Prediction Pipeline on Time Series Data Applied for Usage Prediction on Household Devices

1st Raluca Portase
Technical University of Cluj-Napoca
Cluj Napoca, Romania
email: raluca.portase@cs.utcluj.ro

2nd Ramona Tolas
Technical University of Cluj-Napoca
Cluj Napoca, Romania
email: ramona.tolas@cs.utcluj.ro

3rd Camelia Lemnaru
Technical University of Cluj-Napoca
Cluj Napoca, Romania
email: camelia.lemnaru@cs.utcluj.ro

4th Rodica Potolea
Technical University of Cluj-Napoca
Cluj Napoca, Romania
email: rodica.potolea@cs.utcluj.ro

***Abstract*—Processing time series is wildly used for many real-world applications such as financial market prediction, resource demand forecasting, device maintenance prediction, or environmental state prediction. In this work, we propose a general time series prediction pipeline with a hybrid unit for the relevance intervals on the processing part. The granularity unit is separated based on the intermittency level of the time series. We further apply the pipeline to real data from household appliances for non-intrusive usage pattern modeling and multistep-ahead prediction using machine learning methods.**

Keywords*—*Time series; data filtering; processing pipeline; home appliances data; forecasting devices usage.

I. Introduction

Time series prediction is the subject of multiple studies due to its general applicability to various domains. The existence of null, unrecorded, zero values in time series requires filtering data at different intervals, which still maintains the relevant recordings. The granularity level refers to these intervals of relevance. Changing the granularity could lead to a smaller number of non-relevant entries in the time series but would affect the overall sampling of the result. Depending on the problem, a smaller granularity might enhance the processing step.

Machine learning approaches have various applications in time series processing. One such application is the prediction of future values. Depending on the intermittency level of the dataset, classical regression models, neural networks, or specific ones that target data with multiple zeros are commonly used. Therefore, this work can be classified as an application of machine learning supervised models for knowledge and information extraction and processing.

A substantial percentage of water and energy resources used by a given household comes from household appliance usage [1] [2]. Therefore, extracting and understanding usage patterns would lead to a more accurate prediction of the resources needed in the future.

This work addresses a time series forecasting problem that uses intermittent time series and multistep ahead prediction. First, we propose a sampling rate separation based on the time series' intermittency level. Then, further, we integrate this in a general processing pipeline for prediction, which uses a combination of different sampling rates based on the number of zero entries from the time series. Finally, we apply this for knowledge extraction on real home appliance data from the industry to predict the following usage of given devices for a month. To the best of our knowledge, the previous pipelines used in the literature do not propose a separation of the sampling rate to obtain a better overall combined result. Our proposed pipeline offers a more rigorous approach from the perspective of the sampling rate.

Section 2 presents related work on time series sampling rate and prediction with a focus on forecasting using machine learning methods. Next, we present in Section 3 our proposed strategy for multistep prediction of time series with different intermittency levels. Then, in Section 4, we project the general model to household appliance data to predict the devices' future running time. Finally, we round up this paper in Section 5 with some conclusions and remarks.

II. Related work

A time series is a sequence of consecutive data points over time and is the most commonly used data type [3]. The sampling rate of a time series gives the maximum resolution of any prediction on that data. However, the best results are only sometimes given by using the smallest granularity of the data [4].

Intermittent time series refers to those series that have values equal to zero on multiple entries without obvious patterns of variation [5]. Prediction of their future values has been a subject of interest for numerous studies since long ago. Most of these studies are concerned with predicting intermittent and irregular sales demands [6] [7]. Non-intermittent data can become intermittent at fine-grained decomposition levels, for example, by using the time granularity of minutes or hours instead of days or months.

Univariate time series regression or forecasting is the simplest version and relies only on historical data of a variable to predict future behavior. On the other hand, multivariate analysis and prediction use the relationship between several

 ISBN: 978-1-68558-082-7

variables. Several studies suggest that models with multiple time series perform better than models with a single time series [8] [9].

Machine learning strategies are used to forecast and classify time series. One of the most common methods for multivariate forecasting is Vector Autoregression (VAR), but this has the disadvantage of not capturing non-linearity patterns. Numerous studies are using deep neural networks for prediction due to their capabilities in capturing non-linear interdependencies [10] [11] [12]. On the other hand, more straightforward methods that provide fast results, such as Support Vector Regression (SVR), have been successfully applied in time series forecasting due to their generalization capability in obtaining a unique solution [13] [14] [15].

The random forest regression model can also be used to predict multiple points in the future based on historical data by combining several single-point forecasting [16] [17]. Extreme gradient boosting is a decision tree ensemble learning algorithm similar to the random forest and can be used for classification and regression. Compared to the random forest, it uses a gradient of the data for each tree, which makes the calculation faster and more accurate. XGBoost [18] implementation for extreme gradient boosting method has also been successfully used for time series forecasting [19] [20].

The evaluation metrics are essential to any machine learning linear regression or forecasting problem. The most commonly used ones are Mean Square Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) [21]. For regression problems where the output might be zero, percentage error metrics, such as MAPE [22], are not suitable, instead Symmetric Mean Absolute Percentage Error (SMAPE) [23] can be used.

Time series forecasting applied to data from household appliances is a subject of multiple studies. This is done in the context of predicting the resources needed in the near future for a specific household [1] [24], as well as extracting and understanding usage patterns [25]. A specific time series data from home appliances are the running time information of the devices. This subject is particularly interesting in the cases of appliances with running cycles. Extracting runtime information in a non-intrusive manner from already existing data is tackled in [26].

Electrical energy consumption and peak demand forecasting are vital in planning and maintaining power systems. The appliances give part of the variation of the household consumers. Machine learning approaches have shown the best accuracy in forecasting electrical appliance consumption and are the current state-of-the-art solutions [27].

III. Strategy for dealing with intermittent data

This section covers two dimensions: the data sampling strategy at different granularities on data sets with different levels of intermittency and a general processing pipeline. This pipeline decomposes the processing part in two, based on the number of empty values in the input data. Our strategy proposes the usage of a model selector to decide the model used for the prediction and its corresponding granularity level. Due to the different granularity levels, the prediction result will have a hybrid time series unit.

Time series data from multiple sources can have zero values which could be caused by the nature of the data or the sampling rate used. When data does not offer sufficient initial information, projecting it onto a different subspace could lead to better results.

When applying multivariate forecasting or predicting values based on multiple time series, zero values negatively impact performance. Several strategies could be used to overcome this, such as handling missing data in forecasting or regression, cutting off data portions with multiple zero values, or reducing the overall sampling rate. Removing parts of time series data would lead to a loss of information regarding time dimension and misalignment. Simultaneously, setting the overall sampling rate to a higher value for all time series to overcome the prediction issues of the ones with multiple zeros would impose a prediction with a higher granularity regardless of the level of intermittency of the data series. More than that, it would reduce the dimension of the data set, which might lead to insufficient data in some cases.

To maintain the granularity as small as possible where it does not affect the identification of the objectives in hand and to have proper outcomes over the entire dataset, we propose a hybrid sampling rate based on the time series intermittency level as follows: time series with the number of zero values on initial sampling rate smaller than a given threshold - granularity level set to time series unit. The granularity should be composed of several time units for the other time series. For example, for time series data with a granularity level of a second, if the data has a high level of intermittency, a minute can be used as a time unit in the prediction pipeline. Using a hybrid sampling rate could partially overcome the disadvantages that arise from the sampling rate for portions of the dataset.

The general pipeline is comprised of three main steps: pre-processing, processing, and post-processing. Our method introduces new steps in the processing part.

We propose a multistep-ahead time series prediction pipeline that divides the problem into two parts. A regression problem in the first part predicts the usage of time series with a smaller intermittency level and outputs the prediction for a given period by using the initial time series unit of time. The second model gives the prediction with a higher granularity level for the sampling rate for time series with a smaller number of non-zeros per day. This method is illustrated in Figure 1. Depending on a threshold, the newly added selector component will choose the appropriate model and granularity for the time series. This way, we would achieve the best prediction results in the most suitable unit given a time series used as input in the pipeline.

The threshold used for the decision can vary depending on the nature of the data and the problem at hand. Depending on the initial time series unit, the higher level granularity should be chosen based on the problem to be solved while maintaining

 ISBN: 978-1-68558-082-7

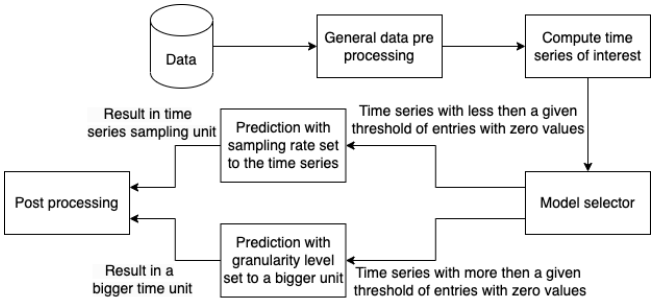

Fig. 1. Proposed pipeline for prediction based on time series intermittency level

a logical predefined time scale such as second, minute, hour, day, week, month, and so on. The post-processing part, as well as the pre-processing, depends on the problem to be solved and the data and is not the subject of this work.

A possible application of this pipeline could be estimating resources - such as the energy or water consumption in our examples. Having an accurate prediction of the future needed resources could lead to better management of the resources.

IV. Method Instantiation on Home Appliances Data

The strategies presented in Section III are applicable in the general context of time series prediction with intermittent demand or missing data. We further applied them in the context of home appliances and present the results in this section. We evaluated three machine learning strategies for prediction: decision trees, extreme gradient boosting, and support vector regression. We particularised the general time series pipeline presented for home appliances running time data based on the results.

A. Context and dataset description

Several types of household appliances have functioning cycles, such as washing machines, tumble dryers, dishwashers, ovens, or microwaves. Given the data's nature, reducing the intermittency in forecasting the sampling period is crucial.

The best-suited granularity from the perspective of the possibility of making an accurate prediction would be of one day because a smaller one would lead to a massive number of samples with values equal to zero. As a consequence, the prediction would be less accurate. According to [25], appliances tend to be used based on a general pattern on the temporal dimension. Therefore, undersampling the devices not so used by cutting off extensive intervals with no usage would lead to a loss of information that arises from the time dimension.

When projecting the appliance usage forecasting into the energy consumption estimation, having a daily prediction could lead to a better estimation of the resources per day. A smaller granularity could be used for a more detailed analysis of the variation of the energy needed for a day.

The methods presented in Section III for data sampling and the pipeline for the prediction can be applied to any type of data. We projected them in our experiments on real operational data from household appliances. More specifically, we used data logs from washing machines recorded over one year. Due to copyright reasons, we will further maintain the data's anonymity. We used a time series unit of one day of usage, pre-processed the data, and computed each device's run time in seconds per day. The result is a time series where one point represents each device's runtime per that day of the year. The proportion of zero values is computed taking into account the entire interval of data samples. Since we are interested in predicting the duration a device would be used during a time interval of a given granularity, we have chosen seconds as the unit of measurement for appliance usage.

Figure 2 presents the usage patterns of appliances investigated - the histogram with the number of days an appliance was used computed for all devices from the initial data set. As can be seen, most of the devices are used for a few days. Thus, we removed them from the investigation.

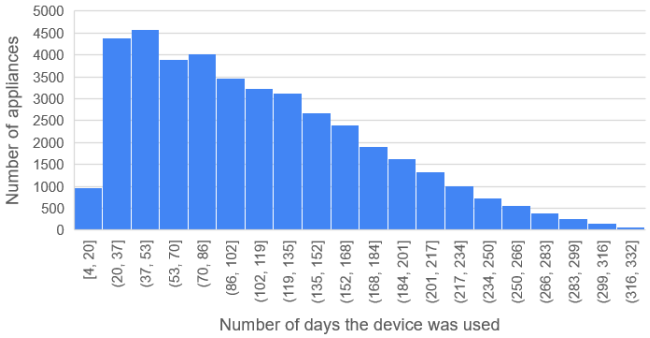

Fig. 2. Histogram of the number of appliances and their numbers of days with running cycles over a year

We removed from the initial data set the devices that have more the 50% of days with no usage since their lack of usage adds a question mark regarding the correctness of their utilization and if it is the usual one. Furthermore, the remaining appliances were sufficient to create a meaningful sample. We separated the remaining entries into two groups based on the average usage of the machines. From the initial dataset of tens of thousands of devices, we obtained a dataset formed of 1.2k devices used at least 70% of the days and a total of 6.3k devices operated at least 50% of the days. The dimensions of the initial dataset compared to the one after selecting instances are presented in Table I.

TABLE I
Dimensions of the initial dataset versus selected instances over one year

Dataset	Size	No of devices	Time period
Raw data	8.1mil	49k	1 year
Selected instances	2.4mil	6.3k	1 year

For our evaluations, we have selected several different appliances. Among the investigated ones, we present the results for six appliances representative of their categories out of the

6.3k selected devices. Their characteristics are summarised in Table II, which we refer to further. All of these appliances have a different number of days without usage per year and different average runtime per day.

TABLE II
NUMERICAL CHARACTERISTICS OF APPLIANCES USED FOR EXPERIMENTS

Appliance	Average usage/ day (s)	No of days without usage/ year
App 1	36.718	39
App 2	7.233	46
App 3	6.976	67
App 4	7.848	120
App 5	5.926	141
App 6	6.483	156

The first three appliances have a lower level of intermittency, while the last 3 have a higher number of zero values. Moreover, they all have a different average usage number of seconds per day. We will further refer to these appliances in the experiments from the next section.

B. Daily prediction of future appliance usage

We designed and implemented several preliminary experiments on our dataset to reduce the search area. From the available tools commonly used for prediction, we selected Random Forest, Support Vector Regression, as well as XGBoost [18] implementation for gradient-boosted trees. We split the dataset and used it for training data from 11 months, while for evaluation, 1-month data.

The purpose of the first set of experiments is included in the multiple-step-ahead prediction category. More specifically, to predict the appliances' daily usage for a month's time window. For each strategy, we investigated the best suited parameters for our dataset. As a result, we identified 125 trees for the random forest as the best configuration, 100 trees for XGBoost, and a linear kernel for SVR.

We made several preliminary investigations to identify the number of zeros from raw data by using levels 10%, 15%, 20%, 30%, 40%, and 50%. We filtered the data and ran several experiments where we varied the dataset used for the model based on the percentage of zeros from the time series. The comparison of the mean average error and symmetric mean absolute percentage error obtained on several appliances using random forest on the three most significant levels from our dataset is presented in Table III. Table II summarizes the appliance characteristics from this experiment. From there, we selected the first three devices due to their low level of intermittency.

In the first experiment, we only used for training the appliances that have less than 15% of days with no usage. Then we added the devices that had up to 30% of days zero runtime seconds. Finally, we added appliances in training set up until half of the entry points were zeros and evaluated the model's performance.

The best results without modifying the granularity of the prediction were obtained for daily prediction of devices for models based on learning data with up to 30% of the days

TABLE III
MEAN AVERAGE ERROR AND SYMMETRIC MEAN ABSOLUTE PERCENTAGE ERROR FOR DAILY PREDICTION OF APPLIANCES BASED ON THE PERCENTAGE OF ZEROS ENTRIES APPLIANCES IN THE TRAINING SET

	15% zeros		30% zeros		50% zeros	
Appliance	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
App 1	8365	10.92	**7871**	**10.28**	7886	10.32
App 2	3345	25.07	**2913**	**22.69**	3134	23.95
App 3	**4292**	27.25	4374	**27.23**	4494	28.01

of a year. However, the prediction was less accurate when we used the appliances with a higher number of zeros for the model. Therefore, further on, we are using the 30% threshold for the daily prediction.

We implemented and compared the results for daily prediction by using random forest, XGBoost, and SVR on the best size for the dataset for training previously found. The results are presented in Table IV. For measuring the prediction, we have used mean average error in seconds and symmetric mean absolute percentage error normalized on [0,100] interval.

TABLE IV
RESULTS OF DAILY PREDICTION OF APPLIANCES USAGE AFTER USING THREE DIFFERENT METHODS

	Classifier					
	Random forest		XGBoost		SVR	
Metrics	MAE	SMAPE	MAE	SMAPE	MAE	SMAPE
App 1	**7871**	**10.28%**	8771	11.38%	9148	12.44%
App 2	**2913**	**22.69%**	3573	26.22%	3810	27.39%
App 3	4374	27.23%	**4321**	**26.61%**	4613	41.81%

In our experiments, XGBoost and Random Forest had similar results, while SVR performed worse for daily usage prediction regardless of the set size.

C. Impact of variation of the granularity level

Generally, predicting time series with a more significant intermittency level using classical forecasting methods does not perform well. For these, several other methods could be used. These scaled on our experiments too, where the average SMAPE was over 50% for daily prediction of devices with more the 30% of the data having values equal to zero when using random forest, SVR, and XGBoost.

According to our previous experiments from Table III, in the case of appliance data, using an upper threshold of 30% for the number of zero values where the daily usage can be predicted would be appropriate. Further, we propose the usage of the next logical time unit as a granularity level. This gives us the time unit of a week instead of a day for devices with a more significant number of missing data or zero values.

We used a granularity level of a week and recomputed the time series for the devices with a higher percentage of zero data. Then, we applied the machine learning strategies from above and recorded the results in Table V. The mean average error represents the number of seconds per week. Although there was no general winner as the best tool for all the appliances, XGBoost and SVR performed well on a subset of devices.

 ISBN: 978-1-68558-082-7

TABLE V
RESULTS OF WEEKLY PREDICTION OF APPLIANCES USAGE AFTER USING THREE DIFFERENT METHODS

	Classifier					
	Random forest		XGBoost		SVR	
Metrics	MAE	SMAPE	MAE	SMAP	MAE	SMAPE
App 4	11339	15.52 %	13760	17.98%	**9148**	**13.89%**
App 5	29483	30.64%	**22435**	**22.65%**	28623	30.70%
App 6	16571	19.65%	**15637**	**17.93%**	16207	19.79%

The initial SMAPE values obtained when using daily prediction on appliances 4-6 were over 50%. However, by changing the granularity of the time series to a week, on all of our experiments, symmetric mean absolute percentage error became under 25%, which means that SMAPE was reduced by at least 50% for devices with a higher number of zeros.

D. Processing pipeline particularized on home appliances data

According to the results for time series data from appliances with running cycles, a daily sampling rate performs well for highly used devices. In our experiments, the machines used at least 70% of the days are part of this category. In the case of the other appliances, using a granularity level of a week gives good results while maintaining a logical time unit, making the results valuable and keeping the dataset size to a reasonable amount.

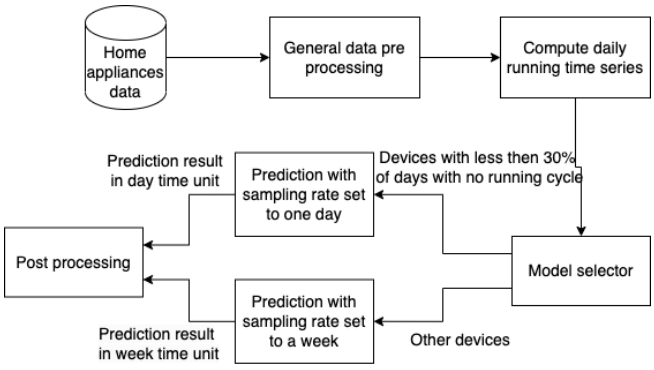

Fig. 3. Proposed pipeline instantiated on home appliances data

Figure 3 presents the corresponding instantiation in the context of home appliances data of the pipeline presented in Section III. The purpose of this pipeline is the non-intrusive usage pattern prediction with a hybrid granularity level.

V. CONCLUSION AND FUTURE WORK

The major contributions of this work consist in the granularity level separation on time series data and the general processing pipeline for time series having different levels of intermittency. We applied these strategies to running time information from real household appliance data. Further on, we instantiated the general pipeline for this particular use case based on our threshold determination experiments to predict the future running cycles of a given appliance in the next month using machine learning. Applying the presented strategies to other types of time series is the subject of future work.

REFERENCES

[1] S. Koop, A. Van Dorssen, and S. Brouwer, "Enhancing domestic water conservation behaviour: A review of empirical studies on influencing tactics," vol. 247. Elsevier, 2019, pp. 867–876.

[2] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," vol. 10, no. 1. IEEE, 2017, pp. 841–851.

[3] J. Lin, S. Williamson, K. Borne, and D. DeBarr, "Pattern recognition in time series," vol. 1, no. 617-645. Citeseer, 2012, p. 3.

[4] R. J. Frank, N. Davey, and S. P. Hunt, "Time series prediction and neural networks," vol. 31. Springer, 2001, pp. 91–103.

[5] J. D. Croston, "Forecasting and stock control for intermittent demands," vol. 23, no. 3. Springer, 1972, pp. 289–303.

[6] M. W. Seeger, D. Salinas, and V. Flunkert, "Bayesian intermittent demand forecasting for large inventories," vol. 29, 2016.

[7] X. Zhuang, Y. Yu, and A. Chen, "A combined forecasting method for intermittent demand using the automotive aftermarket data." Elsevier, 2022.

[8] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," vol. 36, no. 3. Elsevier, 2020, pp. 1181–1191.

[9] K. Bandara, C. Bergmeir, and S. Smyl, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," vol. 140. Elsevier, 2020, p. 112896.

[10] S.-Y. Shih, F.-K. Sun, and H.-y. Lee, "Temporal pattern attention for multivariate time series forecasting," vol. 108. Springer, 2019, pp. 1421–1441.

[11] M. Sousa, A. M. Tomé, and J. Moreira, "Long-term forecasting of hourly retail customer flow on intermittent time series with multiple seasonality," vol. 5, no. 3, 2022, pp. 137–148. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666764922000273

[12] Y. Jeon and S. Seong, "Robust recurrent network model for intermittent time-series forecasting," vol. 38, no. 4, 2022, pp. 1415–1425, special Issue: M5 competition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021001151

[13] K. Lau and Q. Wu, "Local prediction of non-linear time series using support vector regression," vol. 41, no. 5. Elsevier, 2008, pp. 1539–1547.

[14] C.-J. Lu, T.-S. Lee, and C.-C. Chiu, "Financial time series forecasting using independent component analysis and support vector regression," vol. 47, no. 2. Elsevier, 2009, pp. 115–125.

[15] N. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: a survey," vol. 4, no. 2. IEEE, 2009, pp. 24–38.

[16] A. González-Vidal, V. Moreno-Cano, F. Terroso-Sáenz, and A. F. Skarmeta, "Towards energy efficiency smart buildings models based on intelligent data analytics," vol. 83. Elsevier, 2016, pp. 994–999.

[17] E. Mussumeci and F. C. Coelho, "Large-scale multivariate forecasting models for dengue-lstm versus random forest regression," vol. 35. Elsevier, 2020, p. 100372.

[18] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen *et al.*, "Xgboost: extreme gradient boosting," vol. 1, no. 4, 2015, pp. 1–4.

[19] N. Zhai, P. Yao, and X. Zhou, "Multivariate time series forecast in industrial process based on xgboost and gru," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9. IEEE, 2020, pp. 1397–1400.

[20] R. A. Abbasi, N. Javaid, M. N. J. Ghuman, Z. A. Khan, and S. Ur Rehman, "Short term load forecasting using xgboost," in *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33*. Springer, 2019, pp. 1120–1131.

[21] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," 2018.

[22] A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," vol. 192. Elsevier, 2016, pp. 38–48.

 ISBN: 978-1-68558-082-7

[23] S. Makridakis, “Accuracy measures: theoretical and practical concerns,” vol. 9, no. 4. Elsevier, 1993, pp. 527–529.

[24] M. Razghandi, H. Zhou, M. Erol-Kantarci, and D. Turgut, “Short-term load forecasting for smart home appliances with sequence to sequence learning,” in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[25] C. Firte, L. Iamnitchi, R. Portase, R. Tolas, R. Potolea, M. Dinsoreanu, and C. Lemnaru, “Knowledge inference from home appliances data,” in *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2022, pp. 237–243.

[26] E. M. Olariu, R. Tolas, R. Portase, M. Dinsoreanu, and R. Potolea, “Modern approaches to preprocessing industrial data,” in *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2020, pp. 221–226.

[27] E. U. Haq, X. Lyu, Y. Jia, M. Hua, and F. Ahmad, “Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach,” vol. 6, 2020, pp. 1099–1105, 2020 The 7th International Conference on Power and Energy Systems Engineering. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484720314967

 ISBN: 978-1-68558-082-7