Electricity load forecasting at single household level using disaggregated data

Katarzyna Maria Zawadzka

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisor: Prof. Carlos Augusto Santos Silva

Dr. Henrique Ramalho Monteiro Latourrette Pombeiro

Examination Committee

Chairperson: Prof. Susana Isabel Carvalho Relvas

Supervisor: Prof. Carlos Augusto Santos Silva

Member of Committee: Prof. Paulo José da Costa Branco

December 2019

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the *Universidade de Lisboa*.

Abstract

The main purpose of the following research is to further contribute to the study of demand forecasts at the individual household level. In order to do this, nine models were built to predict one-hour-ahead energy consumption. Three different scenarios were tested in terms of distribution and amount of input data: forecast model based on aggregated data, forecast model based on disaggregated data collected during direct metering (with one-hour resolution) and forecast model based on data disaggregated by Non-Intrusive Load Monitoring (NILM) algorithm. In addition, three different machine learning algorithms were used, namely: random forest, multi-layer perceptron and long short term memory. The models were built for one house and additionally tested for four others to check their scope of application.

The results obtained in the analysis showed that models using NILM data provided similar results as models using data with direct measurements of energy consumption by individual devices. Moreover, in almost every case considered data containing only aggregated data provided the most accurate results to perform aggregated consumption forecast obtaining Mean Absolute Percentage Error (MAPE) at the level of 27.6% in the best case scenario. Of the three algorithms tested, random forest provided the best results, and was also the fastest and easiest to implement.

Keywords: load forecasting, disaggregated data, NILM, machine learning, individual household

Resumo

O principal objetivo desta dissertação é contribuir para o estudo das previsões de consumo de energia eléctrica no sector doméstico. Para isso, foram desenvolvidos nove modelos de previsãodo consumo de energia para o horizonte temporal de uma hora. Três cenários diferentes foram testados em termos de distribuição e quantidade de dados de entrada: modelo de previsão baseado em dados agregados; modelo de previsão com base em dados desagregados obtidos atravésde medição direta (com resolução de uma hora) e modelo de previsão com base em dados desagregados por um algoritmo de desagregação não intrusivo (NILM). Par além disso, três algoritmos diferentes de aprendizado de máquina foram utilizados: floresta aleatória, perceptrãode múltiplas camadas e memória de longo prazo. Os modelos foram construídos para uma habitaçãoe adicionalmente testados para outras quatro para verificar seu âmbito de aplicação.

Os resultados obtidos na análise mostraram que os modelos que usam dados NILM forneceram resultados semelhantes aos modelos que usam dados com medições diretas do consumo de energia por dispositivos individuais. Além disso, em quase todos os casos, os dados considerados contendo apenas dados agregados forneceram os resultados mais precisos de previsão agregada de consumo, obtendo um erro percentual absoluto médio (MAPE) de 27,6% para omelhor cenário. Dos três algoritmos testados, a Floresta Aleatória (Random Forest) obteveos melhores resultados, sendo também aquele que é mais rápido e fácil de implementar.

Palavras-chave: previsão de consumo, dados desagregados, NILM, aprendizado de máquina, consumodoméstico de electricidade

Table of Contents

List of Figures

List of Tables

List of Acronyms

AMI - Advanced Metering Infrastructure

ANN - Artificial Neural Network

ARIMA - Auto Regressive Integrated Moving Average

BPNN - Backward Propagation Neural Network

BR - Bayesian Regularization

CBAF - Cluster-Based Aggregate Forecasting

CNN - Convolutional Neural Networks

CV - Coefficient of Variation

EDA - Exploratory Data Analysis

ELM - Extreme Learning Machine

ETC - Extra Trees Classifier

FCRBM - Factored Restricted Boltzmann Machines

FFNN - Feed-Forward Artificial Neural Network

FITNET - Function Fitting Neural Networks

FN - False Negative

FP - False Positive

GMM - Gaussian Mixture Models

GPR - Gaussian Process Regression

GRBFNN - Generalized Radial Basis Function Neural Network

IQR - Interquartile Range

LM - Levenberg-Marquardt Back-Propagation Algorithm

LR - Linear Regression

LSTM - Long Short Term Memory

LTU - Linear Threshold Unit

MAPE - Mean Absolute Percentage Error

MDS - Mean Squared Deviation

MLP - Multi- Layer Perceptron

MLR - Multiple Linear Regression

MSE - Mean Squared Error

NARX - Non-Linear Autoregressive With Exogenous

NILM - Non-Intrusive Load Monitoring

NMAE - Normalized Mean Absolute Error

NRMSE - Normalized Root Mean Square Error

PCA - Principal Components Analysis

R^2 - Coefficient Of Determination

RF - Random Forest

RFE - Recursive Feature Elimination

r-MAPE - resistant MAPE

RMSE - Root Mean Squared Error

RNN - Recurrent Neural Network

S2S - Sequence to Sequence

SAE - Stacked Autoencoders

SARIMA - Seasonal Auto Regressive Integrated Moving Average

SCG - Scaled Conjugate Gradient

SPLF - Similar Profiles Load Forecast

SVM - Support Vector Machines

SVR - Support Vector Regression

TN - True Negative

TP - True Positive

W&D - Washing and Dishwasher

WH - Water Heating

1. Introduction

1.1. Problem definition

According to the data provided by European Commission[1], buildings are one of the largest energy consumers in Europe. In 2017 [2], the residential sector contributed to consumption of 27.2% of final energy in the European Union, which made it the second most energy-intensive sector, right after transport (34.2%). Moreover, numerous factors indicate the rising trend in energy consumption, such as growth of number of electrical appliances in homes, increased demand of comfort as well as overall rise in the number of households [3][4]. The above presented numbers and trends suggest that there is a vast potential for energy efficiency gains in the residential sector.

To achieve the ambitious goal of reaching net-zero carbon economy in Europe by 2050, the European Union has set specific targets to be reached by the Member States by 2030 [1]: at least 40% cuts in greenhouse gas emissions (from 1990 levels), at least 32% share for renewable energy and at least 32.5% improvement in energy efficiency. To achieve this and still be able to retain secure, competitive and sustainable energy systems, significant changes in both energy production as energy consumption sectors must be implemented.

In addition, the use of renewable sources such as solar or wind power is characterized by high volatility in the production of electricity, therefore, one of the biggest challenges to increase the integration of energy coming from renewable sources into transmission networks is to accurately control the system. The installation of reserve capacity, investments in electricity storage units such as batteries, or utilising future electric vehicles as storage are a few of several proposals to deal with negative effects of variable electricity production [5]. To effectively balance real-time demand and supply, and avoid the congestion of the grid, the development of methodologies that allow as accurate as possible to forecast production as well as consumption of the electricity may prove to be crucial.

Moreover, there has been a significant drop of prices of solar panels and battery-based storage systems. Together with a variety of incentives offered by the governments, this created a great demand and interest in installing these systems in the residential sector resulting in an increased number of residential energy self-producers, prosumers and local energy communities. All of these factors, connected with the general trend of switching from energy sales market to services based market, created a new business sector, where the production and sale of energy between a typical producer - consumer model is not the dominant model anymore. The prosumer model has disturbed the current market and opened the way for a new system where energy management will become one of the most important elements of the chain.

Further control over electricity supply and demand can be achieved by continuous investments and improvements in technologies such as Advanced Metering Infrastructure (AMI) [6]. In recent years, projects related to smart networks in Europe have grown significantly - currently, the database contains more than 950 R&D and demonstration projects, which totalled over 5 billion euro investment. Smart meters deployment is already completed in Italy, Finland and Sweden, and is planned or underway by Austria, Denmark, Estonia, France, Greece, Ireland, Latvia, Luxemburg, Malta, Netherlands, Poland, Spain, Romania, Portugal and UK[7]. They can be used as one of the instruments that allow us to get closer to the goals set by the European Union.

With the smart meter deployment, the sector dealing with Non-Intrusive Load Monitoring (NILM) has significantly developed over the last years, as it allows to extract the energy consumption profiles of individual devices from the total energy profile obtained through smart meters.

All this indicates an increasing demand for accurate control of electricity production and consumption, and this requires accurate forecasting of these two dimensions. Smart sensors and increasingly better machine learning algorithms create ideal conditions for the development of this field. And indeed, from year to year, more and more articles on forecasting electricity consumption can be found. And although the majority of them are still focusing on aggregated data on the system demand level, the topic of forecasts with the use of disaggregated data at the level of individual houses is discussed more and more frequently.

1.2. Research goal

The main purpose of the following research is to further contribute to the study of demand forecasts at the household level, striving to improve forecasting by adding consumption data to the models at a disaggregated level, and thus checking whether direct measurements of energy consumption by individual devices or NILM results improve the forecasting of total energy consumption in individual households. In order to obtain the most accurate forecast model, different machine learning methods, several configurations of those and various data representation will be implemented and compared. For that purpose, five different houses will be compared in three following scenarios:

I. Power consumption forecast model based on disaggregated data collected during direct metering.

II. Power consumption forecast model based on data disaggregated by NILM algorithm provided by Watt-IS company.

III. Power consumption forecast model based on aggregated data.

All models will be used to forecast household demand one hour in advance.

Furthermore, answers to the following questions will be sought:

Is the same model sufficient to forecast power consumption of different houses?

A model will be built and optimized specially for one single household. After that, the same model's performance will be tested on the other households with different power consumption patterns. The level of forecast accuracy will be compared.

Is there an advantage in using disaggregated data over aggregated data?

Models built with disaggregated data, thanks to capacity to process more detailed patterns, are expected to perform better than models based on aggregated data. However, this group of models requires more complex collection of the data and results in higher complexity of the model and increased computation time. It will be discussed whether the possible increase in accuracy rewards the additional model costs mentioned above.

Is NILM algorithm necessary to significantly improve the forecast?

Despite the fact that models based on aggregated and NILM data require the same data (total consumption), the latter approach results in higher complexity because of the additional algorithm used for data disaggregation. Again, the level of forecast accuracy of two models will be compared.

1.3. Structure of the document

The presented paper is structured as follow: in the first chapter, an overview of the main power forecasting drivers are explained. This section also presents the main goal of the research as well as the related questions presenting the raised issues in more detail.

The second chapter describes selected scientific articles that focus on the topic of power consumption forecasting. Their analysis will serve as the main driver to choose individual methods and techniques to create the target model.

In the third chapter, the individual steps of model creation and applied methodologies are presented. In the first two parts, the emphasis is placed on the data itself, and in the remaining four on the characteristics and structure of the model.

The fourth chapter presents the results obtained for five houses in three previously defined scenarios - models built on aggregated, disaggregated and NILM data. Identification of the best case is carried out by comparing the selected validation values.

In the fifth chapter all the activities and results obtained in the preceding parts of the work are summarized. Any conclusions resulting from the analysis are described in the final, sixth chapter.

2. Literature review

Despite the growing interest in forecasting power consumption, still relatively few scientific papers focus on models based on disaggregated data. For the overall view, various methodologies and scenarios have been analyzed before proceeding to the target model development. From all the literature, a set of 14 articles have been analysed in detail and present in this chapter. Table 1 summarizes the main characteristics of the papers: the type of model that is used, the time step, the number of samples, the share of training and validation and test data, the features, the type of building, the summary of evaluation indicators, the prediction horizon, the use of disaggregated data or not and finally, the reference indication.

Table 1. Summary of selected scientific articles describing the prediction of electric consumption.

No.	Model	Time step	No. Of samples	Training Data	Validataion/ & Testing Data	Data features	Type of building	Evaluation	Forecast horizon	Disaggregat ion	Ref
1	FFNN SVR LS - SVM GPR GMM Hybrid	5 min	N/A	75%	25%	historical consumption, weather data geometry of buildings	residential	MAPE 8-16 % (1h) 10-19 % (24h)	1h 24h	no	[8]
2	ANN	1h	N/A	75%	25%	historical consumption, weather data, no of people	tertiary	MAPE 4,51-14,81%	1h	no	[9]
3	SVR LR MLP SARIMA	1 h	12 430	67%	33%	historical load, contectual features (day of week, hour of day, weather info)	residential	NRMSE 0,56-0.58% (1h) 0,61-0.67% (24h)	1h 24h	no	[10]
4	LSTM S2S	1 min 1 h	2 060 600 (1 min)	75%	25%	historical consumption	residential	RMSE 0,625 (1h) 0,667 (24h)	1h 24h	no	[11]
5	Proposed NILM ANN ARIMA SPLF	1 min (*)	80 640	52%	58%	historical consumption	residential	MAPE 4,7-10.1%	3 h	yes	[12]
6	LSTM	30 min	17 520	70%	20%	historical consumption daily temperature time of day	residential	MAPE 3,15% (3d) 3,76% (15d)	3 days 15 days	no	[13]
7	ANN	15 min	44 640	70%	15%	historical consumption timestamp	residential	MSE 0,0049 - 0,0056	10 days	yes	[14]
8	ANN	1 year	49	80%	20%	GVA exergy	N/A	MSE 0,000165	10 years	yes	[15]
9	CNN	1 min 1 h	207 5259 (1 min)	75%	25%	historical consumption calendar	residential	RMSE 0,677	60 h	no	[16]
10	ELM	15 min	34 939	N/A	N/A	historical consumption	retail	RMSE 22,9 (30m) 59,18 (1h)	30 min1h	no	[17]
11	ANN	30 min	250	60%	40%	historical consumption air temp realtive humidity solar radiation	institutional	R^2 0,96-0,98	20 days	no	[18]
12	ANN SVR	15 min	190 132	80%	20%	historical consumption weather calendar class schedule	institutional	MAPE 3,5 -10%	24 h	no	[19]
13	FFNN	15 min	93 744	85%	15%	historical consumption apartment area no of occupants	residential	MAPE 10-23% (1h) 4,2% (24h)	1h 24h	yes	[20]
14	RF ARIMA ANN SVR	1 h	8 930	89%	11%	historical consumption behavioral data	individual household	MAPE 2 6.8-56.3%	24 h	yes	[6]

2.1. Residential energy forecast using aggregated data

Dong, Li, Rahman and Vega [8] integrated data-driven and engineering (using laws of thermodynamics) models and built an hybrid model, where they test five different machine learning algorithms (Artificial Neural Network, Support Vector Regression, Least-Square Support Vector Machine, Gaussian Process Regression, Gaussian Mixture Model) to forecast 1 hour and 24 hours ahead. They split the data and run two separate forecasts for AC (air-conditioning) power and non-AC (plug load, lighting, water heater etc.) power consumption simultaneously. The final result was obtained by summing up the two individual predictions. Depending of the tested household, they were able to obtain a mean absolute percentage error around 8.16 % for the hybrid model. In their work, they emphasize the fact that residential buildings, because of the dynamic change of power usage patterns due to stochastic occupancy behaviour, requires better optimized models than commercial buildings with quite regular consumption patterns.

Benedetti, Cesarotti, Introna and Serranti [9] created a model to allow the automatic utilization of a forecasting tool. They propose a method to identify the lack of accuracy in the model and compare two different retraining methods. Moreover, a minimum period of data collection to obtain reliable results and the maximum period of usability of the model are calculated. They investigated three different model structures that differed from each other in the number of hidden layers and neurons. Their model, trained on data obtained from two months of measurements, was able to accurately predict the consumption for 60 days after the training period.

Wijaya, Vasirani, Humeau and Aberer [10] focused on short-term electricity demand forecasting for residential customers at the individual and aggregate level. To create a model for over 782 households, they apply the Cluster-based Aggregate Forecasting (CBAF) strategy, which consists in the initial grouping of similar households, then on forecasting the power consumption of individual clusters separately, and finally on the aggregating separate predictions in the overall result. In addition, they propose the Correlation-based Feature Selection, which allows the automated selection of the most important features for each household.

Stefan and Patrick Hosein [11] investigate two deep machine learning methodologies using standard LSTM algorithm and LSTM based S2S architecture for electricity consumption for a single residential customer. The research was carried out in two different scenarios. While the second method gave satisfactory results in both cases, the standard LSTM failed in the one-minute forecast and was only useful for the hourly forecast.

Al Khafaf, Jalili and Sokolowski [13] propose a model based on LSTM algorithm and use it to forecast electricity consumption of 609 households in Australia, which were grouped into four clusters. They analyze three scenarios that differ in the number of features and the forecasting horizon. In this way they showed that the LSTM algorithm brings satisfactory results in both short and medium term forecasts, and that the number of appropriately selected features significantly affects the accuracy of the model.

Amarasinghe, Marino and Manic [16] analyse the effectiveness of Convolutional Neural Networks (CNN) and compare it against results obtained by LSTM S2S, FCRBM, ANN, SVM. The analysis carried out for one-hour resolution data from an individual household showed that the proposed technique was comparable with other deep learning techniques, but clearly outweighed the forecast obtained under the SVR method.

Li et al. [17] propose a model in which two elements are combined: stacked autoencoders (unsupervised neural network) and the extreme learning machine. The first one is used to extract power consumption features of the building and the other for accurate prediction of power consumption. The results were compared with four different models - BPNN, SVR, GRBFNN and MLR. Despite the fact that the GRBFNN method provided a more accurate forecast than extreme SAE, it was noted that the over-fitting phenomenon in the training process occurred, which is extremely undesirable.

2.2. Residential energy forecast using disaggregated data

Gajowniczek and Ząbkowski [6] focus on 24 hour ahead electricity consumption forecast for individual household. In their paper they confirm that forecast for a single unit generates a larger error than forecasting for the cluster of houses. To increase the accuracy of their model, they create matrix with the probabilities of "appliance turn ON" events in each hour.

Dinesh, Makonin and Bajic [12] first decompose the aggregate power signal into individual appliance signals. Those signals are used to make a forecast of the power consumption of each device separately. Finally, the overall forecast is created by summing up individual projections. Their original model is based on identifying ON and OFF states of the appliances - using graph construction, spectral representation and clustering, they predict the ON-set of devices at timet. The proposed model was compared with four other methodologies (NILM, ANN, ARIMA, SPLF) and in each of the considered scenarios it obtained the best results.

Oprea, Pîrjan et al. [14] used sensor recorded and disaggregated data to forecast electricity consumption of 8 residential houses with renewable electricity generation system and smart metering system. With mixed ANN model using both non-linear autoregressive with exogenous input (NARX) ANNs and function fitting neural networks (FITNETs) they test three different training algorithms (LM, BR, SCG) and apply gap-filling technique based on the interpolation. In their paper two case studies were investigated: the first one for the total electricity consumption from the grid and the second one for the total electricity consumption of all the individual appliances. Both approaches (NARX and

FITNETs) offered a very satisfying predictions with correlation coefficient ranging between 0.961 and 0.998 depending on the scenario and training algorithm.

Sereno and Alves [15] used data on exergy consumption (disaggregated on different types of exergy) collected over nearly 50 years in Portugal to build an ANN model applying the LM algorithm. In data pre-processing, all the values were normalized to obtain null average and unitary standard deviation. After that, for optimal features selection, correlations analysis and principal components analysis were implemented. The accuracy of the model has been compared with models based on multiple and multivariate linear regression and a smaller forecast error was recorded while using the ANN method.

Rodrigues, Cardeira and Calado [20] focused on 93 households in Lisbon, Portugal. Their main objective was to prove that it is possible to obtain satisfactory results without weather data, which is usually used in load forecast models. Using FFNN model and data such as historical consumption, apartment area and the number of occupants, they were able to obtain a mean absolute percentage error at the level of 10 % for daily consumption forecast.

2.3. Application to other types of buildings

Moon, Park, Hwang and Sanghoon [19] use power consumption data of 4 university buildings collected over more than 1 year to build two models using two different methods - Feed-Forward Neural Network (FFNN) and Support Vector Regression (SVR). Using standardized data, adjusted temperature (by annual temperature) and continuous representation of time, they manage to obtain error ranging from 3.46 to 10% in an hour ahead forecast.

Deb, Eang, Yang and Santamouris [18] present a model build for three institutional buildings. To avoid the problem of high variation, they divided the data into five power classes and class numbers, which allows to reduce the level of information in the power consumption values. This approach resulted in significant improvement in the accuracy of the model.

2.4. Conclusions

The analysis of the above papers confirmed the variety of approaches to the issue of energy consumption prediction. At the same time, different similarities can be seen, such as the use of individual algorithms, the choice of the ratio of training to testing data sets, or the horizon of prediction. Therefore, it was decided to continue building a model that would allow forecast of one hour forward energy consumption and to test three algorithms, namely: random forest, multi-layer perceptron and long short term memory.

3. Methodology

In this chapter the methodology adopted in this work is presented, which closely follows a standard machine learning project workflow [21].The workflow with its various sub-steps is presented below (Figure 1):

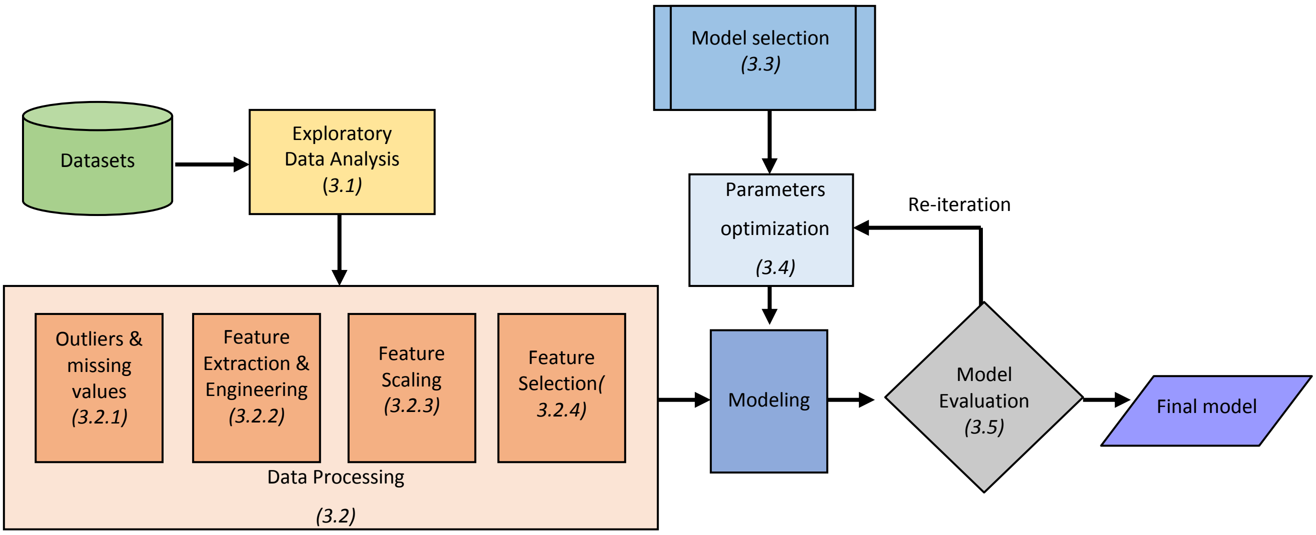

Figure 1. Standard machine learning pipeline.

The first step is to perform exploratory data analysis in data sets, then it is necessary to carry out data processing to prepare the application of the machine learning algorithm. In the next steps, it is necessary to select the models that best suit the problem, as well as define parameters that describe these models. Finally, models are implemented and evaluated to achieve the final model.

3.1. Exploratory Data Analysis

One of the key sub-steps during the machine learning project is to perform an exploratory data analysis (EDA). During this process, a preliminary analysis of various attributes of a given data set is carried out. It is important to become familiar with their properties, characteristic points and dependencies. It is also a step into which data visualization technique should be performed in order to better understand the data specificity, as well as to be able to discover previously invisible patterns [22].

3.1.1. Directly measured datasets

Available data comes from 100 residential homes located in France. The collected data includes the timestamp, total consumption data, appliances consumption data and weather data. The dataset was collected on a similar time span for all houses (slightly over 1 year) with hourly resolution, so the number of timestamps is in range from 8955 to 9136.

Figure 2 displays the total consumption for two of the households. The simple visualization of the data can already show the existence of missing data (in April, July, August in household B and in July in household A), outliers (two out-of-value points in household B data) and basic trends like intensified consumption in winter months (both households).

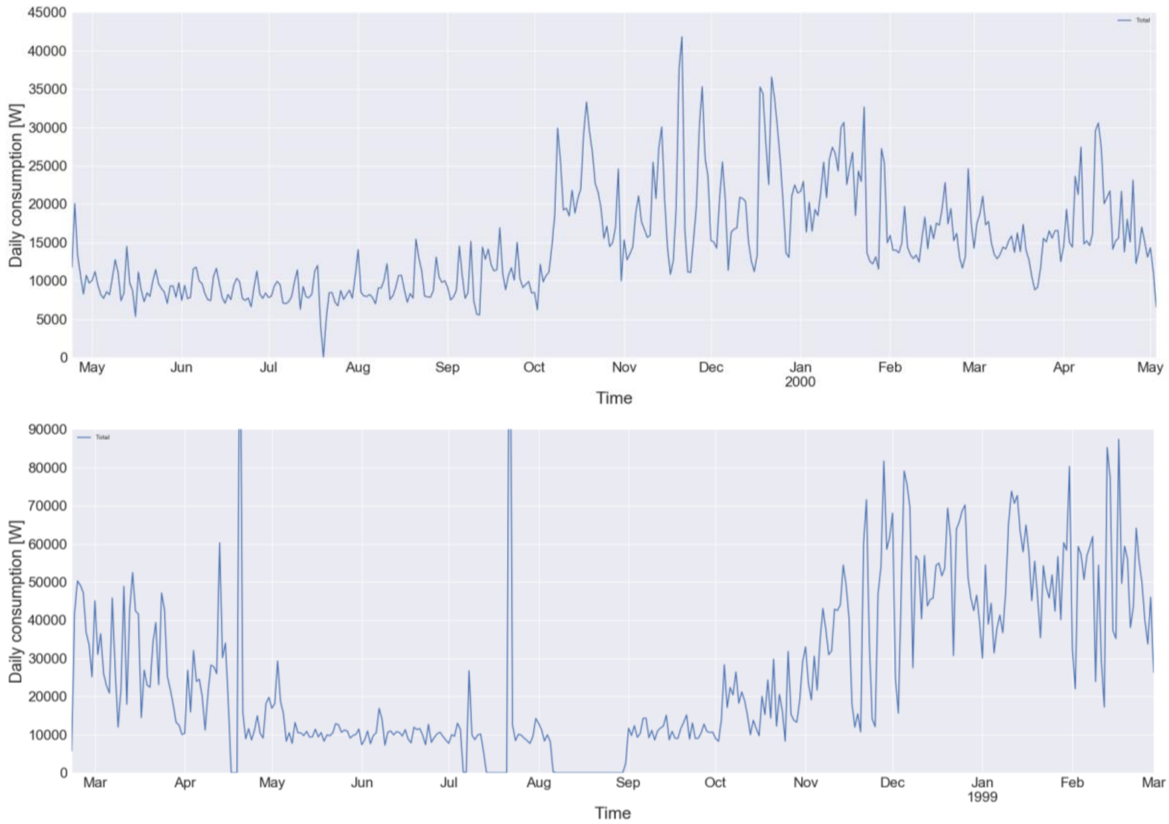

Figure2. Total daily electricity consumption of household A (top) and household B (bottom) over the measurement time span.

From the whole dataset, five houses that significantly differ in the level of electricity use were chosen to develop and test the methodology. Table 2 describes the details of the 5 households that have been chosen. In addition, a different number of devices is monitored - ranging from eight (8) for household A and C to twelve (12) for household B.

Table 2. Overview of the data available for households.

Household	Start date	End date	No. of samples	Daily mean consumption [kWh]	Measured values	Weather data
A (90)	24-04-99 7:00	03-05-00 15:00	9009	14.7	Total consumption Computer site Electric heating (altogether) Fridge (Kitchen, 236l) Microwave oven (2,75kW) TV (55cm) Vertical freezer (Garage, 291l) Washing machine Water heater	Temperature Wind strength Wind direction Humidity Radiance
B (28)	21-02-98 14:00	01-03-99 16:00	8955	25.7	Total consumption Chest freezer Clothes drier (without condenser) Dish washer (12 place setting) Electric Cooker (hot plate+oven) Electric heating (altogether) Fridge freezer Microwave oven Total site light consumption TV (61cm) TV (36cm) Washing machine Water heater	Temperature Wind strength Wind direction Humidity
C (03)	23-01-98 12:00	04-02-99 15:00	9052	12.1	Total consumption Chest freezer Dish washer Halogen lamp 1 Halogen lamp 2 Halogen lamp 3 Halogen lamp 4 Power supply for fuel boiler Washing machine	Temperature Wind strength Wind direction Humidity
D (36)	05-03-98 20:00	19-03-99 8:00	9085	15.2	Total consumption Chest freezer Clothes drier (without condenser) Dish washer (12 place setting) Electric heating (altogether) Fridge Total site light consumption TV (55cm) TV (36cm) Vertical freezer Washing machine Water heater	Temperature Wind strength Wind direction Radiance
E (53)	16-02-99 18:00	03-03-00 9:00	9136	41.7	Total consumption Chest freezer Dish washer (12 place setting) Electric heating (altogether) Fridge freezer Non halogen lamp 1 Non halogen lamp 2 Total site light consumption TV (63cm) TV (36cm) Washing machine Water heater	Temperature Wind strength Wind direction Humidity

Figure 3 plots the average daily profile for the 5 chosen households, which can also be explained by the different patterns of use of appliances of the inhabitants. The average energy consumption during the day varies considerably in each household. For example, in household A, there is a high consumption in the mornings and a increase in the evenings (before and after work / school). In the case of houses B and D, the peak of consumption appears shortly before or after midnight (this may be due to the efforts of residents aware of lower electricity prices at night). Household C is distinguished by the lowest energy consumption of all the analyzed houses, while house E maintains a very high consumption throughout the day, which may be explain by the existence of electric heating.

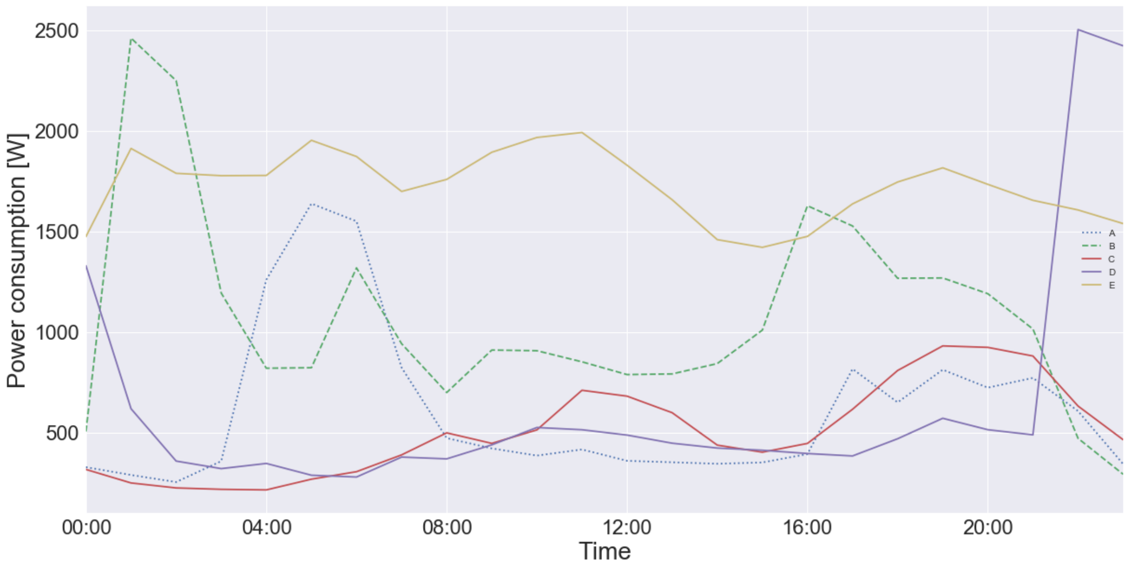

Figure 3. Average power usage during the day in each household.

Figure 4 presents overall statistics for households A to E. In case of households A, B and D only 65% to 70% of the total electricity consumption is described by the direct measurements of the appliances. The remaining part, 29.6% in household B, 34.3% in household D and 34.8% in household A, respectively, is the consumption of unspecified electric appliances. Extreme cases represent the other analyzed households - when it comes to C, as much as 62.3% of energy consumed came from unregistered devices. In the case of E, the opposite is true, namely only 6% of the total energy was consumed by unspecified devices, and as much as 94% by directly measured devices.

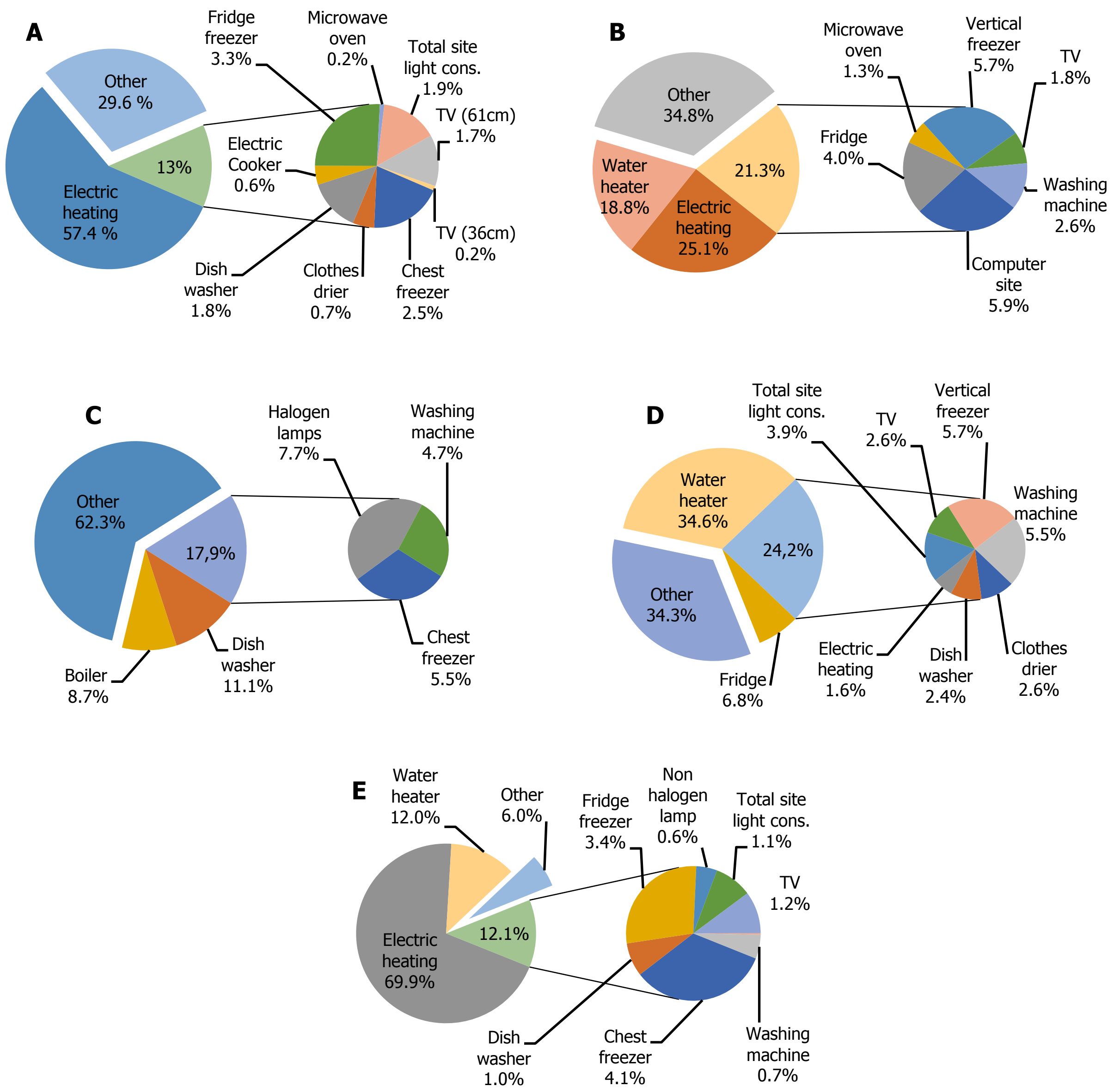

Figure4. Share of appliances in total electricity consumption in households A-E. The "Other" section represents part of the energy consumed by unspecified devices.

The electric heating has a significant impact in the overall consumption (between 25% for household B, 57.4% for household A and 69.9% for household E). In households B and D, the water heater represents also a large share of the total demand (18.8% for household B and 34.6% for household D). Finally, from the other appliances, the white appliances (fridges and freezers, dishwashers and washing machines) represent most of the consumption.

3.1.2. NILM disaggregation datasets

Non - Intrusive Load Monitoring (NILM) is a methodology that allows to identify and extract the load of the individual appliances from the total aggregated data and thus is a low-cost alternative to monitor individual sensors on each device. A file with disaggregated data using NILM was build and delivered by Watt-IS. The algorithm divides the total consumption into eleven separate categories, which do not necessarily match the original appliances. In addition, data from the original files such as temperature and humidity were used. The separated categories are:

1. Fridge
2. Cooking
3. Washing and dishwasher
4. Heating
5. Cooling
6. Water heating
7. Stand by
8. Electric vehicle
9. Lighting
10. TV
11. Other

To check the accuracy of NILM, two of the separated categories were analyzed, namely washing and dishwasher (W&D) and water heating (WH). This decision was made due to the fact that the devices representing these groups were measured separately in each of the selected houses. After comparing available initial data from direct measurements and data generated by NILM, each measurement point was assigned one of the four categories described below [23]:

1. true positive (TP) - model detects something as being true and the actual output is true (e.g., in both cases, NILM and direct measurement, the power consumption of the device was recorded in a given hour) ;
2. true negative (TN) - model detects something as being false and the actual output is false (e.g., in both cases, the device's energy consumption was not recorded in a given hour);
3. false positive (FP) - model detects something as being true and the actual output is false (e.g. NILM records the device's power consumption at a given time, and the device was not turned on in reality);
4. false negative (FN) - model detects something as being false and the actual output is true (e.g. NILM does not record the device's power consumption at a given time, and the device was turned on in reality).

The number of occurrences of individual categories was counted and on this basis the overall accuracy was calculated as the ratio of TP events (the number of events correctly recognized by NILM) to the sum of TP and FN events (the number of all events that took place in reality). The results obtained are presented in the table below (Table 3).

Table 3. The overall accuracy of NILM disaggregation of two categories in each household.

Household	W&D	WH
A	6.2%	30.4%
B	7.0%	42.1%
C	15.0%	22.5%
D	17.4%	66.6%
E	24.1%	47.0%

The results obtained differ significantly (almost 60 percentage points of difference in the most extreme cases), and the inability to evaluate other categories makes it impossible to check the accuracy of the NILM division. Despite this, the numbers above (Table 3) show that there is great potential for improving NILM results at the next iterations. Nevertheless, just like direct measurements, the features extracted by the NILM algorithm are expected to increase the capabilities of forecasting algorithms, by identifying hidden patterns of energy use and thus improve forecasting accuracy. On the other hand, attention should be paid to possible additional restrictions. The extracted energy consumption profiles of individual devices (or groups of devices) can significantly differ from real ones and the consequences can affect the opposite of the intended, namely, confuse the algorithm and degrade forecasting accuracy.

3.2. Data processing

Using directly collected raw data is, in most cases, useless for ML algorithms [21]. This is due to the fact that at the very beginning the data is often unstructured and contains missing and noisy data, which can significantly confuse algorithms in the training process. Therefore, data processing is the process of converting input data into the most suitable form for the ML algorithms. This include the following operations:

1. data cleansing - deleting or assigning new values to the entries that visibly diverge from others (outliers) or have been completely omitted (missing values), deleting duplicates;
2. representation transformation - converting qualitative features to numerical ones or vice versa;
3. feature construction - creating new features using mathematical functions or feature crossing;
4. feature selection - selecting the most important and influencing subset of the input features;
5. feature scaling - bringing values of different features into the same ranges.

Not all of the steps listed above are always required, their application depends on the state of raw data, selected models and the purpose of the study.

3.2.1. Outliers and missing values

Most ML models are highly sensitive to missing values and to outliers, i.e. values that are far from the average data, and therefore it is very important to delete or replace them with other values at the very beginning.

Outliers present in the data were identified using interquartile range (IQR), which is commonly used as a measure of how spread-out the values are. The 1.5 x IQR rule implies[24] that all values that are outside the range [Q_1-1.5xIQR; Q_3+1.5xIQR], where Q1 and Q3 represent the values of the first and third quartiles, and IQR equals the difference of these quartiles, can be considered an outlier (Figure 5).

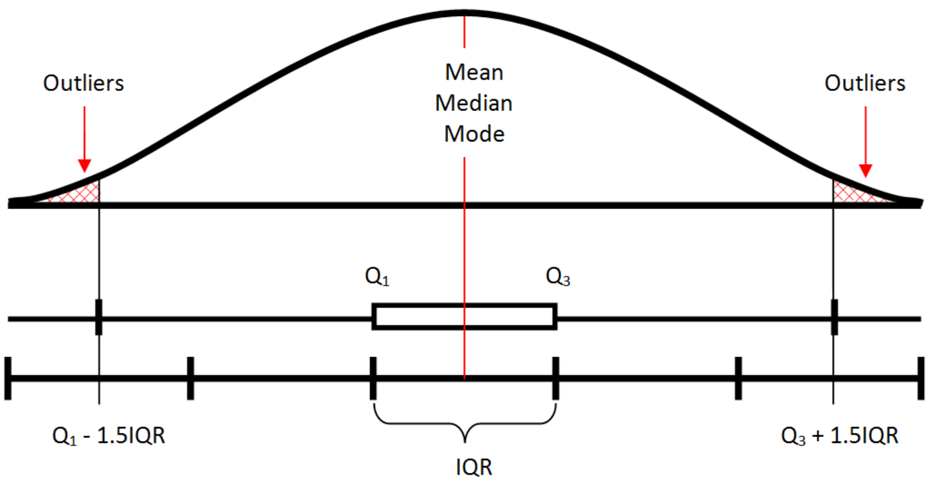

Figure 5. Visual representation of 1.5 x IQR rule.

In our specific case, the aggregated data may hide values beyond the expected scale -e.g. usage patterns at the lowest level vary widely and adding them can result in a gentle line showing overall consumption. In this case, it becomes more difficult to notice any anomalies. Therefore, the fact that measurements of individual devices were available was used, so the outliers were detected and corrected for individual appliances (Figure 6). The following procedure was used for each device:

1. the values of the first (Q_1), third (Q_3) quartiles and their difference (IQR) were calculated;
2. the values smaller than (Q_1-1.5xIQR) were identified and replaced with the value of the first quartile (Q_1);
3. the values larger than (Q_3+1.5xIQR) were identified and replaced with the value of the third quartile (Q_3).

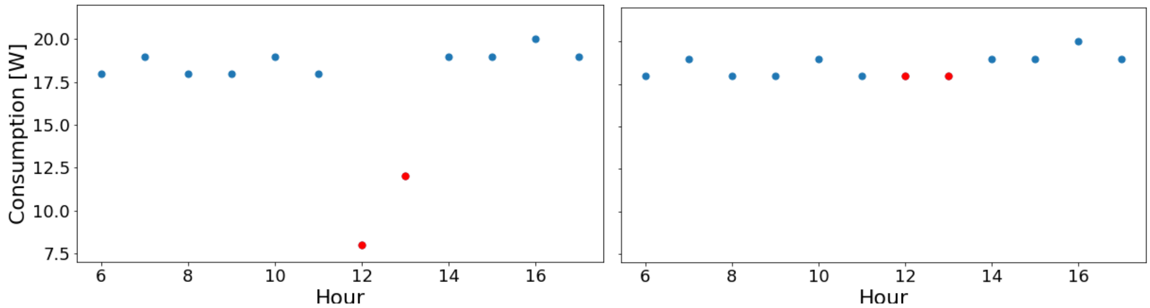

Figure 6. Identified outliers (left) and new values assigned to them (right).

Zero values, that in any of the used datasets did not exceed 1%, were completely removed. The decision was made on the basis of the fact that the number of samples is just a small fraction of the rest of the data and assigning the estimated values to the missing records could introduce errors to the data due to the highly dynamic changing patterns of power use in a single household.

3.2.2. Feature extraction and engineering

Feature extraction and engineering is a set of steps that ensures the definition of the best set of features that depict various representations of the underlying data. It is common to break down the available features into smaller parts, change their form, or combine them into new ones. Thanks to this, greater model accuracy, reduction of the possibility of overfitting, and reduction of learning time can be obtained.

At this stage, the following steps were carried out:

1. extraction of 4 new feature - number of the month (1-12), day (1-31), hour (0-23) and day of the week (0 - Monday, 6 - Sunday) - from the time stamp available in the raw data;
2. addition of the season designation - value 0 for winter months (1,2,12), 1 for spring and autumn months (3-5, 9-11) and 2 for summer months (6-8);
3. adding a column ('Holiday') containing information on non-working days (value 1 for non-working days, value 0 for working days);
4. addition of a column containing the level of power used by unmeasured devices (overall consumption minus the sum of the load of measured devices);
5. combining 'Weekday' and 'Holiday' features into one (by multiplying their value) - this decision was made because of the relationship between these two features found during the EDA;
6. creation of columns containing power consumption in the previous hour of each device (due to data with a one-hour time step);
7. creation of column containing total power consumption at the corresponding day and hour from previous week - it was assumed that the weekly plan of the household is quite repetitive,

and changes in consumption caused by changes in the seasons are not very significant for one week (as opposed to one month).

3.2.3. Feature scaling

Data scaling is another step in the initial data processing. It is not required with every dataset, but ML algorithms usually do not work well when the input data has attributes that differ significantly in scales. For example, in the case discussed in this paper, the power consumption in the previous hour may range from 20 W to 4500 W, and temperature fluctuations are on a scale from -9.6°C to +31.6°C. Here, the first feature with high magnitude will weigh a lot more in distance calculations then the other, smaller in magnitude, feature. To overcome this effect, all features should be brought to the same level of magnitudes. There are different ways to scale the available data, but in this work the three most popular of them were considered:

1. mean normalization,
2. data scaling,
3. data standardization.

Mean normalization scales the data in such a way that all the values fall between -1 and 1 with their mean equals zero (μ = 0). To achieve this, the following equation is used (1) :

$$x'(i) = \frac{x(i) - mean(\boldsymbol{x})}{max(\boldsymbol{x}) - min(\boldsymbol{x})} \quad \textbf{(1)}$$

where x(i) is the sample *i* of the total data set **x**.

Data scaling consists of changing the data into values between 0 and 1. The following formula is used (2) :

$$x'(i) = \frac{x(i) - min(\boldsymbol{x})}{max(\boldsymbol{x}) - min(\boldsymbol{x})} \quad \textbf{(2)}$$

Data standardization replaces all the values by their Z-scores, which are the numbers of standard deviations from their means. These numbers are calculated with the formula (3). After this process the values have mean equals zero and standard deviation equals 1.

$$x'(i) = \frac{x(i) - mean(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \quad \textbf{(3)}$$

The basic models were tested for each of the three forms of input data, and thus the method of data standardization was chosen.

3.2.4. Feature selection

To select the appropriate features, 4 subsets have been prepared: 1 set with all the available features and then 3 subsets that result from the application of 3 feature selection algorithms:

1. Extra Trees Classifier (ETC)
2. Recursive Feature Elimination (RFE)
3. "K-Best" feature selection

ETC, which helps to choose the most informative features by calculating the features importance, is a variation of a random forest, with the difference that the divisions are selected randomly (rather than using specific criteria) and the entire sample is used at each step. "K-Best" function enables selecting the best features based on univariate statistical tests. Finally, RFE selects the desired number of functions recursively considering smaller and smaller sets of features. It requires declaring supervised learning estimator, in this case linear regression was used.

The sets created using the above-described functions ultimately contained the 7 highest-rated attributes.

3.3. Model selection

Wolpert and Macready wrote in 1997 a famous paper that presents the following statement "(...) if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems." [25]. It follows that one cannot be guided by the best results of one model in one field and on this basis apply it to problems in other areas, because each data and each case requires different setups. For this reason, choosing the right algorithms is extremely important and, ideally, testing all possible models to find the one that best reflects the issue should be implemented. Unfortunately, due to limitations such as time and processor capabilities, a preliminary selection should be made, based on the literature reviewed in Chapter 2 for similar problems.

For this reason, looking at the current solutions in the subject of forecasting electricity consumption, three models were selected for analysis:

1. Random Forests (RF)
2. Multi-Layer Perceptron (MLP)
3. Long-Short Time Memory (LSTM)

3.3.1. Random Forest

Random Forests (RF) are the example of ensemble methods and are among the most powerful Machine Learning algorithms available nowadays [26]. The main idea of ensemble learning is to combine (ensemble) weak learners to build a strong learner - a robust model, which has a better generalization error and is less prone to overfitting [27]. Therefore, RF action is based on training many decision-tree models on subsets of the input data and input features (Figure 7). Each decision tree includes a random subset of features and only reaches to a random set of the training data points. These individual trees then contribute to the total outcome. This can be presented by following steps [27] :

1. random selection of *n* samples from the training set for *k* trees,
2. building a decision tree on the randomly selected *d* features,
3. repeating steps 1 and 2 *k* times,
4. aggregating the prediction by each tree.

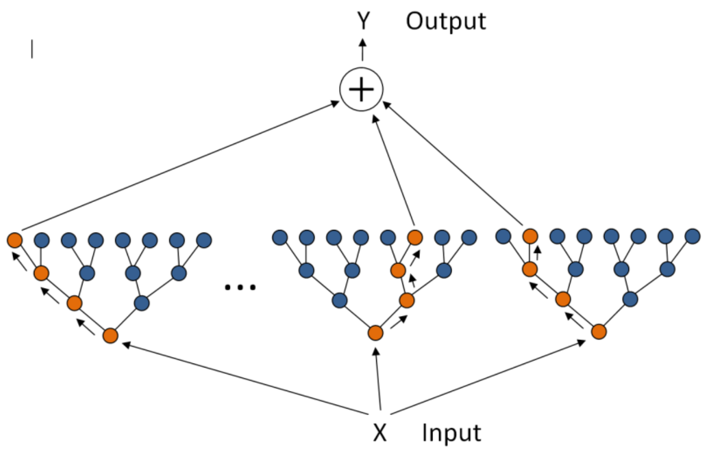

Figure 7. Visualization of Random Forest algorithm.

RF models are quite robust models resistant to interferences appearing in the data. They also do not require fine tuning of many hyperparameters, but in return represent a strong generalization.

3.3.2. Multi-Layer Perceptron

The Perceptron was introduced in 1957 by Frank Rosenblatt and is known as one of the simplest ANN architectures. It is based on the neuron called a linear threshold unit (LTU), where each input connection is associated with a weight. The LTU works in the following way: first it computes the weighted sum of its inputs, and in a second step, it applies a step function to that sum and finally it outputs the result.

Multi-Layer Perceptron (MLP), as the name suggests, is based on combining multiple Perceptrons and is compose of three main elements: one input layer, one (or more) layers of LTUs, called hidden layers, and one final layer of LTUs called the output layer (Figure 8). This architecture is used in wide range of applications from different areas and time series forecasting is among the most important ones. The beginning of vast applicability of those networks started with presentation of the learning algorithm called backpropagation which is now used for MLP's training. MLP networks are the examples of feedforward architecture, which means that the flow within the network occurs only in one direction, namely from the input to the output layer.

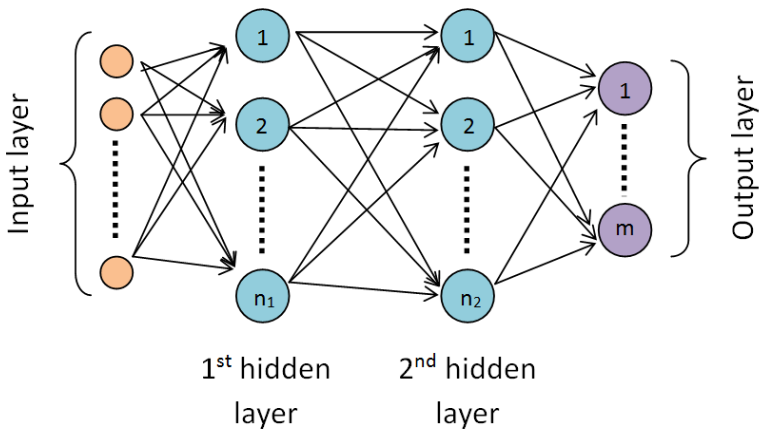

Figure 8. Structure of MLP network with two hidden layers.

MLP network training using the backward propagation algorithm usually takes place through the implementation of two specific stages [28] :

1. forward propagation, in which the values from the training set are inserted into network inputs and are propagated layer by layer, until appropriate results are obtained,
2. the responses generated by the network outputs are compared with the appropriate available desired responses (this is the assumption of supervised learning) and based on this, the weights of all neurons are corrected.

In summary, the use of the above steps enables automatic adjustment of synaptic weights and neuron thresholds in each iteration. This consequently leads to a gradual reduction in the sum of errors generated by network responses relative to the desired responses.

3.3.3. Long Short-Term Memory

Long Short-Term Memory (LSTM) network is an example of a Recurrent Neural Network (RNN). One of the problems of typical RNN networks is that if the sequence fed into the network is extremely long, then there may be a problem with the transfer of information from earlier time steps to later ones. In this way, information from the beginning of the learning process may be lost. LSTM is one of the solutions of the short-term memory mentioned above. Thanks to the built-in internal elements called gates, it can regulate the flow of relevant information in the learning process (Figure 9). These gates can decide which data in the sequence is important to keep or discard. In this way, it can convey relevant information along a long chain of sequences to make predictions.

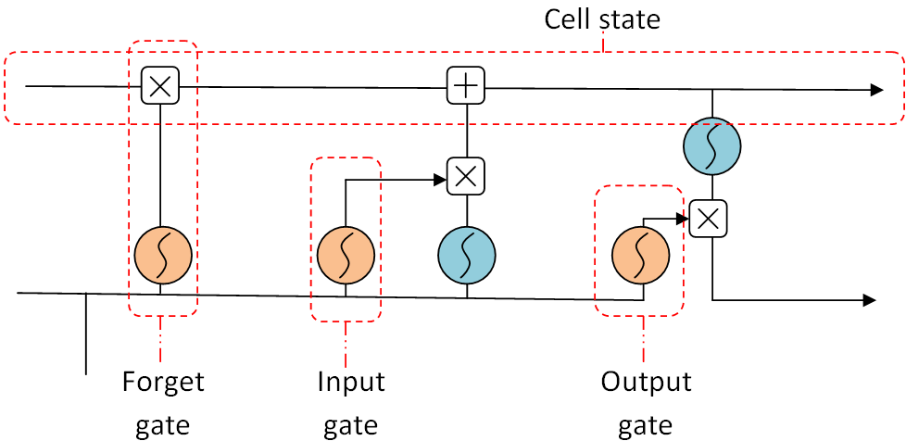

Figure 9. Schematic structure of part of LSTM network.

There are three gates in the architecture of the LSTM network:

1. forget gate - this gate is used to determine what information should be discarded or saved for later stages. The sigmoid function helps to regulate the information from the previous hidden state and the information from the current input. The values obtained range from 0 to 1, where a value of zero means that the information will be forgotten, while the value 1 will keep and pass it on. This gate decides what information is relevant to keep from previous steps;
2. input gate - this gate is used to update the cell status. The previous hidden state and current input are again passed to the sigmoidal and, additionally, hyperbolic tangent functions, and then both outputs are multiplied. The input gate decides what information is relevant to add from the current step;
3. output gate - this gate decides which part of the current cell is passed to the output. Again, both functions (sigmoid and hyperbolic tangent) are used.

3.4. Parameters optimization

To provide the best results, it is rarely possible to use the basic set of algorithms available in various ML dedicated libraries. These algorithms have various specific parameters whose values must be properly tuned to the different data, project requirements, and expected results. Bearing in mind the bias-variance trade-off, a so-called good fit is sought to avoid model underfitting (the algorithm does not learn the hidden data structure) or model overfitting (the algorithm is too adapted to the training data set).

After a preliminary analysis to determine which of the parameters have the greatest impact on the selected models, the following of them were selected to optimization:

1. RF
    a. The number of trees in the forest (a) - higher number of this parameter makes the predictions more accurate and stable. However, it should be remembered that increasing this number is equivalent to extending the model's learning time.
    b. The minimum number of samples required to split an internal node (b) - increasing this parameter makes each tree more constrained (more samples are needed to be considered at each node).
    c. The minimum number of samples required to be at a leaf node (c) - the larger this parameter, the greater the model's resistance to catch noise in the learning data set. [29]
2. MLP
    a. Activation function (d) - different neuron activating functions work differently for different cases. One that maximizes the accuracy of the model and minimizes training time should be found.
    b. Parameter alpha (e) - regularization parameter which helps prevent overfitting by applying weight sizes.
    c. The number of neurons in each hidden layer (f) - this is one of the most important parameters of the model. Too few neurons can result in model underfitting, while too many in model overfitting and unnecessarily long training time.
    d. The number of hidden layers (g) - usually 0 (for linear cases), 1 or 2. Models with 2 hidden layers can represent various functions and there is no need of increasing this number. [30]
    e. Maximum number of iterations (h) - this number determines how many times each data point will be used during the training.
3. LSTM
    a. The number of neurons (i) - same as the MLP model.
    b. The number of epochs (j) - determines the number of uses for each data point.

The optimization process consisted of three phases in which another set of parameters was tested. For RF and MLP models, optimal parameters were found using two additional algorithms: Random Search and Grid Search. The first of these allows you to search faster a larger range of parameters and their values, so it is possible to quickly gain intuition about ranges and the impact of parameters on the model. It was used in the first phase of searching. In the next phase, a second algorithm was used, when the extent to which parameters should be sought was already expected. Following the practices of the authors of various studies (like [6] ,[15], [19], [20]) the data for both models was divided as follows: 80% for training and 20% for testing. In addition, 5-cross validation was used (dividing the training set into 5 subsets and using one of them for validation). In the case of the LSTM model, due to the source code, the search for final parameters was carried out manually, which was considered to be sufficient for prototyping. In this case, the data has been divided into three parts in advance: 70% for the training set, 20% for the testing set and 10% for the validating set.

Hyperparameter optimization was carried out for household A, and then the best configurations were used for the next datasets.

3.5. Model evaluation

After building the model, it is important to subject it to an assessment process based on specific criteria for assessing its accuracy and performance. This is most often done using pre-defined functions that allow you to get a numerical value that helps to decide on the effectiveness of a given model and allows to compare models with each other. In order to accomplish that, the following parameters were measured and used:

1. Mean Absolute Percentage Error (MAPE)
2. Mean Squared Error (MSE)
3. Mean Absolute Error (MAE)
4. Coefficient Of Determination (R^2)

3.5.1. MAPE

The mean absolute percentage error is commonly used in comparing regressive models. It is presented as a percentage and calculated on the basis of the following formula :

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \qquad \textbf{(4)}$$

A_t and F_t represent the actual value and forecast value, respectively, while n stands for the number of observations.

It is important to mention that MAPE imposes a greater penalty on negative errors ($A_t < F_t$) than on positive errors. This results in a biased behaviour of this parameter, which regularly selects models whose forecasts are underestimated in relation to real values. [31]

3.5.2. MSE

The mean squared error, also known as mean squared deviation (MSD) measures the average squared difference between the forecasted values (F_t) and the actual values (A_t) of the observations (n). It is expressed in the squared units of measurements and calculated in the following way:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2 \quad \textbf{(5)}$$

Like MAPE, MSE is also biased and imposes a greater penalty on large errors than on small ones (the result of squaring the value) [32]. Therefore, it is important to pay attention to the existence of any outliers in the data.

3.5.3. MAE

The mean absolute error is calculated by averaging the absolute differences between the forecasted values (F_t) and the actual values (A_t) of observations (n). This value can be represented by the formula:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|F_t - A_t| \quad \textbf{(6)}$$

The advantage of MAE is that every error affects it in direct proportion to its absolute value [33]. The disadvantage is that as it is an absolute value, cannot be used to compare results between different datasets.

3.5.4. R^2

The coefficient of determination, also commonly known as R-squared, is an dimensionless indicator and is determined by the formula:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{\sum_{t=1}^{n}(A_t - \bar{A})^2} \quad \textbf{(7)}$$

Here, again, A_t and F_t represent the actual value and forecast value, respectively, n stands for the number of observations and $\bar{A}$ is equal to the mean value of the observed data. As can be deducted from the formula, a value of R^2 equals 1 indicates a perfect fit ($A_t = F_t$).

4. Results

4.1. Final parameters

Various parameter configurations and evaluations of subsequent optimization steps can be found in Annex I. During the evaluation of the models, the results obtained for the test set were also monitored, and in the event of an increasing tendency to overfitting, the relevant parameters were tuned. The best models were selected based on MSE values. However, when these values differed by only 10%, the time needed for the model learning process was taken into account. The final parameters used to build individual models are presented in the table below (Table 4):

Table 4. Final parameters used for RF, MLP and LSTM models.

		Direct metering	NILM	Aggregated
RF	a	500	750	1700
	b	3	2	2
	c	10	7	10
MLP	d	relu	relu	relu
	e	0.00001	0.05	0.05
	(f,g)	(150, 2)	(100, 2)	(100, 2)
	h	400	400	400
LSTM	i	250	350	3
	j	100	80	200

The final forecasts were based on data consisting of all available features, because the best results were obtained using this particular set. In addition, data standardization has provided better results than normalizing or scaling.

4.2. Forecast

Each of the data sets has been analyzed and processed, to prepare them to be used in optimized models as best as possible. The following results were obtained for forecasting household A one-hour-ahead energy consumption. In addition to the measured evaluation values, the time required for the learning process of each model is given.

From the data presented below (Table 5), it can be seen that the best results for household A were obtained using the RF model for aggregated data. The results are similar for the model using data from direct measurements, and in this case, the values do not differ significantly from those obtained with the MLP and LSTM models. For NILM and aggregated data, both MPL and LSTM present significant higher errors that RF. In addition, RF is the model that required one of the shortest training time. In conclusion, it can be said that RF seems to be better model and that for aggregated consumption forecast the use of direct consumption of appliances or NILM information does not improve the results.

Table 5. Evaluation of household A.

	Direct metering			NILM			Aggregated		
	RF	MLP	LSTM	RF	MLP	LSTM	RF	MLP	LSTM
MAPE	31.9 %	41.6 %	37.0 %	36.4 %	52.2 %	39.2%	**30.9 %**	40.6 %	59.1 %
MSE [kWh^2]	90.59	98.97	110.58	101.77	141.44	125.55	**86.84**	105.72	137.45
MAE [kWh]	0.15	0.18	0.19	0.18	0.22	0.21	**0.16**	0.19	0.24
R^2	0.74	0.71	0.74	0.75	0.66	0.75	**0.64**	0.56	0.43
Time [s]	14.4	15.8	1492.3	14.7	15.5	2539.2	**14.7**	57.1	107.7

The graphs presented below show the prediction of power consumption in household A grouped by data set (Figure 10)- with direct measurement (top), using NILM (middle) and with only total consumption (bottom). It can be seen that all models (for all data sets) had difficulty forecasting energy consumption peaks, and in most cases underestimated the actual value. Graph with the aggregated data deviate from the other two due to the specificity of the method used to deal with outliers.

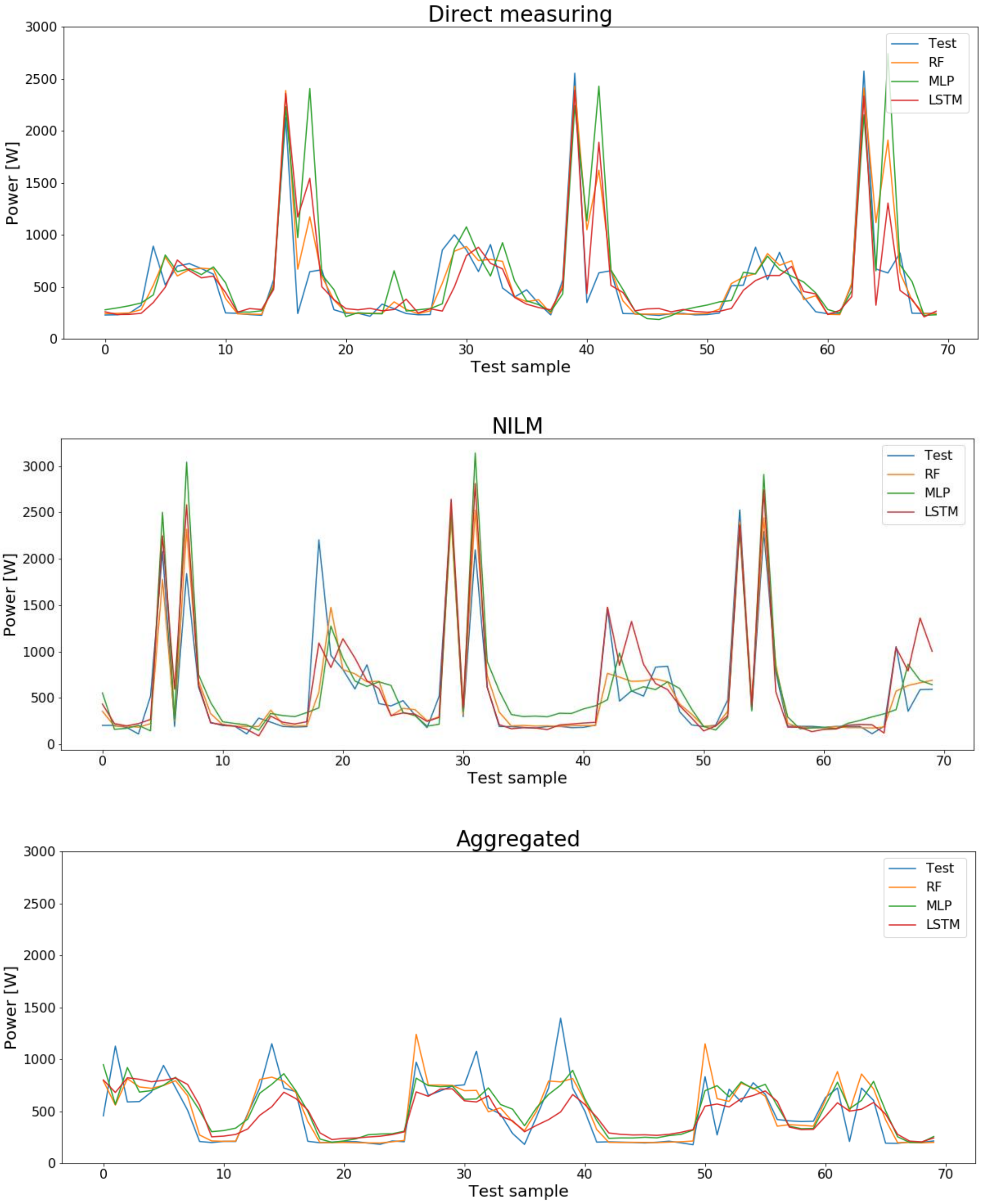

Figure 10. Results of RF, MLP and LSTM models on directly measured (top), NILM (middle) and aggregated (bottom) datasets of household A.

The results from the evaluation of the final models on the other four households are presented below (Table 6-9). In all the cases considered, again, the model that provided the most accurate results was RF with MAPE = 27.6% in the best scenario and an MAE between 0.1 kWh and 0.39 kWh depending on the type of data and the household. The MLP model performed a little worse, but in some cases the results of both models are comparable. Also, the training times of both models were similar (from 9 to 19 seconds with one exception). The LSTM model turned out to be the least accurate, and additionally requires a much longer training time (over 16 minutes on average). The reason for this behaviour of the LSTM model may be that relatively few data points were used (less than 8000 for each household), and these models require more data for proper training than models such as RF.

Table 6. Evaluation of household B.

	Direct metering			**NILM**			**Aggregated**		
	RF	MLP	LSTM	RF	MLP	LSTM	RF	MLP	LSTM
MAPE	76.6 %	97.9 %	80.4 %	63.5 %	97.4 %	93.6 %	63.9 %	80.9 %	118.7 %
MSE [kWh^2]	453.53	601.13	1198.43	372.92	582.62	1321.37	346.05	425.73	1378.38
MAE [kWh]	0.39	0.49	0.82	0.36	0.50	0.78	0.33	0.39	0.92
R^2	0.74	0.66	0.38	0.77	0.64	0.30	0.77	0.72	0.32
Time	13.8	13.9	1253.4	14.4	11.5	1677.3	9.8	12.9	112.6

Table 7. Evaluation of household C.

	Direct metering			**NILM**			**Aggregated**		
	RF	MLP	LSTM	RF	MLP	LSTM	RF	MLP	LSTM
MAPE	37.2%	40.9 %	40.6 %	45.4%	50.2%	49.8%	31.4 %	34.3 %	33.1 %
MSE [kWh^2]	79.87	86.98	122.85	106.14	112.73	110.64	49.38	52.31	107.94
MAE [kWh]	0.17	0.18	0.23	0.21	0.22	0.28	0.13	0.14	0.20
R^2	0.51	0.47	0.34	0.41	0.37	0.31	0.52	0.50	0.31
Time	12.6	14.8	1576.2	15.9	14.6	1543.7	9.9	15.0	111.2

Table 8. Evaluation of household D.

	Direct metering			**NILM**			**Aggregated**		
	RF	MLP	LSTM	RF	MLP	LSTM	RF	MLP	LSTM
MAPE	31.0%	41.9%	49.8%	41.1%	50.6%	50.1%	27.6%	29.9%	37.4%
MSE [kWh^2]	100.03	139.06	233.79	127.46	164.37	172.50	28.53	29.00	75.17
MAE [kWh]	0.15	0.19	0.26	0.19	0.24	0.31	0.10	0.11	0.15
R^2	0.81	0.74	0.63	0.74	0.66	0.53	0.40	0.39	0.27
Time	18.15	14.77	1328.52	17.5	15.0	1683.5	9.2	14.7	118.9

Table 9. Evaluation of household E.

	Direct metering			NILM			Aggregated		
	RF	MLP	LSTM	RF	MLP	LSTM	RF	MLP	LSTM
MAPE	**38.6%**	42.1%	68.8%	42.1%	51.6%	51.7%	41.3%	47.8%	81.7%
MSE [kWh^2]	**255.21**	275.46	700.88	242..82	261.20	274.90	318.88	329.43	1100.00
MAE [kWh]	**0.29**	0.31	0.56	0..30	0.32	0.41	0.33	0.36	0.66
R^2	**0.88**	0.87	0.63	0..88	0.87	0.76	0.85	0.85	0.49
Time	**17.2**	14.6	2013.6	16..3	15.1	1621.2	16.5	59	172.8

Considering only the best performing model (RF), the values of individual evaluation parameters are visualized below (Figure 11). With the exception of house E and the MAE values of house A and B, again, the lowest MAPE, MSE and MAE values were obtained when forecasting using data set only with aggregated data, therefore excluding the power consumption of individual devices.

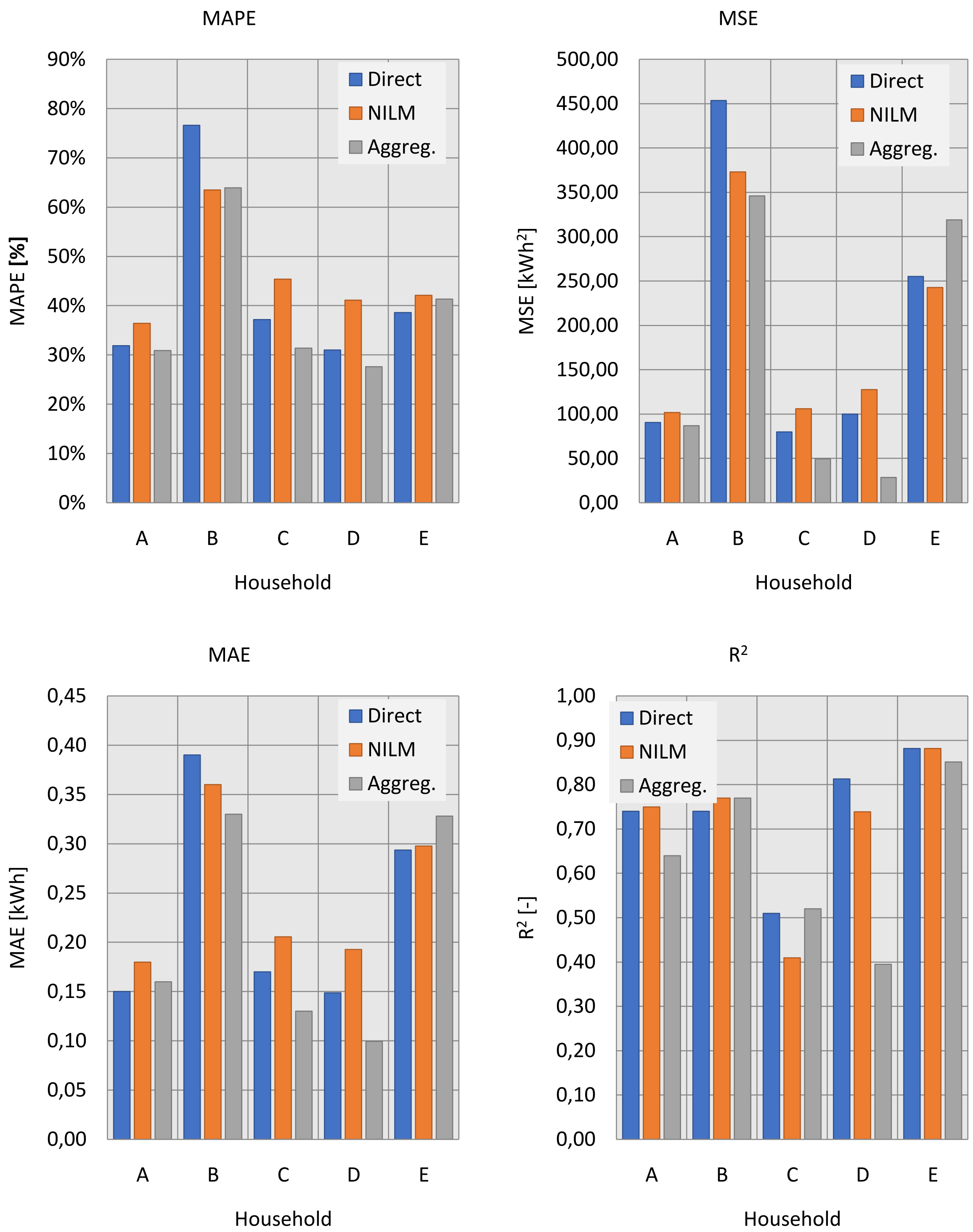

Figure 11. Evaluation parameters visualization : MAPE (top left), MSE (top right), MAE (bottom left) and R^2 (bottom right).

Therefore, in all cases, except household E, the use of aggregated data is better or at least as good as the use of direct data. Only in household B, the use of NILM information is comparable to the aggregated forecast. Therefore, a significant conclusion can be drawn, namely that for aggregated forecast of energy consumption, providing detailed information of the appliances consumption (through direct measurements or NILM estimations) does not improve the forecasting results.

5. Discussion

In the following section, based on the results presented in the previous chapter, an attempt will be made to answer the research questions posed in chapter two.

Is there an advantage in using disaggregated data over aggregated data?

It was expected that models based on disaggregated data, due to the possibility of more detailed pattern recognition, could work better than models based on aggregated data. Looking at the results in figure 11, it turned out that this did not happen in general. In almost every case considered (A-E homes; RF, MLP, LSTM algorithms) data containing only aggregated data provided the most accurate results to perform aggregated consumption forecast. Moreover, it was easier and faster to process, both during data preparation and model training. Less accurate results obtained on disaggregated data may be due to insufficient amount of data, which did not allow individual models to notice and extract more intricate patterns of device behaviour. This caused that these additional features only unnecessarily confuse the model and do not bring the expected benefits. Therefore, the number of devices considered should be limited. The more immediate focus should be only on the most energy-intensive devices (e.g. heating, air conditioning), and those that are switched on sporadically, consume an extremely small amount of energy or are used in the same way throughout the year (e.g. TV) should be discarded.

Also, this encourages to investigate different approach in the future: focusing more on the probabilities of using a given device in a given hour of the day than on the power it consumes. For this it is worth using the measured values at shorter intervals (5-30 min), because a large number of appliances in the household has a duty cycle shorter than one hour. This approach will, probably, allow to better observe different behavioural patterns and obtain more accurate forecasting results.

Is NILM algorithm necessary to significantly improve the forecast?

The results obtained in the analysis showed that models using NILM data provided similar results as models using data with direct measurements of energy consumption by individual devices. Still, the first option is much easier and cheaper to obtain than the second, which requires expensive installation and supervision of a large number of sensors. However, compared to aggregated data, results obtained using NILM data are less accurate. Moreover, the latter requires an additional step, which is the separation of total consumption into individual devices or device subgroups.

Worse results may be due to the fact that the algorithm for obtaining disaggregated data was not tuned enough to individual houses. In addition, different cases should be re-examined using the approach suggested above - using data from only the most energetically intensive devices with less than one-hour time steps.

Is the same model sufficient to forecast power consumption of different houses?

The created model was built and optimized specifically for a single household (A), and then the performance of the same model was tested in other households with different energy consumption patterns. Due to the multitude of models, it was decided to focus on the results provided by the most accurate model (Table 10).

Table 10. Evaluation of RF model based on aggregated data of different households.

	A	B	C	D	E
MAPE	30.9 %	63.9 %	31.4 %	27.6%	41.3%
MSE [kWh^2]	86.84	346.05	49.38	28.53	318.88
MAE [kWh]	0.16	0.33	0.13	0.10	0.33
R^2	0.64	0.77	0.52	0.40	0.85
Time [s]	14.7	9.8	9.9	9.2	16.5
Daily avr. [kWh]	14.7	25.7	12.1	15.2	41.7

The highest values of MAPE, MSE and MAE were obtained for houses B and E, but it should be taken into account that they were also the most energy-intensive cases. On the other hand, they are distinguished by the highest R^2 result (higher than for house A) - various energy peaks appearing in their test data sets have been very well mapped in the model. House D presents a completely opposite situation - although it has a similar average daily energy consumption to the reference house A, it obtained the best results of MAPE, MSE and IEA, and the worst when it comes to R^2.

The results obtained indicate that similar results should not be expected even if the format and amount of data fed with the algorithm is very similar. Individual houses, which in the vast majority of cases, are characterized by exceptional irregularity of energy use and differentiation of the energy level, are extremely difficult to reproduce in one universal model. The next step would be to compare the accuracy of individual models on different subgroups of houses - clustered by similar average energy consumption, a similar pattern of energy use during the day and others.

6. Conclusions

The main goal of the presented research was to examine whether direct measurements of energy consumption by individual devices improve the forecasting of total energy consumption in an individual household. In addition, it was checked whether a single model is able to provide results at a similar level of accuracy for different houses, where the fed data for the algorithm are of similar structure and size. During the search for the most accurate forecast model, various machine learning methods, several configurations and different data representations were implemented and compared. Five different households were used to verify the results, given by three different datasets: aggregated power consumption, disaggregated data collected during direct measurement of power consumption of different devices and disaggregated data according to the NILM algorithm provided by Watt-IS. The results obtained did not show the superiority of disaggregated data over aggregated data. Nevertheless, it is recommended to carry out further research with the following improvements proposed below:

- check larger parameter ranges (more accurate model optimization);
- analyze other machine learning algorithms (e.g. SVR);
- create separate models for forecasting the energy consumption of individual devices and then add them together to obtain the total electricity consumption;
- use the measurements of individual devices to create a map of the probability of switching on particular devices at a given hour during the week and using it as an additional feature fed in the algorithm;
- use data with less than one-hour time step (e.g. 10 minutes or even 1 minute);
- use data collected over a longer time period than one year to avoid under- and overfitting of the algorithms;
- carry out research on a more recent data set (different patterns and energy levels may have changed significantly over the last 20 years);
- analyze models with a larger forecasting horizon (e.g.3 hours / 24 hours ahead).

The use of energy at higher levels, when many consumers are aggregated, forms a smooth pattern with clearly marked and very often repeatable consumption peaks. Accurate estimation of energy consumption at the level of a single household provides exceptional difficulties due to the extremely variable pattern of energy consumption. Still, due to the increasing penetration of renewable energy in the network, increasing efforts to reduce energy consumption, the emergence of individual power generators (e.g. through residential solar panels), as well as the need to ensure the most secure supply of energy, increased efforts to energy management improvements must be undertaken. This in turn may mean a compulsory step towards forecasting energy consumption at the level of a single household and NILM disaggregation may be a key to high accuracy of those forecasts.

References

[1] European Commission, "Clean energy for all Europeans," Publications Office of the European Union, Luxembourg , 2019.

[2] Eurostat, "Final energy consumption by sector," 23 04 2019. [Online]. Available: https://ec.europa.eu/eurostat/web/products-datasets/-/ten00124. [Accessed 13 10 2019].

[3] S. T. Tzeiranaki, P. Bertoldi, F. Diluiso, L. Castellazzi, M. Economidou, N. Labanca, T. R. Serrenho and P. Zangheri, "Analysis of the EU residential energy consumption: Trends and determinants," *Energies,* vol. 12, no. 6, 2019.

[4] Eurostat, "Manual for statistics on energy consumption in households," Publications Office of the European Union, Luxembourg, 2013.

[5] B. Johansson, "Security aspects of future renewable energy systems - A short overview," *Energy,* vol. 61, pp. 598-605, 2013.

[6] K. Gajowniczek i T. Zabkowski, „Electricity forecasting on the individual household level enhanced based on activity patterns," *PLoS ONE,* tom 12, nr 4, pp. 1-26, 2017.

[7] F. Gangale, J. Vasiljevska, C. F. Covrig, A. Mengolin i F. Gianluca, „Smart grid projects outlook 2017: Facts and figures," Publications Office of the European Union, Luxembourg, 2017 The.

[8] B. Dong, Z. Li, S. M. Rahman and R. Vega, "A hybrid model approach for forecasting future residential electricity consumption," *Energy and Buildings,* vol. 117, pp. 341-351, 2016.

[9] M. Benedetti and V. I. V. S. J. Cesarotti, "Energy consumption control automation using Artificial Neural Networks and adaptive algorithms: Proposal of a new methodology and case study," *Applied Energy,* vol. 165, pp. 60-71, 2016.

[10] T. K. Wijaya, M. Vasirani, S. Humeau and K. Aberer, "Cluster- based aggregate forecasting for residential electricity demand using smart meter data," in *2015 IEEE International Conference on Big Data*, Santa Clara, CA, 2015.

[11] S. Hosein and P. Hosein, "Building Energy Load Forecasting using Deep Neural Networks," in *7th IEEE International Conference on Smart Grid Communications*, Sydney, 2016.

[12] C. Dinesh, S. Makonin and I. V. Bajic, "Residential Power Forecasting Using Load Identification and Graph Spectral Clustering," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 7747, 2019.

[13] N. Al Khafaf, M. Jalili and P. Sokolowski, "Application of deep learning long short-term memory in energy demand forecasting," *Communications in Computer and Information Science,* vol. 1000, pp. 31-42, 2019.

[14] S. V. Oprea, A. Pîrjan, G. Cărutasu, D. M. Petrosanu, A. Bâra, J. L. Stănică and C. Coculescu, "Developing a mixed neural network approach to forecast the residential electricity consumption

based on sensor recorded data," *Sensors (Switzerland),* vol. 18, no. 5, 2018.

[15] R. Sereno and V. Alves, "Application of Neural Networks' Models to Predict Energy Consumption," Instituto Superior Técnico, Universidade de Lisboa, Lisbon, 2016.

[16] K. Amarasinghe, D. L. Marino and M. Manic, "Deep neural networks for energy load forecasting," *IEEE International Symposium on Industrial Electronics,* pp. 1483-1488, 2017.

[17] C. Li, Z. Ding, D. Zhao, J. Yi and G. Zhang, "Building energy consumption prediction: An extreme deep learning approach," *Energies,* vol. 10, no. 10, pp. 1-20, 2017.

[18] C. Deb, L. S. Eang, J. Yang and M. Santamouris, "Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks," *Energy and Buildings,* vol. 121, pp. 284-297, 2015.

[19] J. Moon, J. Park, E. Hwang and S. Jun, "Forecasting power consumption for higher educational institutions based on machine learning," *Journal of Supercomputing,* vol. 74, no. 8, pp. 3778-3800, 2018.

[20] F. Rodrigues, C. Cardeira and J. M. Calado, "The daily and hourly energy consumption and load forecasting using artificial neural network method: A case study using a set of 93 households in Portugal," *Energy Procedia,* vol. 62, pp. 220-229, 2014.

[21] D. Sarkar, R. Bali i T. Ghosh, „Machine Learning Fundamentals," w *Hands-On Transfer Learning with Python : Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras*, Packt Publishing, 2018, pp. 9-29.

[22] P. Ankur A., Hands-On Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions From Unlabeled Data, Sebastopol: O'Reilly Media, Inc., 2019.

[23] L. Pereira i N. Nunes, „Performance evaluation in non-intrusive load monitoring: datasets, metrics, and tools - a review," *WIREs Data Mining Knowl Discov.,* 2018.

[24] T. A. Chawsheen and I. S. Latif, "Detection and treatment of outliers in data sets," *Iraqi Journal of Statistical Science,* vol. 9, pp. 58-74, 2006.

[25] D. H. Wolpert i W. G. Macready, „No-Free-Lunch Theorem," *IEEE Transactions on Evolutionary Computation,* tom 1, nr 1, 1997.

[26] G. Aurélien, Hands-On Machine Learning with Scikit-Learn, Sebastopol: O'Reilly Media, Inc., 2017.

[27] S. Raschka, „A Tour of Machine Learning Classifiers Using Scikit-learn," w *Python Machine Learning*, Birmingham , Packt Publishing Ltd., 2015, pp. 90-92.

[28] I. N. d. Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni i S. F. d. R. Alves, „Multilayer Perceptron Network," w *Artificial Neural Networks: A Practical Course*, Springer, 2017, pp. 55-78.

[29] F. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[30] J. Heaton, Introduction to Neural Networks with Java, St. Louis: Heaton Research, Inc, 2008.

[31] S. Makridakis, „Accuracy measures: theoretical and practical concerns,” *International Journal of Forecasting,* tom 9, nr 4, pp. 527-529, 1993.

[32] S. Bermejo i J. Cabestany, „Oriented principal component analysis for large margin classifiers,” *Neural Networks,* tom 14, nr 10, pp. 1447-1461, 2001.

[33] R. G. Pontius, O. Thontteh i H. Chen, „Components of information for multiple resolution comparison between maps that share a real variable,” *Environmental and Ecological Statistics,* tom 15, nr 2, pp. 111-142, 2008.

[34] Y. Oualmakran, J. M. Espeche, M. Sisinni, T. Messervey i Z. Lennard, „Residential Electricity Tariffs in Europe: Current Situation, Evolution and Impact on Residential Flexibility Markets,” w *Sustainable Places 2017*, Middlesbrough, 2017.

Annex I

Table 11. Results of parameters optimization for dataset with directly measured appliances consumptions. Yellow area indicates the best parameters sets.

		Basic conf,		Random Search (RS)	RS result		Grid Search					
							Round 1	R1 results		Round 2	R2 results	
		train	test	Random Grid	train	test		train	test		train	test
RF	a	10		[20, 200, 380, 560, 740, 920, 1100, 1280, 1460, 1640, 1820, 2000]	380		[200, 400, 600, 1000]	600		[500, 600, 800]	500	
RF	b	2		[2, 5, 10]	2		[2, 3]	2		[3, 5]	3	
RF	c	1		[1, 2, 4]	1		[10, 20]	10		[10, 20]	10	
RF	MAPE	12.6%	32.5%	x	11.6%	31.1%	x	25.9%	33.1%	x	25.9%	32.9%
RF	MAE	62.1	168.1	x	55.9	158.1	x	122.8	164.4	x	122.8	164.4
RF	MSE	16597	109149	x	11906	94465	x	61094	106698	x	61173	107361
RF	R^2	0.95	0.7	x	0.97	0.74	x	0.83	0.71	x	0.83	0.71
MLP	d	relu		['tanh', 'relu'],	relu		relu	relu		relu	relu	
MLP	e	0,0001		[0,001, 0,0001, 1e-05]	1,00E-05		[0,0001, 0,05]	0,0001		[0,0001, 0,00001]	0,00001	
MLP	(f,g)	(100,1)		[(50,), (100,), (50, 2), (100, 2)]	(100,2)		[(50,2), (100,2), (150,2)]	(100,2)		[(100,2), (200,2), (150,2)]	(150,20)	
MLP	h	200		[200, 300]	300		[200,400]	400		[400,600]	400	
MLP	MAPE	46.8%	50.8%	x	47.7%	51.5%	x	34.5%	40.5%	x	33.4%	39.6%
MLP	MAE	202.2	204.9	x	197.9	204.1	x	157.0	189.3	x	153.0	185.2
MLP	MSE	132557	132168	x	125295	131681	x	77457	114791	x	73862	112612
MLP	R^2	0.63	0.64	x	0.65	0.64	x	0.78	0.68	x	0.79	0.69
LSTM	i	1		[10,50,100]	100		[80, 100, 150]	150		[180, 250]	250	
LSTM	j	10		[100,500]	100		[100, 150]	100		[80, 100]	100	
LSTM	MAPE	71.0%	64.8%	x	12.7%	44.0%	x	28.0%	38.1%	x	26.3%	35.3%
LSTM	MAE	299.0	326.9	x	48.1	230.1	x	164	211	x	152.0	203.7
LSTM	MSE	249823	234325	x	5315	139458	x	90772	131792	x	9709	120438
LSTM	R^2	0.33	0.24	x	0.99	0.55	x	0.72	0.57	x	0.86	0.73

Table 12. Results of parameters optimization for dataset with NILM disaggregation. Yellow area indicates the best parameters sets.

		Basic conf,		Random Search (RS)	RS result		Grid Search					
							Round 1	R1 results		Round 2	R2 results	
		train	test	Random Grid	train	test		train	test		train	test
RF	a	10		[20, 200, 380, 560, 740, 920, 1100, 1280, 1460, 1640, 1820, 2000]	1820		[600, 1000, 1700, 1900]	600		[550, 650, 750]	750	
RF	b	2		[2, 5, 10]	2		[2, 3]	2		[2, 3]	2	
RF	c	1		[1, 2, 4]	1		[10, 20]	10		[7,10]	7	
RF	MAPE	13.9%	34.4%	X	12.6%	33.4%	x	28.43%	35.11%	x	25.65%	34.37%
RF	MAE	72.7	175.9	X	63.8	166.0	X	142.7	175.1	X	129.42	172.29
RF	MSE	20371	100166	X	13439	92900	X	69048.65	101191.19	X	58133.26	98579.68
RF	R^2	0.95	0.76	X	0.97	0.78	X	0.83	0.76	x	0.86	0.76
MLP	d	relu		['tanh', 'relu'],	relu		relu	relu		relu	relu	
MLP	e	0,0001		[0,001, 0,0001, 1e-05]	1,00E-05		[0,0001, 0,05]	0,05		[0,05, 0,1]	0,1	
MLP	(f,g)	(100,1)		[(50,), (100,), (50, 2), (100, 2)]	(50,2)		[(50,2), (100,2), (150,2)]	(100, 2)		[(100,2), (200,2), (180,2)]	(200,2)	
MLP	h	200		[200, 300]	300		[200,400]	400		[400, 600]	400	
MLP	MAPE	51.0%	48.7%	X	50.9%	48.6%	X	38.70%	41.37%	X	37.12%	42.33%
MLP	MAE	233.4	218.6	X	227.4	213.5	X	178.48	191.65	X	161.33	193.6
MLP	MSE	157873	128944	X	148490	122686	X	88981.5	99021.16	X	67747.18	94130.81
MLP	R^2	0.62	0.68	X	0.64	0.69	X	0.79	0.75	x	0.84	0.76
LSTM	i	1		[10,50,100]	100		[80, 100, 150]	150		[250, 350]	350	
LSTM	j	10		[100,500]	100		[100, 150]	100		[80, 100]	80	
LSTM	MAPE	62.7%	95.0%	X	12.6%	53.2%	X	43.7%	49.98%	X	8.92%	47.38%
LSTM	MAE	289.0	383.6	X	44.7	287.4	X	194.9	274.21	X	34.96	240.63
LSTM	MSE	241506	265837	X	3937	208003	X	162948	189086.99	X	3336.89	145719.26
LSTM	R^2	0.43	0.27	x	0.99	0.43	X	0.78	0.48	x	0.99	0.6

Table 13. Results of parameters optimization for dataset with only total consumption. Yellow area indicates the best parameters sets.

		Basic conf,		Random Search (RS)	RS result		Grid Search					
							Round 1	R1 results		Round 2	R2 results	
		train	test	Random Grid	train	test		train	test		train	test
RF	a	10		[20, 200, 380, 560, 740, 920, 1100, 1280, 1460, 1640, 1820, 2000]	1820		[600, 1000, 1700, 1900]	1700		[1750, 1800, 1850]	1850	
RF	b	2		[2, 5, 10]	2		[2, 3]	2		[2, 3]	2	
RF	c	1		[1, 2, 4]	1		[10, 20]	10		[7,10]	7	
RF	MAPE	12.78%	32.10%	x	11.34%	30.78%	x	25.74%	31.68%	X	23.16%	31.17%
RF	MAE	61.07	162.65	x	53.34	153.20	x	123.7	161.1	X	111.41	157.92
RF	MSE	14484.56	90232.00	x	9585.79	81356.65	x	53595.69	86392.72	X	44117.72	82932.92
RF	R^2	0.94	0.66	x	0.96	0.70	x	0.77	0.67	X	0.81	0.69
MLP	d	relu		['tanh', 'relu'],	relu		relu	relu		Relu	relu	
MLP	e	0,0001		[0,001, 0,0001, 1e-05]	0.001		[0.0001, 0.05]	0.05		[0.05, 0.1]	0.05	
MLP	(f,g)	(100,1)		[(50,), (100,), (50, 2), (100, 2)]	(100,2)		[(50,2), (100,2), (150,2)]	(100,2)		[(100,2), (200,2), (180,2)]	(200, 2)	
MLP	h	200		[200, 300]	300		[200,400]	400		[400, 600]	600	
MLP	MAPE	46.06%	45.06%	X	42.76%	41.84%	x	38.71%	40.06%	X	34.08%	38.72%
MLP	MAE	195.59	191.92	x	187.31	183.67	x	165.58	170.45	X	146.84	167.48
MLP	MSE	113943.37	112524.49	x	107143.93	105746.69	x	76107.46	83821.83	X	58961.95	80341.82
MLP	R^2	0.52	0.54	x	0.55	0.56	x	0.68	0.65	X	0.75	0.67
LSTM	i	1		[10,50,100]	10		[5, 10, 20]	20		[20, 30]	20	
LSTM	j	10		[100,500]	100		[50, 150]	50		[20, 50]	30	
LSTM	MAPE	102.12%	82.03%	x	33.45%	39.89%	x	39.60%	39.53%	X	40.77%	40.33%
LSTM	MAE	347.89	299.44	x	142.07	198.83	x	164.05	186.63	X	167.81	186.16
LSTM	MSE	277282.39	146674.73	x	57246.10	89162.02	x	77873.49	78347.26	X	80565.84	77711.29
LSTM	R^2	-0.04	-0.11	x	0.79	0.33	x	0.71	0.41	x	0.7	0.41