	0.8488
			Cluster 2				
3	Electricity	1 min	2,075,259	8	0.0000	0.7028	0.9088
30	Air Quality	15 min	79,559	11	0.0000	0.5948	0.8852
41	Electricity	5 min, 1 h	4,826,760	6	0.0000	0.6134	0.9773
52	Weather	1 h	35064	12	0.0000	0.7422	0.9557
			Cluster 3				
11	Mortality	1 yr	21,201	8	0.0179	0.0851	0.0896
17	AD Exchange Rate	1 h	9610	3	0.0032	0.0085	0.0000
20	Cloud Load	5 min	67,740	3	0.0000	0.0496	0.0584

Furthermore, the second cluster's mean number of data points was the largest, with 1,705,981, followed by the first cluster, with a mean of 528,922 data points. Cluster three had the lowest average data points of 32,850. Both non-stationary datasets could be found in the outliers.

The first cluster had AC values lower than 0.56 and PRV values higher than 0.76. Then, the second cluster had AC values higher than 0.59 and PRV values higher than 0.88. Next, the third cluster had PRV and AC values below 0.1. Based on these clustering results, we defined the ranges of categories. First, we defined high PRV values as higher than 0.75, low values as lower than 0.25, and medium values as within the range of low and high. AC values were high if the value was higher than 0.59 and low to medium if the values were lower than 0.59. We combined low and medium for AC because the clustering suggested that a clear border between the groups did not exist.

Thus, we derived the following categories, which include a grouping of similar characteristics.

1. **stationary/high PRV/low to medium AC**: This category is a time series that is stationary and has many repeating values which are not distributed in regular patterns or distributed in some regular patterns. Similar datasets could be found in cluster one.
2. **stationary/high PRV/high AC**: This category is a time series that is stationary and has many repeating values which are distributed in regular patterns. Similar datasets could be found in cluster two.
3. **stationary/low PRV/low AC**: This category is a time series that is stationary and has many unique values which are distributed in irregular patterns. Similar datasets could be found in cluster three.
4. **stationary/low PRV/high AC**: This category is a time series that is stationary and has many unique values which are distributed in regular patterns. We only identified the dataset with ID 18 in the outlier cluster. This indicates that this category does not naturally appear in datasets used in research. This could be caused by the multiple domains which are combined in that dataset.
5. **non stationary**: Due to the small number of datasets we identified which were not stationary, this category could not be used for comparison. It is possible that there were multiple additional clusters that we did not identify. Nevertheless, the work done in this paper could be an indicator that there are not many stationary time series datasets for forecasting used in publications.

These categories can be utilized to find a public dataset with similar characteristics by computing the AC, ADF value, and PRV on the original dataset and categorizing the resulting values. We provide the code to compute the AC, ADF, and PRV values on GitHub (https://github.com/tmdt-buw/Time-Series-Dataset-Survey-4-Forecasting-with-Deep-Learning) (accessed on 1 February 2023).

8. Conclusions

In this paper, we reviewed publicly available datasets used in publications in the field of time series forecasting using deep learning. We provided a cross-domain overview of the different time series forecasting datasets published in the context of deep learning. Furthermore, we analyzed these datasets regarding their domain, file, and data structure, as well as statistical characteristics and similarity measures, to enable other researchers to choose a dataset on a profound base of knowledge. Additionally, we provided links to all of these datasets to facilitate easy access. Finally, we categorized datasets and provided a method to find similar datasets within a group of similar characteristics, which can be utilized to publish comparable results if the researcher cannot publish their datasets.

The reviewed studies showed that many different time series domains have available public datasets. We did not find a single domain from which most of the public datasets originated. However, a big part of the research is still domain-driven. As a result, publications dealing with time series forecasting use different datasets, leading to a lack of comparability. This may be part of the slowdown in deep learning progress in this area. Even if publications used multiple datasets to test their models, they differed so that there were no commonly used datasets.

To summarize, we identified the research gap of a strongly needed general time series forecasting benchmark dataset, which would improve the progress made in the field of time series forecasting. Furthermore, this analysis shows five categories of datasets. Thus, to construct a representative benchmark dataset, one should consider covering a combination of these categories in a conglomerate of datasets so that all possibilities are included. Likewise, this work can be seen as a first step towards creating a representative cross-domain benchmark dataset for forecasting, focusing on providing an overview of the current state of research. This current state has shown a lack of non-stationary and weakly patterned datasets, which should be included in a benchmark dataset. Therefore, the next steps of creating a benchmark dataset include researching which datasets are used outside scientific publishing and combining multiple datasets from different domains with different characteristics to incorporate a representative cross-domain benchmark dataset.

More statistical measurements could extend the method of finding datasets with similar characteristics to cover additional characteristics of time series in future work.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

References

1. Längkvist, M.; Karlsson, L.; Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* **2014**, *42*, 11–24. [CrossRef]
2. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5243–5253.
3. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
4. Web of Science. Available online: https://www.webofscience.com/wos/woscc/basic-search (accessed on 19 October 2021).
5. Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **2012**, *29*, 141–142. [CrossRef]
6. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
7. Chen, Y.; Keogh, E.; Hu, B.; Begum, N.; Bagnall, A.; Mueen, A.; Batista, G. The UCR Time Series Classification Archive. 2015. Available online: www.cs.ucr.edu/~eamonn/time_series_data/ (accessed on 1 February 2023).
8. Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **2017**, *31*, 606–660. [CrossRef] [PubMed]
9. Laptev, S.A.N.; Billawala, Y. S5-A Labeled Anomaly Detection Dataset, version 1.0 (16M). Available online: https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&%20did=70&guccounter=1 (accessed on 1 February 2023).
10. Ahmad, S.; Lavin, A.; Purdy, S.; Agha, Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* **2017**, *262*, 134–147. [CrossRef]
11. Wu, R.; Keogh, E. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022.
12. Ahmed, R.; Sreeram, V.; Mishra, Y.; Arif, M.D. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renew. Sustain. Energy Rev.* **2020**, *124*, 109792. [CrossRef]
13. Aslam, S.; Herodotou, H.; Mohsin, S.M.; Javaid, N.; Ashraf, N.; Aslam, S. A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renew. Sustain. Energy Rev.* **2021**, *144*, 110992. [CrossRef]
14. Chandra, R.; Goyal, S.; Gupta, R. Evaluation of Deep Learning Models for Multi-Step Ahead Time Series Prediction. *IEEE Access* **2021**, *9*, 83105–83123. [CrossRef]
15. Chen, C.H.; Kung, H.Y.; Hwang, F.J. Deep Learning Techniques for Agronomy Applications. *Agronomy* **2019**, *9*, 142. [CrossRef]
16. Dikshit, A.; Pradhan, B.; Alamri, A.M. Pathways and challenges of the application of artificial intelligence to geohazards modelling. *Gondwana Res.* **2021**, *100*, 290–301. [CrossRef]
17. Ghalehkhondabi, I.; Ardjmand, E.; Young, W.A.; Weckman, G.R. Water demand forecasting: Review of soft computing methods. *Environ. Monit. Assess.* **2017**, *189*, 313. [CrossRef] [PubMed]

18. Lara-Benitez, P.; Carranza-Garcia, M.; Riquelme, J.C. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *Int. J. Neural Syst.* **2021**, *31*, 2130001. [CrossRef]

19. Liu, H.; Yan, G.; Duan, Z.; Chen, C. Intelligent modeling strategies for forecasting air quality time series: A review. *Appl. Soft Comput. Soft Comput.* **2021**, *102*, 106957. [CrossRef]

20. Mosavi, A.; Salimi, M.; Ardabili, S.F.; Rabczuk, T.; Shamshirband, S.; Varkonyi-Koczy, A.R. State of the Art of Machine Learning Models in Energy Systems, a Systematic Review. *Energies* **2019**, *12*, 1301. [CrossRef]

21. Sengupta, S.; Basak, S.; Saikia, P.; Paul, S.; Tsalavoutis, V.; Atiah, F.; Ravi, V.; Peters, A. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowl.-Based Syst.* **2020**, *194*, 105596. [CrossRef]

22. Somu, N.; Raman, G.M.R.; Ramamritham, K. A deep learning framework for building energy consumption forecast. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110591. [CrossRef]

23. Sun, A.Y.; Scanlon, B.R. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environ. Res. Lett.* **2019**, *14*, 073001. [CrossRef]

24. Wang, H.; Liu, Y.; Zhou, B.; Li, C.; Cao, G.; Voropai, N.; Barakhtenko, E. Taxonomy research of artificial intelligence for deterministic solar power forecasting. *Energy Convers. Manag.* **2020**, *214*, 112909. [CrossRef]

25. Wei, N.; Li, C.; Peng, X.; Zeng, F.; Lu, X. Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. *J. Pet. Sci. Eng.* **2019**, *181*, 106187. [CrossRef]

26. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. [CrossRef]

27. Zambrano, F.; Vrieling, A.; Nelson, A.; Meroni, M.; Tadesse, T. Prediction of drought-induced reduction of agricultural productivity in Chile from MODIS, rainfall estimates, and climate oscillation indices. *Remote Sens. Environ.* **2018**, *219*, 15–30. [CrossRef]

28. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]

29. Paper with Code. Available online: https://paperswithcode.com/ (accessed on 25 February 2022).

30. Cheung, Y.W.; Lai, K.S. Lag order and critical values of the augmented Dickey–Fuller test. *J. Bus. Econ. Stat.* **1995**, *13*, 277–280.

31. Nason, G.P. Stationary and non-stationary time series. *Stat. Volcanol.* **2006**, *60*, 129–142.

32. Cheung, Y.W.; Lai, K.S. Power of the augmented dickey-fuller test with information-based lag selection. *J. Stat. Comput. Simul.* **1998**, *60*, 57–65. [CrossRef]

33. Mushtaq, R. Augmented dickey fuller test. *Econom. Math. Methods Program. Ejournal* **2011**. [CrossRef]

34. Moineddin, R.; Upshur, R.E.; Crighton, E.; Mamdani, M. Autoregression as a means of assessing the strength of seasonality in a time series. *Popul. Health Metrics* **2003**, *1*, 10. [CrossRef]

35. Percival, D.B. Three curious properties of the sample variance and autocovariance for stationary processes with unknown mean. *Am. Stat.* **1993**, *47*, 274–276.

36. Chen, Y.; Wang, Y.; Kirschen, D.; Zhang, B. Model-Free Renewable Scenario Generation Using Generative Adversarial Networks. *IEEE Trans. Power Syst.* **2018**, *33*, 3265–3275. [CrossRef]

37. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Multivariate time series forecasting via attention-based encoder-decoder framework. *Neurocomputing* **2020**, *388*, 269–279. [CrossRef]

38. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. *IEEE Trans. Knowl. Data Eng.* **2021**, *33*, 2412–2424. [CrossRef]

39. Li, T.; Hua, M.; Wu, X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). *IEEE Access* **2020**, *8*, 26933–26940. [CrossRef]

40. Huang, G.; Li, X.; Zhang, B.; Ren, J. PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. *Sci. Total. Environ.* **2021**, *768*, 144516. [CrossRef] [PubMed]

41. Kim, T.Y.; Cho, S.B. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* **2019**, *182*, 72–81. [CrossRef]

42. Xu, W.; Peng, H.; Zeng, X.; Zhou, F.; Tian, X.; Peng, X. A hybrid modelling method for time series forecasting based on a linear regression model and deep learning. *Appl. Intell.* **2019**, *49*, 3002–3015. [CrossRef]

43. Jin, X.; Park, Y.; Maddix, D.; Wang, H.; Wang, Y. Domain adaptation for time series forecasting via attention sharing. In *Proceedings of the International Conference on Machine Learning, Paris, France, 29–31 April 2022*; PMLR: London, UK, 2022; pp. 10280–10297.

44. Kim, K.; Kim, D.K.; Noh, J.; Kim, M. Stable Forecasting of Environmental Time Series via Long Short Term Memory Recurrent Neural Network. *IEEE Access* **2018**, *6*, 75216–75228. [CrossRef]

45. Wu, S.; Xiao, X.; Ding, Q.; Zhao, P.; Wei, Y.; Huang, J. Adversarial sparse transformer for time series forecasting. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17105–17115.

46. Alexandrov, A.; Benidis, K.; Bohlke-Schneider, M.; Flunkert, V.; Gasthaus, J.; Januschowski, T.; Maddix, D.C.; Rangapuram, S.; Salinas, D.; Schulz, J.; et al. GluonTS: Probabilistic Time Series Models in Python. *arXiv* **2019**, arXiv:1906.05264.

47. Feng, M.; Zheng, J.; Ren, J.; Hussain, A.; Li, X.; Xi, Y.; Liu, Q. Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. *IEEE Access* **2019**, *7*, 106111–106123. [CrossRef]

48. Fang, K.; Shen, C.; Kifer, D.; Yang, X. Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental US Using a Deep Learning Neural Network. *Geophys. Res. Lett.* **2017**, *44*, 11030–11039. [CrossRef]

49. Nigri, A.; Levantesi, S.; Marino, M.; Scognamiglio, S.; Perla, F. A Deep Learning Integrated Lee-Carter Model. *Risks* **2019**, *7*, 33. [CrossRef]

50. Sagheer, A.; Kotb, M. Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. *Sci. Rep.* **2019**, *9*, 19038. [CrossRef]

51. Munir, M.; Siddiqui, S.A.; Dengel, A.; Ahmed, S. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. *IEEE Access* **2019**, *7*, 1991–2005. [CrossRef]

52. Raissi, M. Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations. *J. Mach. Learn. Res.* **2018**, *19*, 357.

53. Shih, S.Y.; Sun, F.K.; Lee, H.y. Temporal pattern attention for multivariate time series forecasting. *Mach. Learn.* **2019**, *108*, 1421–1441. [CrossRef]

54. Liu, M.; Zeng, A.; Lai, Q.; Xu, Q. Time Series is a Special Sequence: Forecasting with Sample Convolution and Interaction. *arXiv* **2021**, arXiv:2106.09305.

55. Madhusudhanan, K.; Burchert, J.; Duong-Trung, N.; Born, S.; Schmidt-Thieme, L. Yformer: U-Net Inspired Transformer Architecture for Far Horizon Time Series Forecasting. *arXiv* **2021**, arXiv:2110.08255

56. Shen, L.; Wang, Y. TCCT: Tightly-Coupled Convolutional Transformer on Time Series Forecasting. *Neurocomputing* **2022**, *480*, 131–145. [CrossRef]

57. Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; Hoi, S. Etsformer: Exponential smoothing transformers for time-series forecasting. *arXiv* **2022**, arXiv:2202.01381.

58. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.

59. Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; Xu, B. TS2Vec: Towards Universal Representation of Time Series. *arXiv* **2021**, arXiv:2106.10466.

60. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 11106–11115.

61. Deng, J.; Chen, X.; Jiang, R.; Song, X.; Tsang, I.W. A Multi-view Multi-task Learning Framework for Multi-variate Time Series Forecasting. *arXiv* **2021**, arXiv:2109.01657.

62. Du, W.; Côté, D.; Liu, Y. Saits: Self-attention-based imputation for time series. *Expert Syst. Appl.* **2023**, *219*, 119619. [CrossRef]

63. Lai, G.; Chang, W.C.; Yang, Y.; Liu, H. Modeling long-and short-term temporal patterns with deep neural networks. In Proceedings of the The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 95–104.

64. Minhao, L.; Zeng, A.; Chen, M.; Xu, Z.; Qiuxia, L.; Ma, L.; Xu, Q. SCINet: Time Series Modeling and Forecasting with Sample Convolution and Interaction. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.

65. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; Zhang, C. Connecting the dots: Multivariate time series forecasting with graph neural networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 23–27 August 2020; pp. 753–763.

66. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the International Conference on Machine Learning, Paris, France, 29–31 April 2022; PMLR: London, UK, 2022; pp. 27268–27286.

67. Liu, Y.; Wu, H.; Wang, J.; Long, M. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November 2022.

68. Wojtkiewicz, J.; Hosseini, M.; Gottumukkala, R.; Chambers, T.L. Hour-Ahead Solar Irradiance Forecasting Using Multivariate Gated Recurrent Units. *Energies* **2019**, *12*, 4055. [CrossRef]

69. Zuo, G.; Luo, J.; Wang, N.; Lian, Y.; He, X. Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. *J. Hydrol.* **2020**, *585*, 124776. [CrossRef]

70. Samal, K.K.R.; Babu, K.S.; Das, S.K. Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach. *Urban Clim.* **2021**, *36*, 100800. [CrossRef]

71. Zhang, Z.; Zeng, Y.; Yan, K. A hybrid deep learning technology for PM2.5 air quality forecasting. *Environ. Sci. Pollut. Res.* **2021**, *28*, 39409–39422. [CrossRef]

72. Harutyunyan, H.; Khachatrian, H.; Kale, D.C.; Ver Steeg, G.; Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **2019**, *6*, 96. [CrossRef]

73. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [CrossRef]

74. Kang, Y.; Hyndman, R.J.; Li, F. GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. *Stat. Anal. Data Mining Asa Data Sci. J.* **2020**, *13*, 354–376. [CrossRef]

75. Ng, E.; Wang, Z.; Chen, H.; Yang, S.; Smyl, S. Orbit: Probabilistic Forecast with Exponential Smoothing. *arXiv* **2021**, arXiv:2004.08492v4.

76. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. *arXiv* **2020**, arXiv:1905.10437v4.

77. Bhatnagar, A.; Kassianik, P.; Liu, C.; Lan, T.; Yang, W.; Cassius, R.; Sahoo, D.; Arpit, D.; Subramanian, S.; Woo, G.; et al. Merlion: A Machine Learning Library for Time Series. *arXiv* **2020**, arXiv:2109.09265v1.

78. Redd, A.; Khin, K.; Marini, A. Fast ES-RNN: A GPU Implementation of the ES-RNN Algorithm. *arXiv* **2019**, arXiv:1907.03329v1.

79. Klimek, J.; Klimek, J.; Kraskiewicz, W.; Topolewski, M. Long-Term Series Forecasting with Query Selector—Efficient Model of Sparse Attention. *arXiv* **2021**, arXiv:2107.08687v2.

80. Deshpande, P.; Sarawagi, S. Long Range Probabilistic Forecasting in Time-Series using High Order Statistics. *arXiv* **2021**, arXiv:2111.03394.

81. Yang, L.; Hong, S.; Zhang, L. Iterative Bilinear Temporal-Spectral Fusion for Unsupervised Representation Learning in Time Series. Available online: https://openreview.net/forum?id=MjbdO3_ihp (accessed on 25 February 2022).

82. Koochali, A.; Schichtel, P.; Dengel, A.; Ahmed, S. Probabilistic forecasting of sensory data with generative adversarial networks–forgan. *IEEE Access* **2019**, *7*, 63868–63880. [CrossRef]

83. Bondarenko, I. More layers! End-to-end regression and uncertainty on tabular data with deep learning. *arXiv* **2021**, arXiv:2112.03566.

84. Malinin, A.; Band, N.; Chesnokov, G.; Gal, Y.; Gales, M.J.F.; Noskov, A.; Ploskonosov, A.; Prokhorenkova, L.; Provilkov, I.; Raina, V.; et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv* **2021**, arXiv:2107.07455.

85. Choudhry, A.; Moon, B.; Patrikar, J.; Samaras, C.; Scherer, S. CVaR-based Flight Energy Risk Assessment for Multirotor UAVs using a Deep Energy Model. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 262–268.

86. Rodrigues, T.A.; Patrikar, J.; Choudhry, A.; Feldgoise, J.; Arcot, V.; Gahlaut, A.; Lau, S.; Moon, B.; Wagner, B.; Matthews, H.S.; et al. In-flight positional and energy use data set of a DJI Matrice 100 quadcopter for small package delivery. *Sci. Data* **2021**, *8*, 155. [CrossRef] [PubMed]

87. Patrikar, J.; Moon, B.; Oh, J.; Scherer, S. Predicting Like A Pilot: Dataset and Method to Predict Socially-Aware Aircraft Trajectories in Non-Towered Terminal Airspace. *arXiv* **2021**, arXiv:2109.15158.

88. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* **2018**, *34*, 802–808. [CrossRef]

89. Spiliotis, E.; Assimakopoulos, V.; Makridakis, S.; Assimakopoulos, V. The M5 Accuracy competition: Results, findings and conclusions. *Int. J. Forecast.* **2022**, *38*, 1346–1364.

90. Khodayar, M.; Wang, J. Spatio-Temporal Graph Deep Neural Network for Short-Term Wind Speed Forecasting. *IEEE Trans. Sustain. Energy* **2019**, *10*, 670–681. [CrossRef]

91. Peng, Z.; Peng, S.; Fu, L.; Lu, B.; Tang, J.; Wang, K.; Li, W. A novel deep learning ensemble model with data denoising for short-term wind speed forecasting. *Energy Convers. Manag.* **2020**, *207*, 112524. [CrossRef]

92. Gharghabi, S.; Imani, S.; Bagnall, A.; Darvishzadeh, A.; Keogh, E. Matrix profile xii: Mpdist: A novel time series distance measure to allow data mining in more challenging scenarios. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 965–970.

93. Madrid, F.; Imani, S.; Mercer, R.; Zimmerman, Z.; Shakibay, N.; Keogh, E. Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile. In Proceedings of the 2019 IEEE International Conference on Big Knowledge (ICBK), Beijing, China, 8–11 November 2019; pp. 175–182.

94. TS-Fresh. Available online: https://tsfresh.readthedocs.io/en/latest/ (accessed on 25 February 2022).

95. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the kdd, Portland, Oregon, USA, 2–4 August 1996; Volume 96, pp. 226–231.