TÉCNICO LISBOA

Data-driven Prediction of Optimal Operating Strategies for Residential Energy Systems

Ladislava Lipinová

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisors: Prof. Paulo Jose da Costa Branco

Dipl.-Ing. Ganzhou Wang

Examination Committee:

Chairperson: Prof. Duarte de Mesquita e Sousa

Supervisor: Prof. Paulo Jose da Costa Branco

Member of the Committee: Prof. João Filipe Pereira Fernandes

November 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.

Acknowledgements

I would like to express my gratitude to everyone who made this thesis work possible. Firstly, I would like to thank to my supervisor Mr. Ganzhou Wang for his guidance, suggestions and patience; and to Robert Bosch GmbH, Germany for hosting the topic of my thesis.

I would like to thank for the support and motivation to professor Paulo Jose da Costa Branco from Department of Electrical and Computer Engineering at Técnico Lisboa.

I would also like express my appreciation to InnoEnergy Master School for giving this opportunity to study this double master program and to both institutions Karlsruhe Institute of Technology and Instituto Superior Técnico.

Last but not least I would like to appreciate my friends and family for giving unlimited support and motivation throughout the two years of master studies.

Abstract

Future residential energy systems composed by decentralized roof-PV generation, electrification of heating, and storage installation aim to offer flexibility at the price of increased interdependencies and higher complexity to determine optimal operation conditions. An optimization-based algorithm such as model predictive control, accounts these multiple factors, however, it requires complex implementation and intensive computational effort. Concurrently, the detection of interdependencies within multiple variables has improved due to recent expansion of machine learning application. This sheds a light on designing and implementing smart controllers with the help of data analytics.

The aim of this work is to determine whether selected machine learning algorithms (ML) could detect a relation between multiple factors in residential energy system and thus improve prediction of an optimal operation of a bivalent heat pump system. Four system configurations are selected. Optimal operating profiles are obtained from an energy-planning tool that is developed at Bosch. Inputs of the tool are besides economical boundary conditions and technical specifications: demand profiles, weather and irradiation conditions.

Impacts of PV installation, thermal and battery storage are analysed. Three machine learning algorithms: random forest regression (RF), neural network (NN) and long short-term memory network (LSTM) are tested to analyse which features could improve the accuracy of prediction.

The results demonstrate capability as well as potential of implementing ML algorithms for residential energy management. The analysis also shows that RF and NN outperform LSTM on selected data. Additionally, these algorithms predict the operation of the system with average accuracy of R2 0.93 for all configurations.

Keywords: Energy Management, Hybrid Heat Pump, Machine Learning, Prediction of operation,

Resumo

Os futuros sistemas de energia residencial com produção descentralizada, paineis fotovoltaicos no telhado, eletrificação do aquecimento e armazenamento de energia, oferecem maior flexibilidade quanto ao preço dos vetores de energia usados mas maior complexidade para determinar as condições ideais de operação de cada um. Um algoritmo de otimização como controlo preditivo baseado em modelos usa esses múltiplos fatores. Entretanto requer implementação complexa e intenso esforço computacional. Simultaneamente, a detecção de interdependências entre variáveis mellhorou nos últimos anos devido ao aumento de algoritmos de aprendizagem.

O objetivo desta tese consiste no uso de algoritmos de aprendizagem na deteção das relações entre fatores determinantes do sistema de energia residencial e assim melhorar a operação otimizada de um sistema de bombeamento de calor tipo bivalente. Os perfis operacionais ideais vieram de um programa de planejamento energético da Bosch. Como entradas do programa, além das condições econômicas e das especificações técnicas, há perfis de demanda, condições climáticas e de irradiação.

Analizam-se os impactos da instalação fotovoltaica, armazenamento térmico e bateria. Três algoritmos são empregues para analisar quais recursos podem melhorar a precisão da previsão: regressão aleatória da floresta (RF), rede neural (NN) e rede de memória de curto prazo (LSTM).

Os resultados demonstram o potencial de implementação de algoritmos de aprendizagem para gerenciamento de energia residencial. A análise mostra que os algoritmos RF e NN superam LSTM em dados selecionados. Além disso, eles prevêem a operação do sistema com precisão média de R2 0,93 para todas as configurações.

Palavras-chave: Bomba de Calor Híbrida, Predição de operação, Gestão de Energia, Algoritmos de aprendizagem

Table of Contents

Acknowledgements V
Abstract VII
Resumo VIII
List of figures XI
List of tables XII
List of Abbreviations XIII
1. Introduction 1
1.1 Motivation 1
1.2 Objective 3
1.3 Contribution of This Work 3
1.4 Thesis Outline 4
2. Fundamentals 5
2.1 Heat Pump 5
2.2 Control of Heat Pump Systems 6
2.2.1 Base Control 6
2.2.2 Energy Management Strategies 7
2.3 SystemFinder 8
2.4 Machine Learning 11
2.4.1 Correlation Analysis 11
2.4.2 Supervised machine learning 12
2.4.3 Parameter Tuning 14
2.4.4 Feature Scaling 17
2.4.5 Model Evaluation Metrics of Regression Problems 17
3. Literature review 19
4. Methodology and Program Development 22
4.1 Methodology 22
4.2 Program Development 24
4.3 Description of Configurations 27
4.4 Analysis of Configuration 28
4.4.1 Configuration 1 28

4.4.2 Configuration 2 33
4.4.3 Configuration 3 34
4.4.4 Configuration 4 35
4.5 Data Preparation 37
4.6 Parameter Optimization 38
5. Results 40
6. Conclusion 48
References 49
Annex 53

List of figures

Figure 1 The bivalent operation of a hybrid heat pump system: working principle (left) and system illustration (right) [36]. 2
Figure 2 The scheme and operational diagram of a complex residential energy management [36]. 2
Figure 3 The schematic explanation of the objective of the thesis [36]. 3
Figure 4 Overview of the SystemFinder modelling approach [22]. 9
Figure 5 The scheme of the model System Finder [25] 10
Figure 6 Illustration of working principle of random forest [42]. 12
Figure 7 Illustration of working principle of artificial neural network [43]. 13
Figure 8 Illustration of the basic concept of LSTM network [15]. 14
Figure 9 Distribution of ambient temperature for the input profile 24
Figure 10 Annual distribution of ambient temperature and space heating demand 25
Figure 11 Daily hot water demand profile. 25
Figure 12 Mean hourly electricity demand of household. 26
Figure 13 Average monthly irradiation of the dataset. 26
Figure 14 Heating curve function of the house. 29
Figure 15 Occurrence of operational hours of each generation technology. 31
Figure 16 Distribution of input power of the heat pump for price ratios 2,5; 3,5; and 5. Maximum input power equals to 1 and minimum equals to 0. 31
Figure 17 Priority chart of PV generation flow 33
Figure 18 Operation of heat pump for configuration with and without PV installation with respect to PV generation. 34
Figure 19 Annual distribution of space heating demand [kW] and thermal storage discharge [kW]. Hourly range between 3624-5832 corresponds to months June-August. 36
Figure 20 Level of thermal storage [kW] and space heating demand [kW] annual distribution. 36
Figure 21 Result evaluation of Mean Absolute Error (MAE) for each model at each configuration. 42
Figure 22 Result evaluation of Mean Square Error (MSE) for each model at each configuration. 42
Figure 23 Result evaluation of coefficient of determination (R^2) for each model at each configuration. 42
Figure 24 Comparison of results of each algorithm for each configuration 43
Figure 25 Random Forrest Regression Prediction Results. Configuration 2 44
Figure 26 Neural Network Regression Prediction Results Configuration 2. 45
Figure 27 Long Short-term memory network training and testing results configuration 2. 46

List of tables

Table 1 Overview of possible control method for the heat pump system. Adapted from [9]. 6
Table 2 Dimensioning of the generation and storage systems. 27
Table 3 Configuration setup. 27
Table 4 Correlation matrix for configuration 1. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating 28
Table 5 The operation of hybrid system with respect to ambient temperature based on its mode. From the top left to right: Gas Boiler mode operation for electricity/gas price ratio 2.5; Gas Boiler mode operation for electricity/gas price ratio 3; 30
Table 6 Correlation table for configuration 2: T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating, D_hw: Demand hot water, PV_total: Photovoltaics generation, Storage_hw: Hot water storage level 33
Table 7 Correlation table for configuration 3. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_hw: Demand hot water, D_sh: Demand space heating, Storage_hw: Hot water storage level, PV_total: Photovoltaics generation, Bat_dischar: Battery discharging. 35
Table 8 Correlation table for configuration 4. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating, PV_total: Photovoltaics generation, PS_level: Thermal storage level, PS_char: Thermal storage charging, PS_dischar: Thermal storage discharging. 36
Table 9 Operating modes table of features 37
Table 10 Parameters for grid search – Neural Network algorithm. 38
Table 11 Parameters for grid search – Random Forest 39
Table 12 Parameters for grid search – LSTM algorithm 39
Table 13 Optimized parameters for algorithms development. 39
Table 14 The complete evaluations of all algorithms and models for both datasets. 53
Table 15 The list of variables generated from the SystemFinder model. 54
Table 16 List of features for each model of Configuration 1. 55
Table 17 List of features for each model of Configuration 2. 56
Table 18 List of features for each model of Configuration 3. 57
Table 19 List of features for each model of Configuration 4. 58

List of Abbreviations

ANN	Artificial Neural Network
COP	Coefficient of Performance
CNN	Convolutional Network
EU	European Union
LSTM	Long short-term Neural Network
MAE	Mean absolute error
MAPE	Mean absolute percentage error
ML	Machine Learning
MLP	Multilayer Perceptron
MPC	Model Predictive Control
MPD	Markov Decision Process
MSE	Mean Square Error
NN	Neural Network
PDC	Predictive Demand Control Algorithm
PLS	Partial Least Square
PV	Photovoltaics
RBC	Rule Based Controller
RF	Random Forrest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SCOP	Seasonal Coefficient of Performance
SVR	Support Vector Machine Regression
RT	Regression Tree
TDNN	Time Delay Neural Network

1. Introduction

1.1 Motivation

The strategic long-term vision of European countries is a competitive and climate-neutral economy by year 2050. That leads to strong emphasis on the reduction of the CO_2 emission, reduction in energy consumption and efficient energy management. According to data provided by Eurostat, the building sector was responsible for 27.2% of final energy consumption in European Union in 2017. Most of the residential energy consumption was used for space heating (64.1%) and water heating (14.8 %). Additionally, most of the space heating and water heating consumption was covered by gas (up to 42 % and 47 % respectively) [1]. These facts bring a lot of potential for energy savings and CO_2 reduction and, they have also been considered in European legislation. According to the European directive 2010/31/EU [2] on energy performance of buildings, all new buildings built by 2021 are encouraged to be zero net energy buildings. Electrification of the heating sector presents itself as an essential part to meet that goal and to progress in the overall effort to decarbonize energy supply replacing the fossil fuel-based heating systems. The heat pumps are considered energy-efficient and low-carbon alternative to conventional heating systems. The advantage of the heat pumps is that their efficiency is much higher comparing to conventional electrical resistance heaters. That is as heat pump transfers heat from external environment and only fraction of energy is needed to run the device. The transferred heat can be up to 3-4 times higher than the electric input power of the heat pump. That makes the efficiency of the system, which is defined as coefficient of performance (COP), 3-4 larger comparing to COP of 1 of the electrical resistance heaters. Additionally, the EU Directive on energy efficiency considers heat pump as a renewable heat source [3].

However, transition to electrification of heating would cause steep growth of electricity demand resulting in increase in conventional electricity production and risks for electrical grids. As a transitional solution to tackle these challenges hybrid systems could be considered [4]. Hybrid systems are composed of an electrically driven heat pump paired with a fossil-based heat generator such as gas boiler. Both components can be operated simultaneously or in a serial operation [5]. Connecting heat pump into hybrid system enables reduction in sizing of the heat pump in order to cover partial load of the total maximum thermal load required. Thus, the seasonal coefficient of operation (SCOP) increases as the heat pump can operate during the heating season with higher load factors, reducing the annual cycling losses [6]. Hybrid systems are also appealing from the economical aspect as they decrease the investment cost and can flexibly adjust to current fuel costs. Additionally, the hybrid system could be a solution for implementation of heat pumps into old buildings which require high temperature at the inlet of the heating system.

Operation of the Hybrid system

Currently the operation of the hybrid system is dictated by the ambient temperature. When the ambient temperature drops under certain determined value (bivalent point) heat pump is not able to cover the whole load on its own. Thus, heat pump starts to operate in parallel with the secondary heat generation technology. That is due to decreasing COP of the heat pump with decreasing temperature and presumable increase in heating demand. When the ambient temperature and consequently COP are too low for heat pump to operate then gas boiler provides all the heating supply. The bivalent operation of the hybrid system with respect to both ambient temperature and load is depicted on Figure 1.

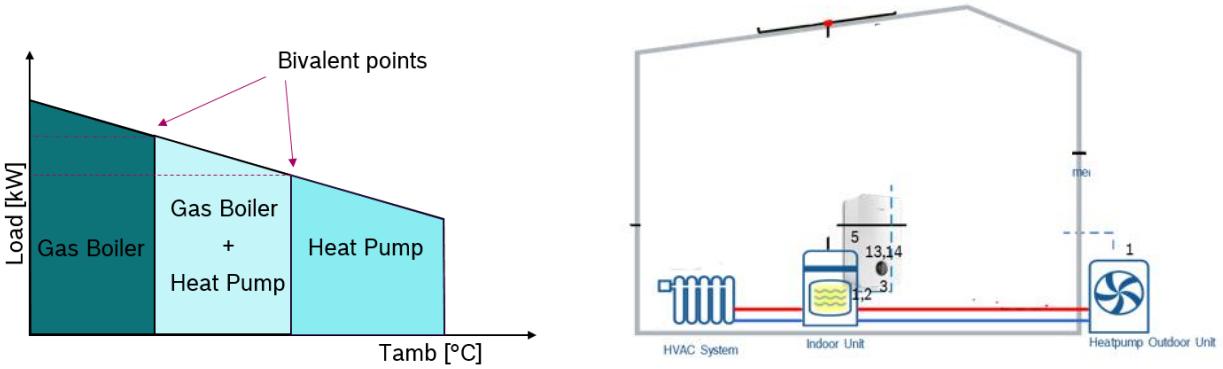


Figure 1 The bivalent operation of a hybrid heat pump system: working principle (left) and system illustration (right) [36].

On the other hand, future residential energy systems will present much larger complexity caused by several trends. First is the decentralization of energy generation. Second, the energy market flexibility will dynamically shift loads based on the current energy generation. The operation of such complex system depends on multiple factors such as generating technology implemented in the house, available storage technologies, price tariffs, demand profiles, weather conditions, unpredictability of the renewable sources etc. These technologies might shift the optimal operation of bivalent point from being ambient temperature dependent to dependency on multiple variables. All these factors might cause that a complex prediction model for residential energy management will be necessary to consider, as illustrated in Figure 2.

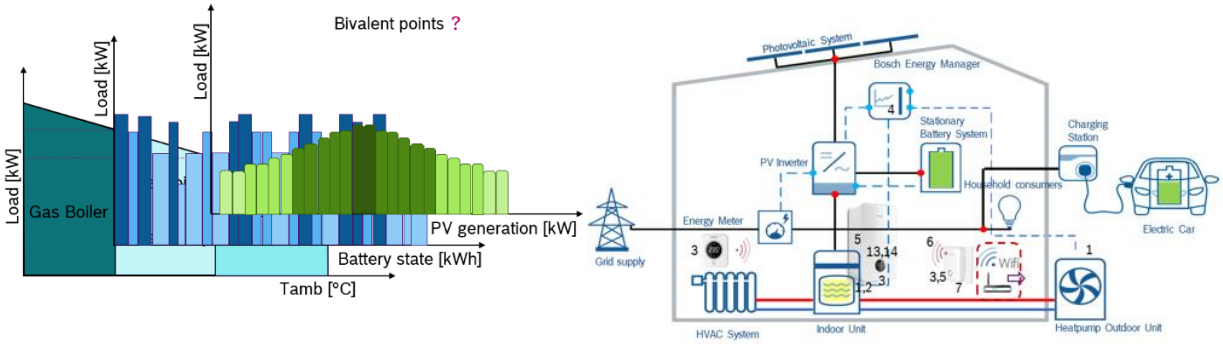


Figure 2 The scheme and operational diagram of a complex residential energy management [36].

1.2 Objective

The objective of this work is to investigate if the operation of the hybrid heat pump system can be improved using machine learning algorithms which are trained with data obtained from simulation model. The simulation generates the optimal operation of the system by calculating the energy balance at the time step nodes having all relevant information in defined time frame including past, current and future time steps. The thesis investigates if the ML algorithms can capture the relations between the variables and learn the optimal operation of the system in order to be able to predict the operation of the system in next time step with the data available at current and past steps. The illustration of the objective is presented on Figure 3.

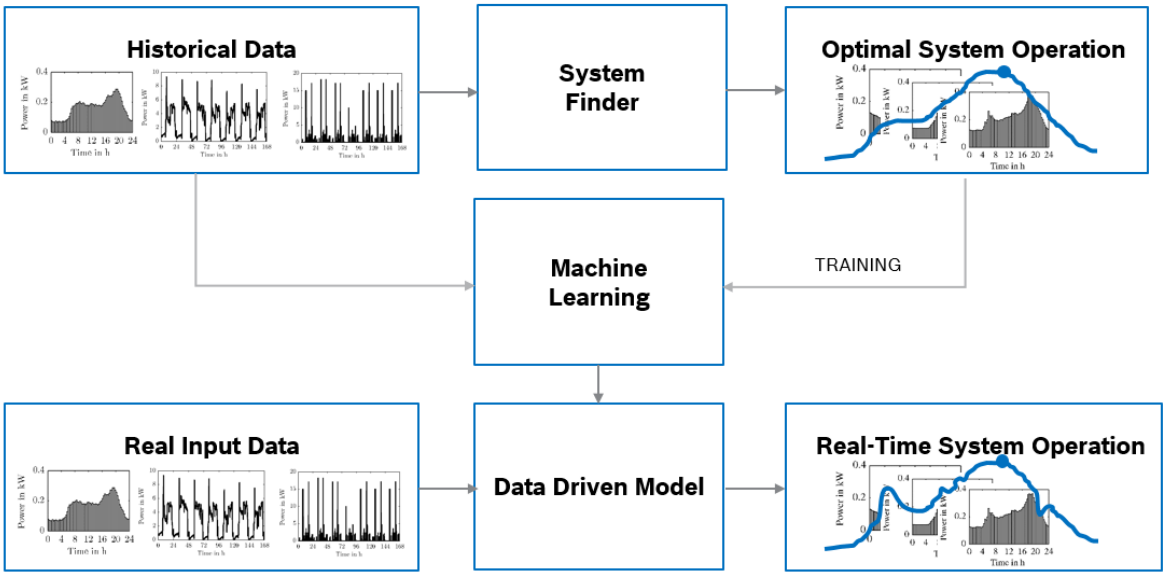


Figure 3 The schematic explanation of the objective of the thesis [36].

1.3 Contribution of This Work

The main purpose of this work can be divided into two parts:

I. First one is to analyze how various factors influence the optimal operation of the hybrid heat pump system. The factors that are considered include the price of electricity, installation of photovoltaics panel, battery storage and thermal storage.

II. Secondly, the work investigates the potential implementation of machine learning algorithms into the control of hybrid generation system operation in relation to energy management of the house. The aim is to select appropriate machine learning algorithms, analyze which algorithm would be the most appropriate for this purpose and what features are relevant for the performance of the algorithms.

1.4 Thesis Outline

The structure of this document is organized into six chapters as follow:

Chapter 1 Introduction – The first chapter gives an overview of the problem researched in this paper and states the motivation behind the examined matter. The hybrid heat pump and its position in residential energy management are introduced.

Chapter 2 Fundamentals – describes theoretical background necessary in order to obtain better understanding of the discussed topic. It presents current state of heat pump control, describes SystemFinder model which is used as a source for data generation and essential theory for building up a machine learning algorithm.

Chapter 3 Literature review – discusses scientific articles which focus on the topic of heat pump operation and prediction of hybrid system using machine learning algorithms. The analysis is used in order to determine the methodology and select machine learning algorithms.

Chapter 4 Methodology and Program Development – describes the general procedure and reasoning leading to model creation and evaluation. Further it describes the input profiles used for data generations necessary to understand in order to get better understanding of the optimal operation. Explanatory analysis on data is performed, algorithm development and parameter optimization are carried.

Chapter 5 Results – presents results obtained from the algorithms, evaluates each machine learning algorithm and discusses the outcomes.

Chapter 6 Conclusion – summarizes the results obtained in previous chapters and gives recommendations for future research.

2. Fundamentals

2.1 Heat Pump

Heat pumps is a form heat engine that uses mechanical work to transfer heat from a low temperature source to a higher temperature sink. The heat transfer occurs by extracting the heat from the external environment (source) and providing that heating to the building (sink). The process reserves natural heat transfer using refrigerant [7].

The efficiency of the heat pump operation is denoted as coefficient of performance (COP).

Carnot Efficiency

Carnot efficiency is defined as a coefficient of performance of a reversed Carnot cycle. It is theoretical maximum efficiency limited by the absolute temperature of the sink (T_{sink}) and source (T_{source}). This implies that the efficiency of the heat pump is higher when the temperatures of the sink and source are closer [7]:

$$COP_{Carnot} = \frac{T_{sink}}{T_{sink} - T_{source}} \quad (1)$$

Theoretical COP

COP_{carnot} indicates the maximum theoretical potential performance of the system given by the environments, however in real operation such performance will never be achieved. The operational performance corresponds better to theoretical COP_{th} which takes into account efficiency of the system [7]:

$$COP_{th} = \eta \cdot COP_{Carnot} \quad (2)$$

For which is η evaluation of heat pump system performance. it is related to particular heat pump model and it is given by the manufacturer. It is dependent on modulation of the compressor, losses in the hydraulic system etc.

Coefficient of Performance

COP can also be defined as a ratio between quantity of the heat delivered by the heat pump and the electricity input power of the heat pump [7]:

$$COP = \frac{Q_{HP}}{W_{HP}} \quad (3)$$

Seasonal Coefficient of Performance (SCOP)

Seasonal Coefficient of Performance is defined as an average COP during the heating season.

2.2 Control of Heat Pump Systems

Control of system is defined as method of operating device in a consistent, economical and safe way which could not be achieved by human manual control. [8]

2.2.1 Base Control

Control of the heat pump system consist of the decision and action agent. The decision agent is an operational system, which analyses the current state of the system and implements decision-making process. Action agent implements the required change on the device. In case of the heat pump, the action agent is the operational speed of compressor [8].

The control of the heat pump system is implemented by the operation of the compressor. There are two approaches to operate the compressor:

Table 1 Overview of possible control method for the heat pump system. Adapted from [9].

Type	Pros	Cons
Rule-Based	Easy to implement Low computational demand Robustness of the system No external information needed Low economic price	Inflexible operation based on rules Expert knowledge needed for proper operation Does not consider the future behavior of the system
Model Free Predictive Control	Uses information from predictions Low computational demand Compromises between complexity and performance Better performance comparing RBC	Prior expert knowledge needed
Model Based Predictive Control	Uses information from predictions Optimal or close to optimal solutions Enables handling constrains Better performance comparing RBC	Complex in design Model of the physical system required Modelling errors Uncertainty of the prediction not considered Computationally demanding
Stochastic Predictive Control	Considers the uncertainty of the prediction Robustness of the system Better performance comparing to non-stochastic MPC	Complex in design Computationally demanding

On/Off compressor operation – the operation of compressor is driven by the on and off state of the constant speed motor, any adjustment in change of the heat load can be only done by intermitting of the motor. [8]

Inverter adjusts variable compressor speed **–** the speed of the compressor motor in order to meet the demanded heat load [8]**.**

2.2.2 Energy Management Strategies

There are several strategies which are used for control of the operation system which can be divided into two groups - non-predictive methods and predictive methods. The basic overview of the control method is listed in Table 1. In the following, a brief introduction of these methods is given.

Non-predictive methods

Rule Based Control (RBC) – it is the most widely used control for heat generation systems. The control principle is based on simple *if condition – then action* rule-based mechanism. Operation of most rule-based controllers for energy management is determined by the ambient temperature, which is used as set point for the operation of the heat pump or secondary generation source. Additionally, ambient temperature is also used to determine required supply temperature to heating distribution system – heating curve function. The drawback of the rule-based controller is that the action is determined by current state of the system. The current state is defined by the output of sensors in real time or its average along the determined past time period. There could be various sensors consider for the control of the system (ambient temperature, indoor temperature, temperature at the inlet and outlet at the space heating, temperature level of hot water storage, temperature level of thermal storage etc.). The limitation of the rule base operation is that it only takes into account the current state of the system and the response value (usually indoor temperature). However, it does not predict the possible future behavior of the system with respect to future heating demand, change of the heat inertia of building, occupancy, time of the day or irradiation. This can lead to overheating or overcooling of the dwelling and therefore it is not considered as the most economical way to operate energy management system. [8]

Certain controllers also consider the usage of PV generated electricity by giving a fixed order for example heat pump runs only when PV production exceeds certain threshold, or when electricity is fed into grid. The limitation of such system also consists of operation based exclusively on current generation not considering the prediction of the PV generation which leads to similar disadvantages as mentioned for the temperature-based controller. [8]

Model Predictive Control

Model predictive control (MPC) is an advance method of control which can manage complex dynamic systems. The principle of the MPC is to optimize the current timeslot while considering the following future time slots. MPC can be divided based on the complexity of computation and input data into:

Model based predictive control is model based meaning that model of an actual physical system is designed and improved with the data obtained from the real measurement or the real system. Modelling approach which considers both the physical model of the system and measured data is known as gray box modelling. [8]

Model free predictive control is based on purely the data from the real system measurements which are used to determine the operation of the system. The physical model of the system is not implemented.

Stochastic Predictive Control is a model which is based on taking into account the uncertainty of the prediction of the model. [8]

The data provided for the MPC prediction can be provided entirely from the system (autonomous system) or from third parties. External values such as ambient temperature, irradiation or market price can be very accurately obtained from third party provider. Additionally, ambient temperature and irradiation are used for predicting the thermal demand of the building (heating curve) and PV generation and are determined individually for each household. The prediction of the energy demand for hot water and space heating demand is based on the internal historic data with pattern recognition. If data from the third parties are not available or the system aim to be autonomous various statistical methods are used such as auto regressive method, moving method, ANN, generalized mixture models and others. [8], [9]

2.3 SystemFinder

In order to obtain better understanding of the dataset it is necessary to explain basic modelling design and constrains of the simulation model. The data provided for this thesis was obtained using an optimization-based energy-planning tool called SystemFinder, which was developed by Bosch. SystemFinder was created in order to provide optimal sizing, configuration and operation of the system for the residential housing taking into consideration interdependencies between end-use domains of electricity, hot water and space heating. The optimization of the operation mode is calculated by meeting the end-use demand with combination of power flows of various generation units and storage technologies in each time step. The overview of the modelling approach scheme can be observed on Figure 5. The target function of the optimization is the minimization of total annualized discounted costs of an entire house energy system including the fixed, variable operational cost and investment cost.

The input information can be divided into three main groups – demand profiles, technology input and economic information as depicted on Figure 4. The detailed overview of model input used for the data generation for the purposes of this thesis are listed in chapter 5.1.

There are two main outputs that the system can provide with respect to demand profile and fixed and operational cost. First one is an optimal sizing of technologies and storages for a given house. Second is an optimal operation of the given system.

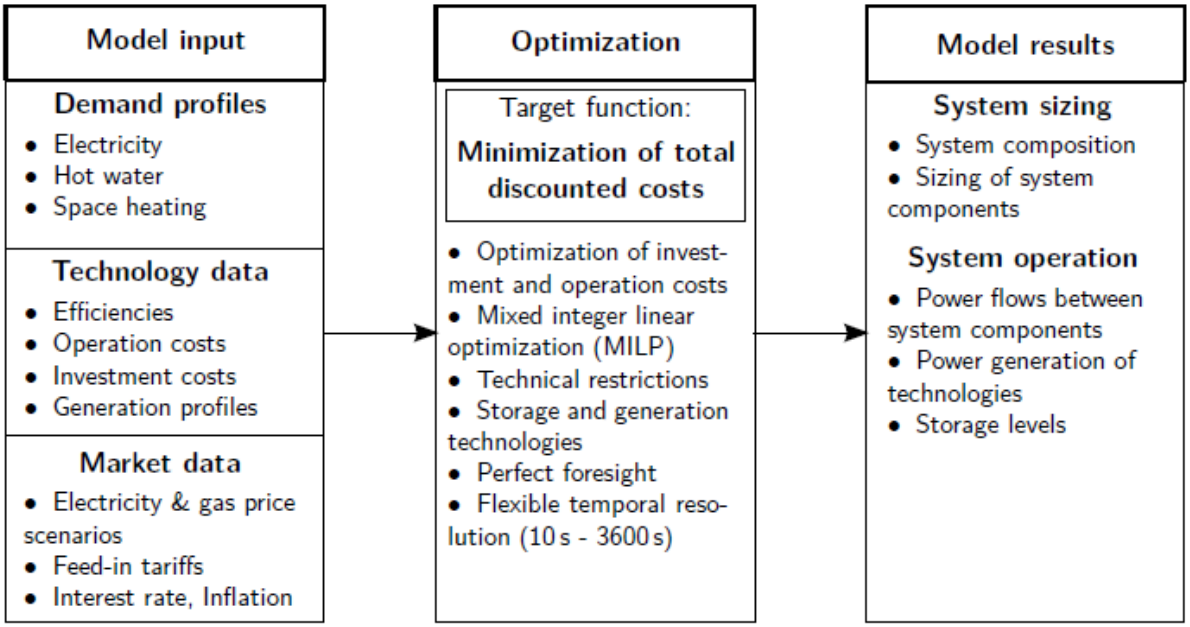


Figure 4 Overview of the SystemFinder modelling approach [22].

The specific constrains of modelling

Optimization time steps

There are two different time steps implemented in the system with the respect to different demand behavior. The optimization of the power flows are modeled in the step of 5 minutes whereas the time step for space heating demand is optimized in 60 minutes time step taking into consideration the inertia of the building and the heat transfer system. At each time step, the end-use demand (load) must be met.

Variable Flow Temperature

When modelling the temperature dependent demand profiles – (hot water demand and space heating demand) not only the load at each time step must be fulfilled but also the temperature demand.

When there is more than one source providing the heat supply simultaneously, it assumed that the all units provide the same temperature level for the simulation purposes of linear operation (parallel connection). However, in the real operational system it is more beneficial to use a serial connection of technologies in order to make use of the high COP of the lower flow temperature of the heat pump. Therefore, in order to make the model more corresponding to the actual physical state, discretization of the temperature levels is introduced for the technologies with non-linear behavior with respect to flow temperature.

The required heat demand is divided on three levels at each time step and consequently the required temperature at each level and COP of heat pump is calculated. Consequently, each generation technology can supply the heat up to the temperature level which is currently the most economical.

Indoor Temperature

SystemFinder does not take into consideration indoor temperature as it only calculates with energy flows. Therefore, if the final energy demand of space heating is met, it is considered as indoor temperature being achieved.

General Economic Parameters

The optimization of dimensioning of the technologies and investment will be not discussed further as there are not objective of this work. However, as the optimization of the system operation is dependent of the target function of the simulation it is necessary to state cost for investment and operation [Basic scenario].

Further assumptions

Comparing to data obtained from real system there are assumptions that are not necessary to be taken into account. Those would be delay of the measurement system and disturbances caused by weather and house occupancy.

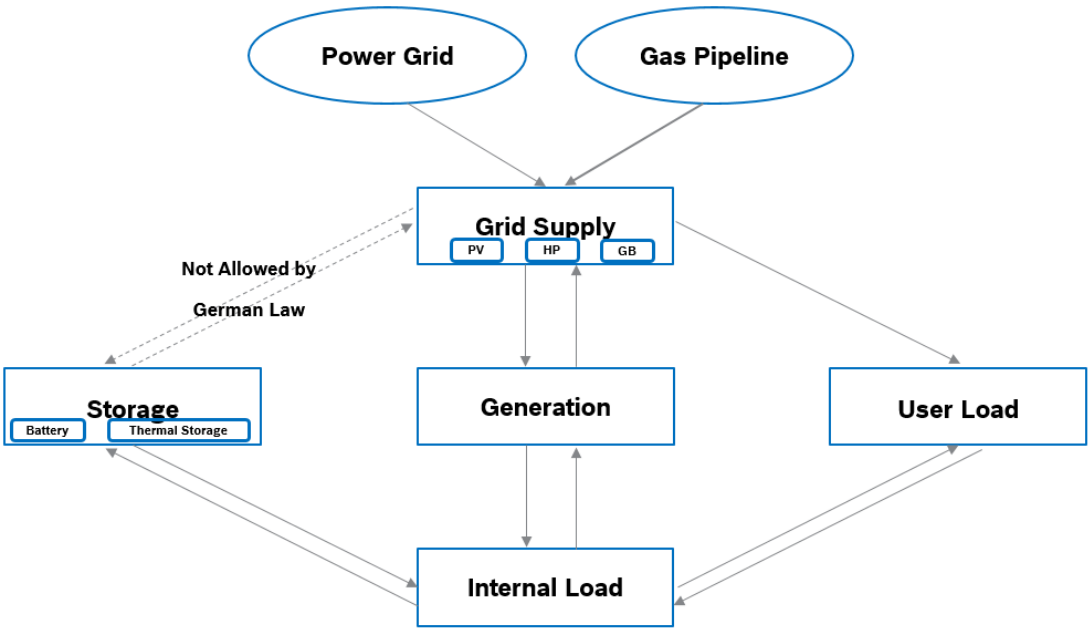


Figure 5 The scheme of the model System Finder [25]

2.4 Machine Learning

Machine learning (ML) is statistical algorithm method which can perform a specific task without being explicitly programmed by capturing patterns between the input variables during training period.

2.4.1 Correlation Analysis

Correlation analysis is an essential tool in order to determine the eventual relation between variables (features). It is valuable to learn the existence of relation between the features for several reasons. First, if two variables present close correlation, they can be predicted one from another. This can potentially eliminate the number of features and simplify the model. Second, certain algorithms might give misleading results for cases when several features possess perfect correlation. Additionally, the correlation analysis provides us with better perspective to understand the data. It ranges between - 1 to 1. Correlation acquires negative values when the relation between features is indirect, meaning that when one variable increases the other one decreases. Variables present no correlation if result value is approaching 0 [10].

Covariance

Covariance determines if two variables present linear relationship. However, it does not establish how strong the relationship is. Covariance is calculated as following [11]:

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)(Y_i - \overline{Y})}{n-1} \tag{4}$$

where X_i, Y_i present the sample X,Y in observation i respectively; $\overline{X}, \overline{Y}$ present mean of samples respectively and n represents number of samples.

Correlation

Correlation build on covariance principle and additionally determines how strong the linear relationship between two variables is [11]:

$$COR = \frac{COV\ (X,Y)}{\sigma_X * \sigma_Y} \tag{5}$$

where COV (X,Y) is the covariance calculated in Equation 4. and σ stands for standard deviation.

Pearson correlation

Pearson correlation indicates linear relation (strength and direction) between two continuous variables. It seeks a line that best fits between the data points of two features and consequently calculates the distance of each point to the line. The relationship is linear only when the change in one variable is proportional to the change in another variable [12].

$$r = \frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{\sqrt{[n\,\sum x^2 - (\sum x)^2]\,[n\,\sum y^2 - (\sum y)^2]}} \quad (6)$$

x,y represent the respective variables, and *n* stands for number of samples.

2.4.2 Supervised machine learning

Supervised machine learning is a task of machine learning that trains the interdependencies between inputs based on given output. In other words, the predicted output of the given inputs is known during the training phase and directly affects the training process [13].

Random Forest

Random forests are a type of ensemble learning algorithm in which each model (tree) is trained individually with different subset of dataset and different features. Comparing to other models each tree is trained individually and there is not weighing scheme in between the trees. The final prediction is an average of prediction of several independent base models - trees. The scheme of working principle of RF can be observed on Figure 6 [13]. The working principle can be explained as following. Firstly, *n* samples of dataset is used for training set for *k* trees. Secondly, decision tree is built on the randomly selected *d* features. Then, steps 1 and 2 are repeated *k* times. Final prediction is a result of overall average of trees output.

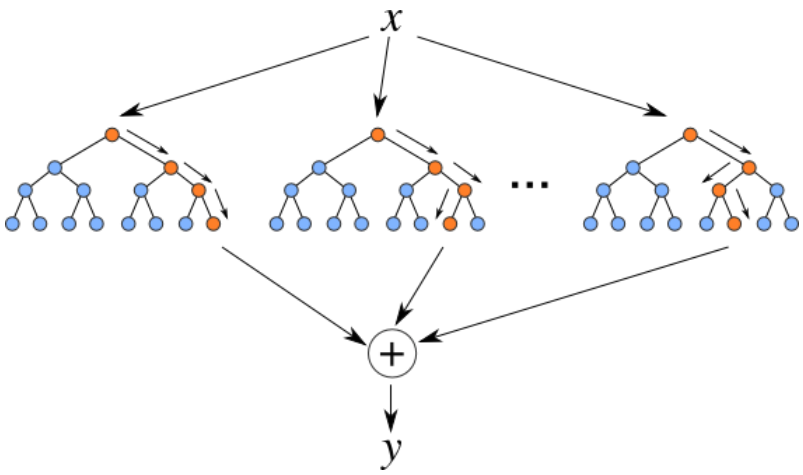


Figure 6 Illustration of working principle of random forest [42].

There are several advantages of using random forest model. Firstly, RF is very easy to implement. Secondly, decision trees are non-parametric meaning that it can model complex relations between inputs and outputs without any previous assumption [10]. Additionally, RF does not require complex data processing. It can work with both categorical and continuous data even combined in one dataset. Moreover, RF works with certain in build feature selection which eliminates noisy variables and therefore it makes RF robust to outliners.

Artificial Neural Network

The artificial neural networks are inspired by the functioning of biological neural system. Such neural system consists of connected basics units called neurons. Biological neurons work on a principle of receiving a signal from the network, processing it and eventually transmitting it forward through electrochemical signals [14].

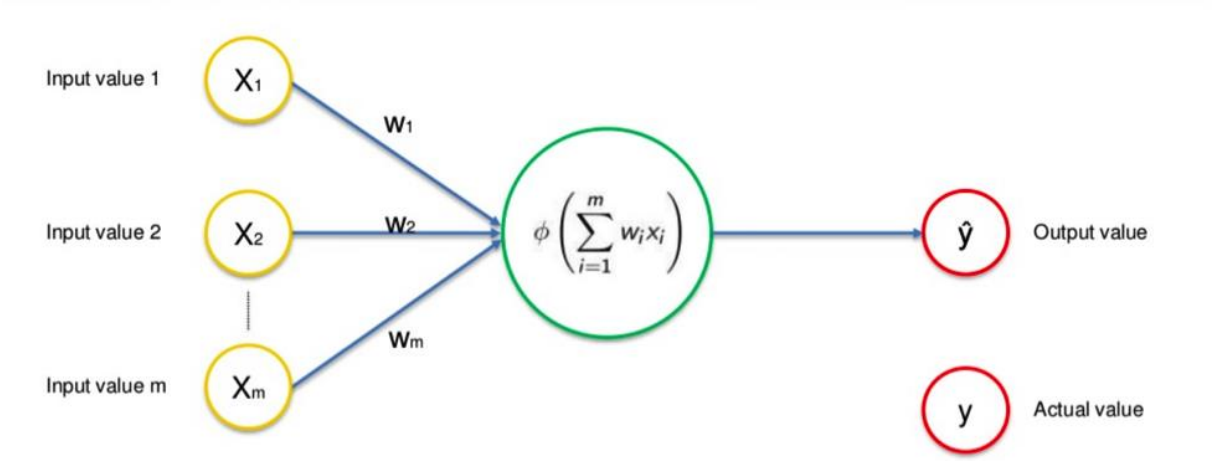


Figure 7 Illustration of working principle of artificial neural network [43].

The artificial neuron works on a similar principle. The network consists of signal inputs, which are fed into artificial neuron, processed in neuron and forwarded as an output value. The basic principle of artificial neural network is depicted on Figure 7. The independent input variables are processed by the weight function and feed into neuron layer. In the neuron layer all the weighted inputs are added up and activation function is applied. The processed value is forwarded as an output value ŷ. This process is called a forward propagation when the input variable is processed by the neuron resulting into output variable ŷ. This output variable is then compared with the actual value using the cost function. The aim of the optimization of the neural network is the minimization of the cost function which represents the error between the prediction of ANN and the actual value. The result from the cost function is then feed back into network and back propagated. During backpropagation the weighted functions are updated based on the output from the cost function. This process of forward and back propagation is repeated until the values of the weighted functions are optimized and the error is minimized [14].

Long Short-Term Memory Network

Long Short-Term Memory network (LSTM) is a type of recurrent neural network (RNN). RNN is a class of neural networks that allow previous outputs to be used as inputs while having hidden states. RNN might be affected by short memory while processing long sequence of data. Consequently, an important information might get lost while processing a longer sequence. Additionally, RNN might suffer from vanishing gradient problem during back propagation. That means that the function reaches local minimum and the network stops updating value of the weights. As a solution to this LSTM, have been created with

a system of gates which regulate the flow of information by selecting information based on its relevance [15] as illustrated in Figure 8.

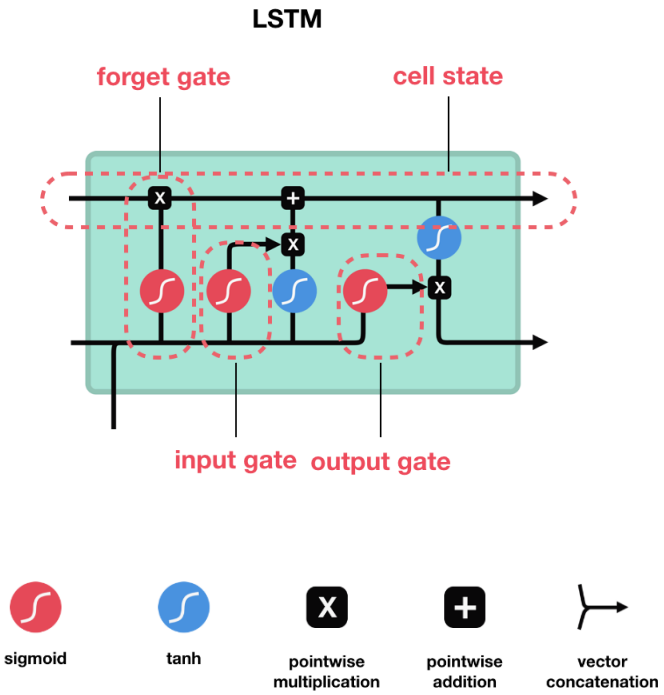


Figure 8 Illustration of the basic concept of LSTM network [15].

Tanh activation function – transform the vector values in range between -1 and 1.

Sigmoid activation function – is used to rescale the data between 0 and 1 (0 for data to be forgotten and 1 to be kept).

Input gate – is used to update the state of cell. Previous hidden state and current cell input are firstly separately processed by both the sigmoid function and tanh function. Then the outcomes of the functions are multiplied and based on its value, it is passed forward of forgotten.

Forget gate – processes the information from the previous hidden state and current cell input and determines if the value should be kept or discarded using the sigmoid function. In order words, it decides if the value from previous should be kept in network.

Cell state – is updated using the forget and input gate.

Output gate – determines the state for the next hidden state and new cell state using both tanh and sigmoid function. Hidden layers consist of hidden states and memory cells. The value of hidden state is passed into output layer whereas memory cell state stays internal.

2.4.3 Parameter Tuning

Parameter tuning is a method to determine the optimal parameters for the models training. It mainly depends on experimental results. Parameters are tuned by testing all possible combinations and selecting

those with the lowest training error. Evaluating the optimized parameter results exclusively only on training data of the model may lead to overfitting. This results in very good performance for the training data but performs poorly on testing data. Therefore, cross validation is introduced. K-cross validation divides the training dataset into number of groups. During the parameter optimization, each combination of parameters is trained on the number of groups of and tested on the data from other group. The testing and training groups for each parameter set is changing. The number of possible combinations increase exponentially with an increment of one hyperparameter. Therefore, certain parameters are preselected with respect to the solving task in order to minimize the number of possible combinations and decrease the optimization time [10].

Parameters for Random Forest [16]

Number of Trees (n_estimators) - number of trees in the forest, higher number of trees improves the accuracy however it slows down the predicting time.

Maximum Features (max_features) - number of features to consider when looking for best way of splitting a node.

Maximum Depth (max_depth) - maximum number of levels in each decision tree, the deeper tree is, more capable the model is in capturing more information about the data. However, overfitting may occur for too high number of depths

Minimum samples split (min_samples_split) - minimum number of data points required to split the node, increasing this parameter may lead to underfitting.

Minimum samples leaf (min_samples_leaf) - minimum number of data points allowed in a leaf node at the base of the tree, increasing this parameter may lead to underfitting.

Resampling data (*bootstrap*) - method for sampling data points (with or without replacement).

Parameters for Neural Network [17]

Activation Function **–** decides if the given neuron should be activated based on the weighted sum. There are number of activation functions, some improve the accuracy, others training time. The selection of activation function is dependent on the type of problem (classification or regression), data and type of layer (input, output, hidden).

Activation Functions for input and output layers for regression problems

Linear – one of the simplest functions. It is limited when it comes to capture complexity in parameters. However, it is still recommended for regression problems.

Hyperbolic Tangent – it is similar to commonly used sigmoid activation function. However, the advantage of this function is that it also considers negative values. It introduces non-linearity into data and smoothens the gradient preventing the values from jumping.

Activation Function for hidden layers

ReLu - non-linear activation function recommended for hidden layers as it quickly converges.

Optimizer **–** when algorithm is trained it is an iterative process with the aim to minimize (or maximize) an objective function, in order to that an optimizer:

Adam – an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The advantage of the Adam optimizer is quick convergent and efficient learning method.

Learning Rate **–** parameter which determines how much the model should change in response to estimated error each time weights of model are updated. Too small rate could lead to long computational times and too large rate could lead for model not to converge or to omit the optimal weights.

Batch size – number of samples passed in a single batch before the weight of the network is updated. Small batch size might increase the training time. On the other hand, too large batch size might reduce the ability of the algorithm to generalize its training.

Epochs – number of times the full dataset is passed through training, too high number of epochs can lead to overfitting and too low number of epochs leads to network not being properly trained.

Number of Neurons – it is a crucial parameter for neural network training, low number of neurons can lead to underfitting and opposite high number of neurons can lead to overfitting and too long training time.

Number of Neurons in input layer – it is common to use as an input number of neurons number of features.

Number of Neurons in hidden layers **–** there is no universal rule of thumb for the number of neurons in hidden layers, certain literatures recommended number of neurons between the ranges of 2/3 of input layer neurons to maximum double of number of input layer neurons. In order to find optimal number of neurons in hidden layers, grid search optimization was applied with variables 2/3 of input layer neurons, double of input layer neurons, and sum of the previous two mentioned.

Number of Neurons in output layer - is equal to number of output parameters.

Number of Hidden Layers – it is common to use up to 2 hidden layers. That is as two hidden layers show to be good enough to cover most complex relations between features, more layers might improve the model but also makes the model harder to train. Generally, the number of layers can be selected accordingly:

0 number of hidden layers - for linear problems

1 number of hidden layers - for simple problems

2 number of hidden layers - for complex relations

Parameters for Long Short-Term Memory Network [17]

Number of Hidden Layers – see Number of Hidden Layers for ANN
Number of Neurons in Hidden Layers - see Number of Neurons in Hidden Layers for ANN
Epochs – see Epochs for ANN.

2.4.4 Feature Scaling

Scaling is important for gradient-based models as these are designed to work with data close to zero. Using unscaled data for such algorithms may result in longer computing times or even promote the non-convergence of the algorithm. The models which are not gradient based do not require features scaling – for example Random Forest [10].

Standard Scaling

Standard scaling was used in order to scale the features for ANN and LSTM models [10]:

$$x' = \frac{x - mean\ (x)}{\delta} \tag{7}$$

where x is the current value, mean (x) is the mean of the training samples and δ is the standard deviation of the training samples.

2.4.5 Model Evaluation Metrics of Regression Problems

In order to evaluate the accuracy of the regression model the following metrics are used to measure the accuracy for continuous variables.

Mean Absolute Error (MAE)

It is determined as an average over test sample *n* of absolute difference between each predicted value (P_j) and actual observation (O_j). All the individual differences are weighted equally in average that means that the error for outliners is not additionally penalized. MAE describes the average error [18].

$$MAE = \frac{1}{n}\sum_{j=1}^{n} \left|P_j - O_j\right| \tag{8}$$

Mean Square Error (MSE)

It is defined as an averaged squared difference between predicted value (P_j) and actual observation (O_j) over the test sample *n*. Higher the value of MSE the worse the performance of model is. The perfect model would have a value of MSE equal to zero [18].

$$MSE = \frac{1}{n}\sum_{j=1}^{n} (P_j - O_j)^2 \tag{9}$$

Disadvantage of this metrics MSE is that it gives higher weight (penalty) to bigger errors, which may lead to misevaluation of a good model when outliners in dataset occur.

Coefficient of Determination (R^2)

Coefficient of Determination is defined as ratio between how good the prediction of the model is comparing to the model predicting the mean.

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \left(P_j - O_j\right)^2}{\sum_{j=1}^{n} \left(O_j - \overline{O_j}\right)^2} = 1 - \frac{MSE\ (model)}{MSE\ (baseline)} \quad (10)$$

where O_j is mean of observed data

It usually ranges between 0 to 1 with 1 being the perfect model. However, it might also score negative values of R^2 for very poor models when the model is worse than predicting the mean. Advantage of this model is that it is scale free model which means that its evaluation is independent of whether the output values are very small or large [18].

3. Literature review

This chapter shortly reviews previous work carried on the topic of advance control of heat pump in recent years. When analyzing the research papers main focus was concentrated on data origin [19-22] algorithm choice [19], [22-26] features selection [20-22], [24], [26-29] metrics evaluation [22], [24], [26] and general implementation of the control system [8], [19-21], [23], [27], [29], There have been several researches conducted on developing complex controls which can minimize the cost, shift the load or decrease the energy consumption.

First, research papers with focus on heat pump optimized operation were reviewed:

Drgoňa, Picard, Kvasnica and *Helsen* [19] investigate the benefits of replacing the rule-based controller for the predictive control based on machine learning in residential building. A real building is modeled using the Modelica library and it is linearized into a linear time-invariant state space model. Afterwards, the model is used to be trained and controlled using MPC. Lastly the data obtained from the simulation and MPC are used for machine learning model. The paper uses Deep Time Delay Neural Network (TDNN) and Regression Trees (RT) to determine dependency of multiple parameters. The aim is to achieve thermal comfort required by the user while minimizing energy consumption. The research showed that the TDNN can maintain climate comfort while contribute to energy savings. The performance of RT was not as good as TDNN yet better than traditional RBC.

Urieli and *Stone* [20] implement adaptive reinforcement learning agent for heat pump thermostat. The paper developed the thermostat control based on the Markov Decision process (MPD). MPD explored the effects of each state-decision for three days. The following variables were tracked as state T_{indoor}, $T_{outdoor}$, Time of the day, energy consumed by last action. In order to maximize the reward function in long term, random tree regressor was applied. Finally, the thermostat was tested using realistic simulator applied to various typical houses. The simulator was build using linear regression model to fit the model of the building. The paper states that it was achieved between 7 -14.5% of energy savings while keeping the comfort level of an existing thermostat.

Patyn, Ruelens, Deconinck [23] also develop a batch reinforcement learning algorithm and explores implementation of three various neural network types for heat pump controller. The simulation considered long short-term neural network (LSTM), convolutional network (CNN) and multilayer perceptron (MLP). The learning agent maintain the indoor temperature in required ranges. Additionally, agent minimizes electricity cost by shifting load based on day-ahead market prices for next day. According results all algorithms overperformed the average thermostat after 20-25 days of training. As for the ML performance MLP and LSTM presented equally good results outperforming CNN model. MLP was recommended due to shorter training times compared to LSTM.

Vázquez-Cantelia, Kämpfb and *Nagy* [27] researched the optimization of the heat pump operation using two thermal storages and free control reinforcement-learning model. The model is replacing T_{min} and T_{max} of standart control with T_{target} and uses these following variables – heat tank temperature, target temperature, outdoor temperature, indoor temperature and hour of the day. The model is using the

outdoor temperature prior to 22 hours in order to forecast the temperature in next 2 hours. During reinforcement learning and operation, a penalty is given for deviation of the target temperature and for electricity consumption of the heat pump at each hour. Consequently, neural network is used to map the states and actions of the learning in order to find the maximum future reward. The trained model has been tested by implementing into building energy simulator (CitySim).

Qing Zhang, Nan Lv, Shengpeng Chen, Hao Shi and *Zhenqian Chen* [28] explore and compare three different operating and control strategies for the ground hybrid heat pump with cooling tower used for cooling and heating office building in Shanghai. The control strategies take into account metrological parameters, characteristics of the building load, fluid temperature at the inlet/outlet of the heat pump and effect of the storage. The paper concluded that the optimal operation of the system occurs when as a main control parameter is chosen the inlet fluid temperature of the ground loop exchanger and as the subordinate control parameter the difference between the outlet fluid temperature of the heat pump and local wet bulb temperature.

D'Ettore, Conti, Schito, Testi [29] analyzed cost optimal sizing and control strategy for hybrid heat pump system. Additionally, explores the energy savings and operational cost savings between rule-based controller and MPC while using thermal storage. As method for optimal sizing is used mixed linear programming and for MPC considering an ideal forecast of ambient temperature and thermal load. Two cases were taken into account, system without thermal storage, which is operated by rule-based control based on calculated value of COP economic equivalence. And case with thermal storage which is controlled by MPC with target function of minimizing the operation cost of the system, taking into consideration the information of the current system and prediction of the future step. The paper concluded that the system operated by MPC using thermal storage presents energy savings up to 8% compared to system operated by RBC controller with no thermal storage.

Fisher, Madami [8] reviewed current state of the possible implementation and its impact of heat pumps in smart grids. Paper discusses in detail the advance control on various levels – power system, building level, heat pump unit level. Further paper argues that for proper implementation of heat pump in the smart grids advance control is essential. Currently the predictive control is mostly model based and there are only few examples of research using reinforcement learning algorithms

Sun, Djapic, Aunedi, Pudjianto, Strbac [21] research the implementation of smart hybrid heat pump within smart grids. The operation of the smart hybrid heat pump has been analyzed using the data from the pilot project of 75 households (FREEDOM trial). The aim of the control optimization was to reduce the energy price and maximize the operation of the heat pump as a low carbon generation technology. The research also considers scenarios for various gas/electricity price ratio. The method used for control was the predictive demand control algorithm (PDC) which learns the demand characteristics of the house in order to optimize the operation of the system for upcoming day with the aim of minimizing energy consumption and energy cost. The users were able to set their preference temperature in the provided app. The control was using between the hybrid technologies under these factors – outdoor temperature, flow temperature and price of both gas and electricity. The research proved an improvement of the COP using the Smart

Hybrid Heat Pump and obtained flexibility for smart grid. Part of the paper is also a diversity analysis which is an optimal operation of heat pump for various price ratios and household types.

Reinforcement learning algorithm is the most common algorithm used. However, it is not considered as an option for the purpose of the thesis as the indoor temperature is not available variable from SystemFinder. Therefore, research papers focusing on forecasting heating demand were also reviewed:

Dalipi, Yayilgan, and *Gebremedhin* [24] use machine learning model for prediction of heat load in district heating system. The study used three ML algorithms – support vector machine regression (SVR), partial least square (PLS) and random forest (RF). The dataset contains hourly data from buildings in several locations for period of 29 weeks with following features – time of day, forward temperature, return temperature, flow rate and heat load. As a metrics to evaluate the performance of the algorithms have been used mean absolute error (MAE), mean absolute percentage error (MAPE) and correlation coefficients. The results show that SVR overperformed other two algorithms with PLS having the poorest performance.

Ekicki and *Aksoy* [25] are using backpropagation three-layer neural network in order to predict heat demand for various buildings. As features are used transparency ratio, orientation angles and insulations thickness. Neural network achieves to predict building heating energy needs with average accuracy of 94,8- 98,5 %.

Kato, Sakawa, Ishimaru, Ushiro, and *Shibano* [22] used the recurrent neural network in order to tackle the dynamic variation of heat load in district heating system. The paper compares the neural network and RNN for heat load prediction with data of 37 days. Paper investigates if time series of RNN could capture the trend in heat loads which worsen accuracy of prediction of neural networks. As input features were used heat load at each time, day of the week, predicted highest open-air temperature of the day, and predicted lowest open-air temperature of the predicted day in time series of several days. As an evaluation was used mean squared error. RNN have proved to improve the prediction obtaining the MSE of 11,82 comparing to NN which score MSE of 21,05.

Idowu, Saguna, Ahlund and *Schel* [26] conducted a forecast for heat load of several apartment buildings in district heating system. Four predictive algorithms were tested - Support Vector Regression (SVR), Regression Tree (RT), Feed Forwards Neural Network (FFNN) and Multiple Linear Regression (MLR). The data period for training data occurred between 6 to 11 weeks. The paper used external input parameters (outdoor temperature, hour of the day) and internal input parameters (supply temperature, return temperature, flow rate) to predict the thermal load. Paper also studies the effects of combining the internal and external factors. To evaluate the algorithms the following metrics were used - Root Mean-Square Error (RMSE), the Mean Absolute Percentage Error (MAPE) and Correlation coefficient (Corr Coef). The results show that SVR has the best accuracy of 5.6% MAPE.

4. Methodology and Program Development

This chapter describes theoretical procedure carried in order to determine the scope, design and implementation of predictive models. To accomplish this, methodology is introduced as first, followed by detailed specification leading to algorithm development and implementation.

4.1 Methodology

I. Objectives and determination of required data

At first, a scope of required data must be selected based on a determined objective. The objective is to analyze how various factors affect the optimal operation of a hybrid heat pump system and how to predict its operation based on selected machine learning algorithms. To conduct such analysis, a set of typical household energy demand profiles are selected over one year. Dimensioning of system is conducted using SystemFinder. The simulation enables to find the optimal sizing of generation technologies and storage elements tailored to house demands.

To analyze how various technologies and storage elements affect the operation of the hybrid heat pump operation, configurations using different technology setups have been designed. The basic configuration is designed consisting only of a hybrid heat pump system. Photovoltaics, battery and thermal storage are added one by one to the other three configurations, respectively.

Next, SystemFinder is performed for each configuration using demand profiles of selected household and sizing of technologies as input parameters. The output of the simulation are optimal operating profiles of the hybrid heat pump system for the given year. Additionally, parameter variations on price ratio between electricity and gas are conducted since it has significant influence on the optimal operation.

II. Exploratory data analysis and model determinations

The generated data from SystemFinder are object to analysis, which enables to understand properties of various attributes and the interdependencies between them. This analysis is also conducted in order to determine which features are crucial for model implementation. At first, Pearson correlation analysis is carried to establish linear relationship between the features. Additionally, data visualization is performed in order to obtain the specificity of the data and to discover previously invisible patterns. Three models varying in features are designed for all configurations as a result of the performed analysis. These models are designed to analyze how features improve the performance of the prediction.

III. Machine learning algorithm development

Based on the literature review, three machine-learning algorithms are selected based on their implementation on similar applications.

Random Forest – for its easy implementation and good results [19], [24], [26],.
Neural Network – for its common application used for prediction of operation and good results [22], [23], [26].

Long short-term memory – for its option to determine the dynamics trends in prediction based on [22] and [23].

IV. Data preparation

The performance of the machine-learning model is highly dependent on the quality of the data provided. It is essential that the algorithm is provided with filtered, processed data with consistent input and output data at each time step. Since data is provided from simulation, datasets are coherent and no disturbance such as delay of measurement or error in measurement is existing. Steps to be taken to prepare the data for implementation include:

A. Outliners elimination

B. Feature construction

C. Feature scaling

D. Division of the dataset into training and testing data

V. Evaluation metrics

In order to be able to compare the models among each other, a metrics of evaluation must be determined. The description of each metrics is provided in chapter 2.4.5.

VI. Implementation of models

Selected machine learning algorithms are implemented in python script using selected libraries for neural network development. Consequently, algorithms are fed with features based on the model's setup design in step II.

VII. Parameter optimization

To increase the accuracy of the algorithms, the right set of parameters of each algorithm to a given model must be determined. Despite many expert discussion and recommendation, there are no universal rules to determine optimal parameters for each algorithm before being tested on actual data. Therefore, parameter grid optimization must be conducted, testing various combination of preselected parameters, in order to improve the performance and accuracy of the models and to avoid both underfitting and overfitting of the data. Parameters for optimization are described in chapter 2.4.3.

VIII. Training and testing optimized models

Algorithms with optimized parameters are trained with training dataset and subsequently the trained models are tested with testing dataset.

IX. Evaluation and interpretation

The output models are evaluated using the metrics selected in step V. The results are interpreted and presented visually and numerically.

4.2 Program Development

Selection of programming language and libraries

Python has been selected as programming language for its easy implementation and existence of variety libraries supporting machine learning and data analytics. The library packages described below are used:

Pandas [30] – open source library for data analysis

Matplotlib [31] – open source library for data visualization

Sckitlearn [32] – open source machine learning library

Keras [33] – open source neural networks library

TensorFlow [34] – supporting library for neural networks

Description of input profile variables

Ambient temperature

The input temperature profile is an average hourly temperature for the year of 2013 in area of Stuttgart, Germany. The profile is a typical modern climate with four seasons. The temperature ranges from -16.8 °C to 32.4 °C and the mean temperature is 8.9 °C. The hourly occurrence of each temperature level for that year is presented in the histogram [Figure 9]. From the histogram distribution, it can be deducted that the dataset distribution is not symmetric with respect to temperature as one of the main features.

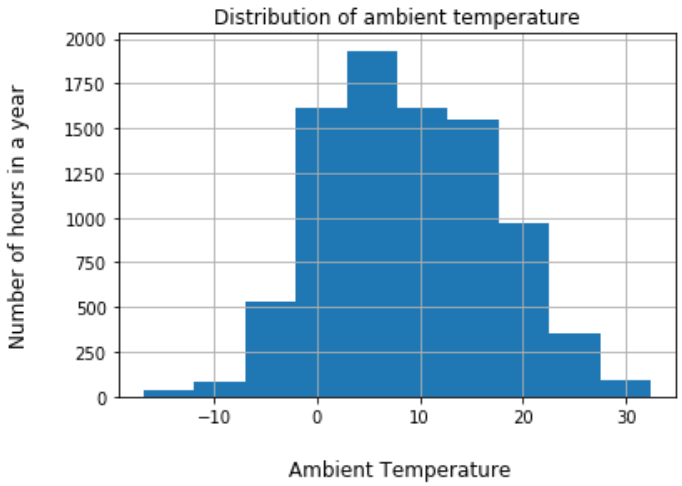


Figure 9 Distribution of ambient temperature for the input profile

Space heating demand and supply temperature

Space heating demand takes a major part of the total energy consumption of the household. The annual space heating demand profile is depicted on Figure 10. The offset temperature for switching off space heating operation is the ambient temperature of 20 °C. For the period between months June, July and

August, space heating is off despite the occurrence of cold days with an ambient temperature below 13 °C.

Space heating demand corresponds to supply heating temperature at the inlet of the heating system of the house. The relation between inlet space heating temperature with respect to ambient temperature is called *heating curve* and can be determined for each household taking into account thermal inertia of the house. Thus, household demand of spacing heating can be modelled using its heating curve.

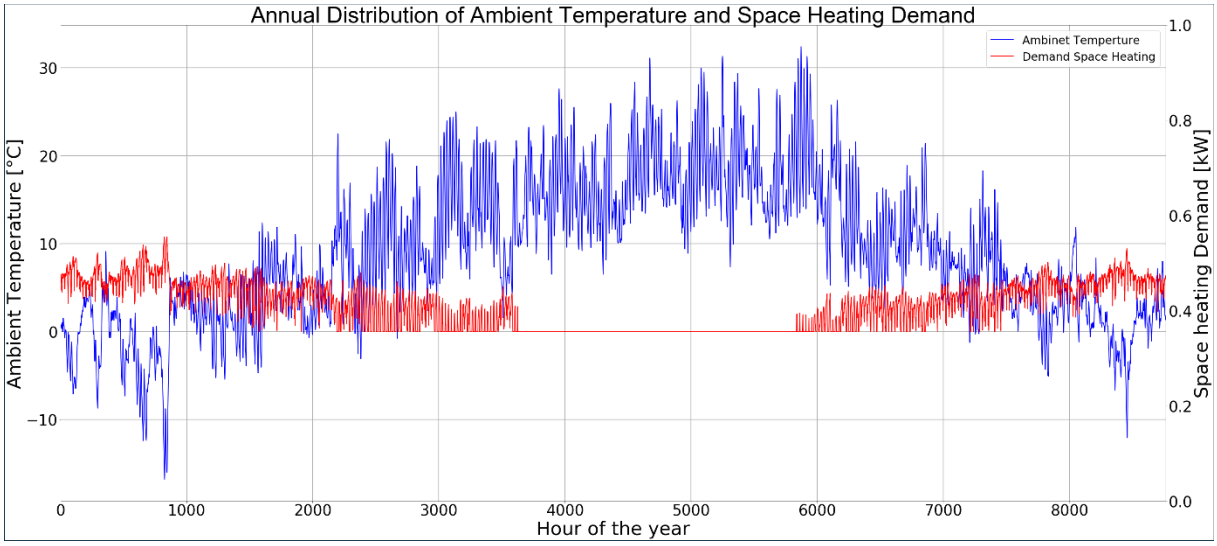


Figure 10 Annual distribution of ambient temperature and space heating demand

Hot water demand

The input profiles of hot water demand of the household is assumed identical for every single day in a year. Figure 11 presents the hot water demand distribution during the day with the peak hours in the morning (7-9 hr) and evening (20-23 hr). During the periods without hot water demand, hot water storage is still maintained in operation, thus requiring heating power supply.

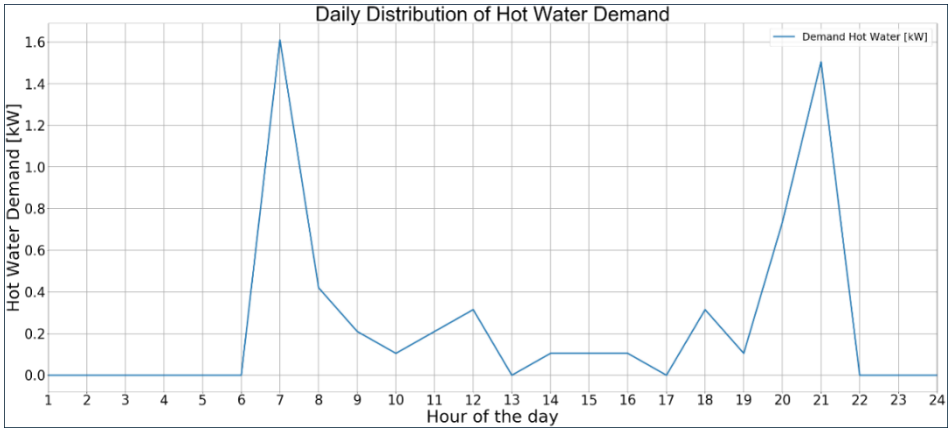


Figure 11 Daily hot water demand profile.

Electricity demand

The electricity demand profile of the household is depicted on Figure 12. In this case, heat pump consumption is not included. Therefore, this demand is identical for all datasets. The dataset has an average hourly electricity consumption of 0.45 kW and maximal electricity consumption 3.7 kW. From the daily mean electricity demand of the household, it can be deducted that the load is lowest during night hours and increases during the day with its peak at 18 o'clock. The electricity consumption of the household is thus important for configurations with photovoltaics installation in order to determine any residual electricity generation for heat pump use.

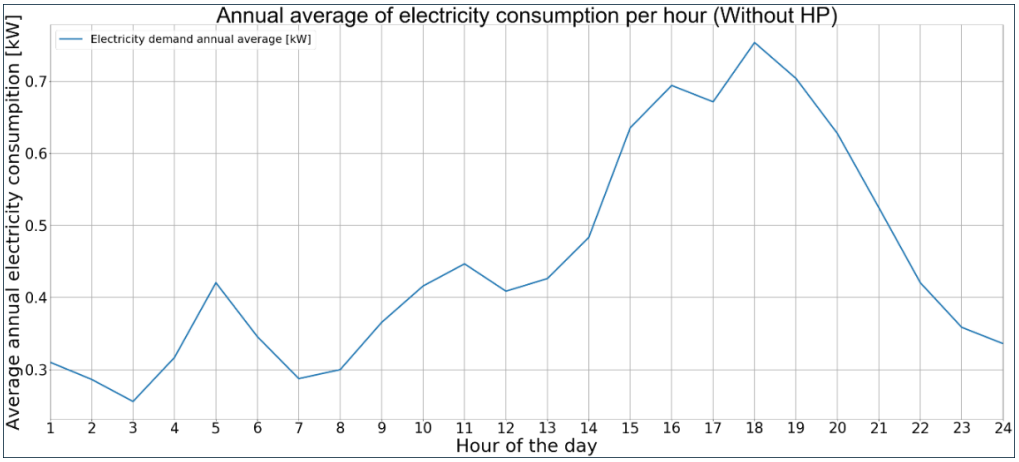


Figure 12 Mean hourly electricity demand of household.

Irradiation profile

Irradiation profile for the Stuttgart area is used as an input for configurations with photovoltaics installation. The average monthly irradiation of the demand profile is presented on Figure 13.

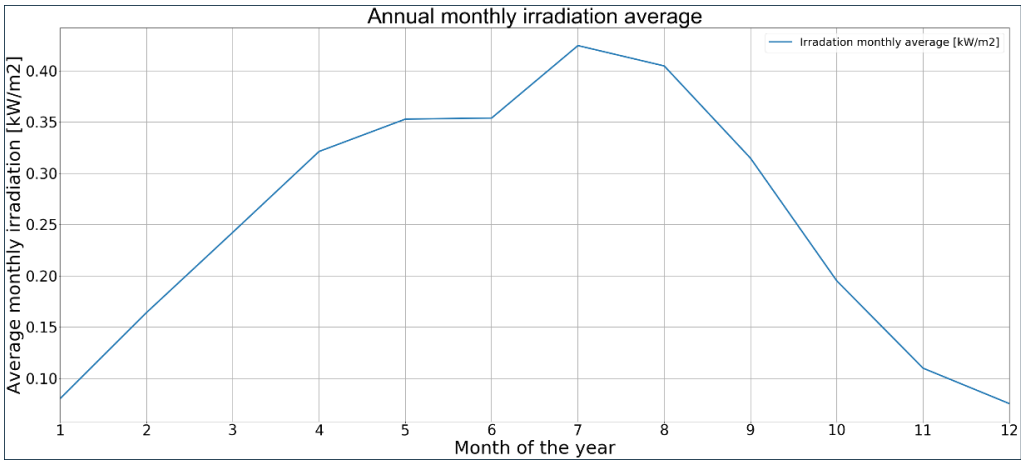


Figure 13 Average monthly irradiation of the dataset.

4.3 Description of Configurations

Dimensioning of generation and storage technologies

Based on the input demand profiles of the household, dimensioning of generation technologies and energy storage elements have been conducted using the SystemFinder. The final sizing of the technologies is presented in Table 2.

Table 2 Dimensioning of the generation and storage systems.

Parameters	
Maximum Heat Pump Power Input [kW]	0,96
Boiler Nominal Power [kW]	8,9
Bivalent Storage Capacity [l]	200
Photovoltaics Maximum Power [kW]	5
Capacity Battery [kWh]	5
Capacity Thermal Storage [l]	400

Configuration setup

Four different configurations have been selected in order to investigate the influence of various factors. The setup of all configurations is listed in Table 3. Firstly, basic configuration (Configuration 1) is designed consisting only of the hybrid heat pump system. This configuration is used as baseline model to compare with other configurations having additional PV installation and storage units. All configurations also include bivalent water storage. For generating data for all configurations, the component size of each generation technology remained constant. Each configuration is generated for two gas/electricity price ratios. The description of all features available from SystemFinder are enclosed in Annex.

Table 3 Configuration setup.

	Gas Boiler	Heat Pump	PV Panel	Battery	Thermal Storage
Configuration 1	✖	✖			
Configuration 2	✖	✖	✖		
Configuration 3	✖	✖	✖	✖	
Configuration 4	✖	✖	✖		✖

4.4 Analysis of Configuration

4.4.1 Configuration 1

Configuration 1 explores operation of the hybrid heat pump system. Within such systems, it is usually the ambient temperature that determines the switching between heat pump and gas boiler.

Based on the correlation matrix for dataset 1 [see Table 4], it can be observed that ambient temperature has a strong negative correlation in relation to gas consumption (*-0.83*). Heat pump operation presents slight negative correlation with respect to ambient temperature (-0.28). The linear relation is not that straight forward as the heat pump operation increases with increasing temperature. However, once crossing below the threshold temperature, the operation of heat pump decreases caused by decreasing COP.

Additionally, there is an almost perfect negative correlation between ambient temperature and the temperature at the inflow of the space heating (T_sh_flow) -0.99 and also between ambient temperature and space heating demand (D_sh) -0.9. This implies that the space heating demand can be determined using these two variables (T_sh_flow and T_amb). Similarly, one can also verify this relation as control by using the heating curve function of house (see chapter 2.2). Heating curve function of the house is a relation between the ambient temperature and the required supply heat that needs to be provided in order to allow the heating system to meet the building thermal loads, as shown at Figure 14.

It can be assumed that the gas generation contributes the most to the heating demands of the house from the correlation 0.84 between the gas consumption (Gas_consum) and the inflow of the space heating (T_sh_flow). To conclude, the correlation analysis shows that ambient temperature is an important feature with respect to control of the bivalent system. Additionally, the relation between the flow temperature of space heating and ambient temperature can determine the space heating demand for respective house.

Table 4 Correlation matrix for configuration 1. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating

	T_amb	Gas_consum	HP_consum	D_sh	T_sh_flow
T_amb	1.0	-0.83	-0.28	-0.9	-0.99
Gas_consum	-0.83	1.0	-0.054	0.89	0.84
HP_consum	-0.28	-0.054	1.0	0.37	0.3
D_sh	-0.9	0.89	0.37	1.0	0.91
T_sh_flow	-0.99	0.84	0.3	0.91	1.0

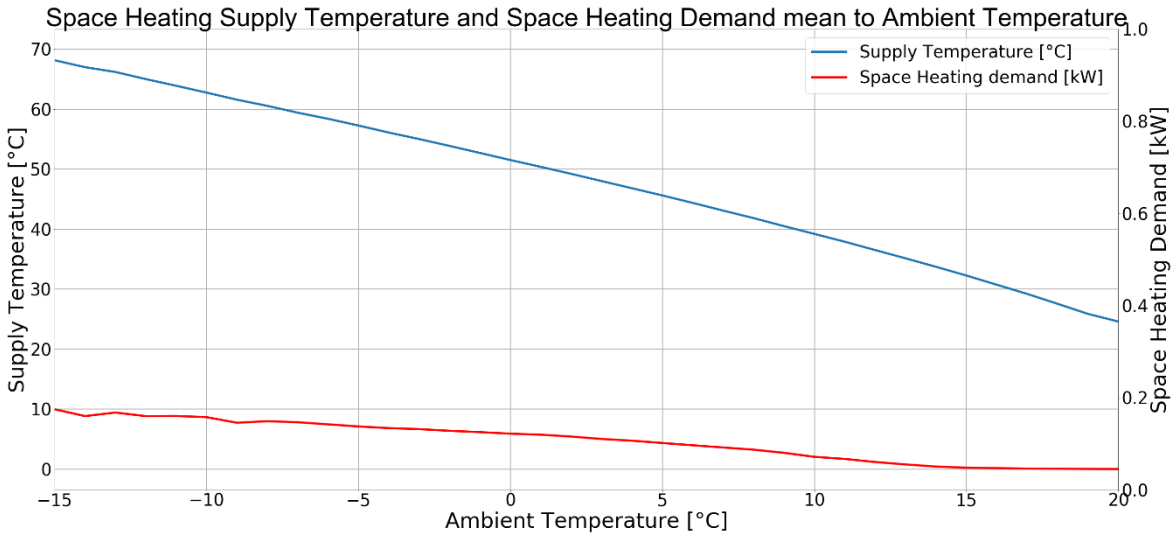


Figure 14 Heating curve function of the house

In order to better understand the operation of the optimized trajectory, a classification into three modes has been introduced.

Gas Boiler Mode is for operation of gas boiler only. As depicted on picture in Table 5 a. and b., there are two cases when the gas boiler mode operation. Firstly, for very low temperature below -12 °C when the efficiency of the heat pump is very poor and uneconomical. Secondly, for ambient temperature range between 6-20 °C when the space heating demand is not requested. However, the temperature level in water storage needs to be maintained and so is the request for hot water demand. COP of heat pump is too low to work in parallel with gas boiler in bivalent mode. Therefore, the low thermal heat needs to be supported by operation of gas boiler. With the increasing price of electricity, gas boiler mode becomes more frequent.

Bivalent Mode is the most common operation when both heat pump and gas boiler operate in parallel as observed at Table 5 c. and d. For the very low temperature range (below -10 °C), bivalent mode is only used for cases when the household demand exceeds the dimensioning of the gas boiler and the heat pump needs to support the operation of the system despite low COP. For the temperatures higher than -5, °C the proportion of the heat pump generation increases not only due to more frequent operation but also due to increasing COP, which results in higher heat supply despite the same heat pump power input. Simultaneously, the proportion of gas boiler generation decreases with the increasing temperature.

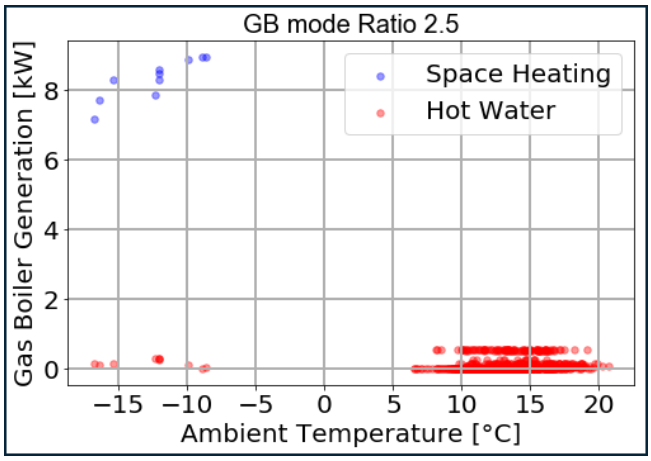
Heat Pump Mode is for operation of heat pump only. It can be divided into two cases. The first case is for temperature ranges between 10 - 15 ° C in which heat pump mode operates mostly to supply space-heating demand. Second case for temperatures above 20°C (which is also the threshold temperature for space heating), in which the generation for hot water demand increases. Comparing the heat pump mode figures for ratio 2.5 and 3, it can be observed that the electricity cost is the major reason for shifting of the heat pump operational ranges and decrease of the heat pump operation in this mode.

Table 5 The operation of hybrid system with respect to ambient temperature based on its mode. From the top left to right: a) Gas Boiler mode operation for electricity/gas price ratio 2.5; b) Gas Boiler mode operation for electricity/gas price ratio 3;

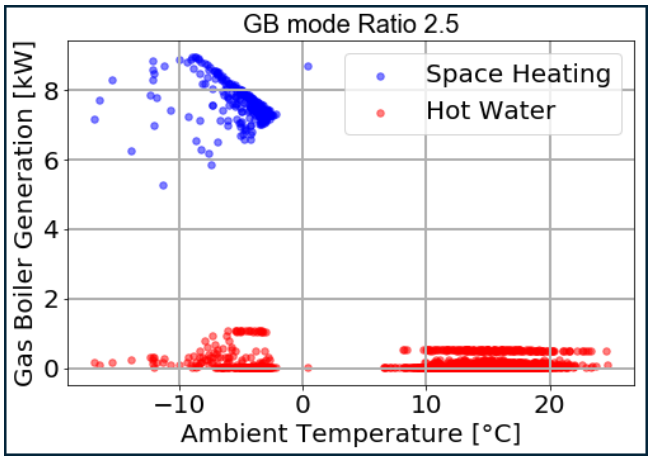
c) Bivalent mode operation for electricity/gas price ratio 2.5; d) Bivalent mode operation for electricity/gas price ratio 3; e) Heat Pump mode operation for electricity/gas price ratio 2.5; f) Heat Pump mode operation for electricity/gas price ratio 3

GB mode

a.

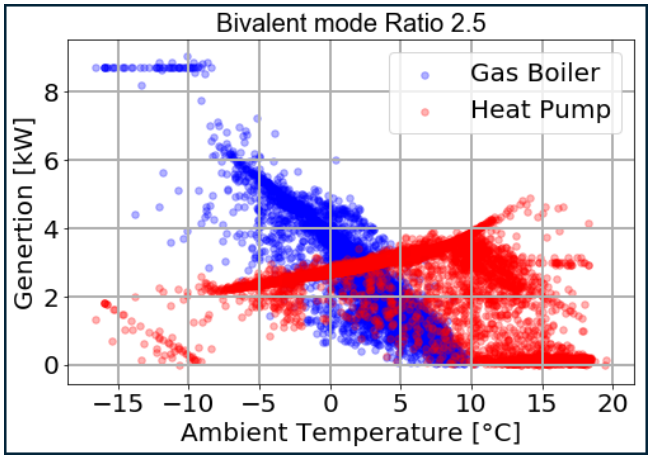


b.

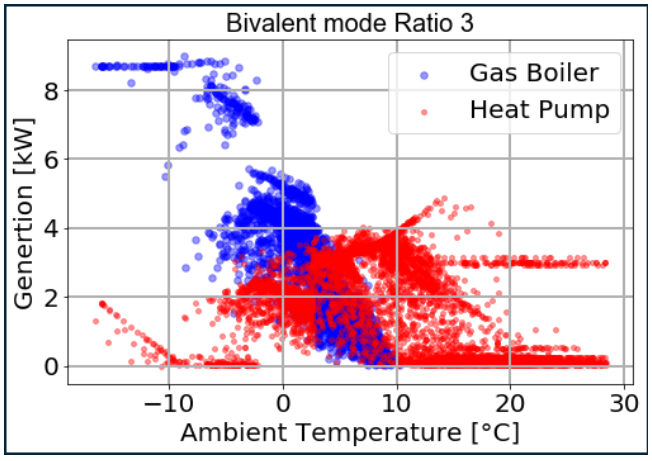


Bivalent mode

c.

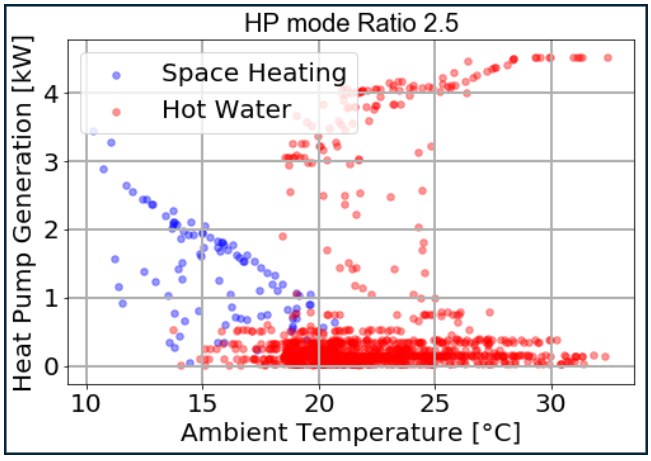


d.

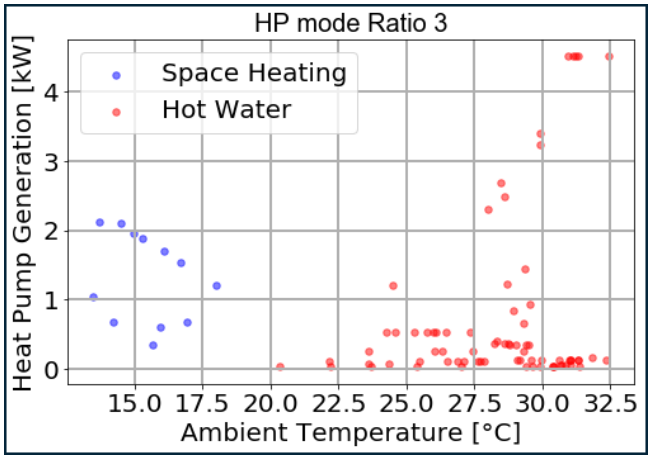


HP mode

e.



f.



Price Impact on the Operation of the system

As the objective function of the model is the minimum investment and minimum operational cost, the variable cost of each generation technology plays a major role in order to determine the optimal trajectory of the device. The histogram below (Figure 15) shows the number of hours of operation of heat pump and gas boiler with the increasing ratio between the electricity price compared to gas price. Additionally, as it can be observed in Figure 16, the increasing electricity price also has an impact on the power input of the heat pump. While heat pump runs on maximum power for the small price ratios, with increasing price of electricity it mostly operates on partial load.

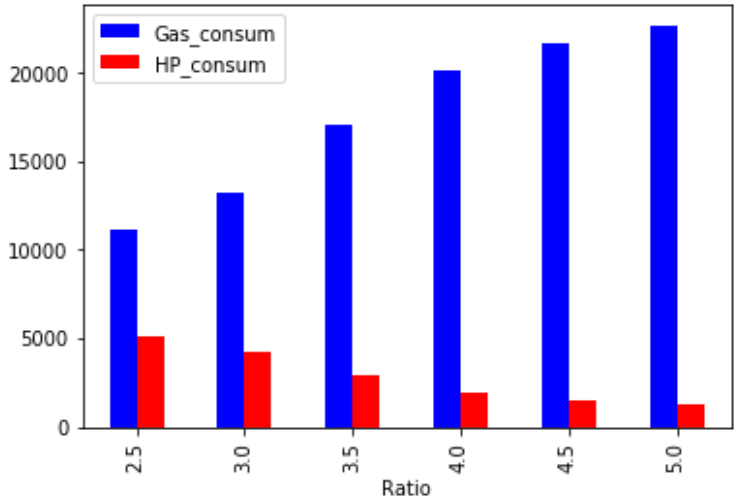


Figure 15 Occurrence of operational hours of each generation technology.

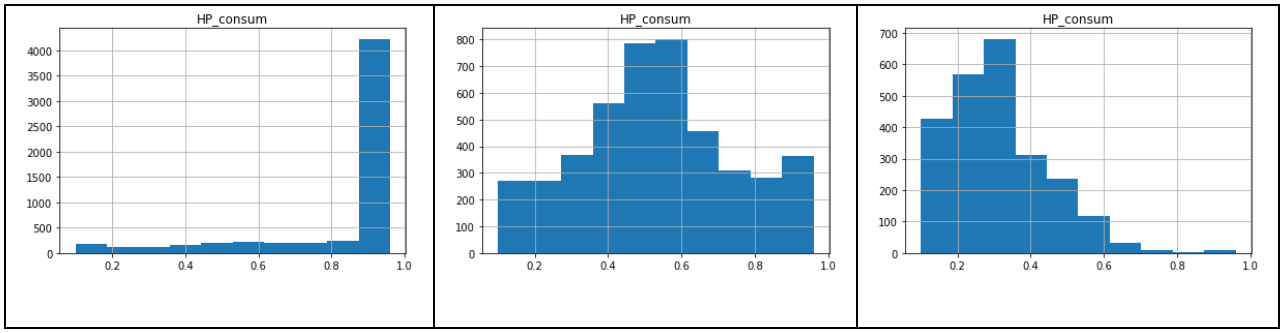


Figure 16 Distribution of input power of the heat pump for price ratios 2,5; 3,5; and 5. Maximum input power equals to 1 and minimum equals to 0.

To implement the effect of price, the feature ratio was initially built into dataset both as numerical or categorical value, respectively. However, the model did not perform well with this feature as it did not capture the relevant dependencies with the categorical value. Therefore, other features must have been introduced into dataset. One representing the efficiency coefficient of heat pump at each hour (COP) provided directly from SystemFinder model. Second representing the price of the heat pump generation taking into account the electricity price and the COP efficiency at each hour. The current dataset is present only with one electricity and gas price ratio, which is identical for the whole year. However, these features could be also implemented for flexible pricing or tariffs in future.

Configuration 1: conclusion

The analysis of the scenario for *configuration 1* was performed in order to determine what features are the most relevant in order to predict in next hour the operation of heat pump and gas boiler.

Several conclusions can be drawn from the performed analysis:

- As the space heating demand takes major part of the energy consumption, it is necessary to identify the main variables, which enable the most accurate prediction of space heating demand.
- From the analysis, one can conclude that demand for the space heating is strongly correlated with ambient temperature (*T_amb*) and temperature at the inlet of the space heating (*T_sh_flow*).
- Additionally, feature displaying whether the space heating is on or off was added. That is, since ambient temperature and space heating demand are strongly correlated. However, during summer the space heating is turn off despite occasional drop below threshold temperature 20°C, which is an offset temperature to turn space heating on during the rest of the year.

As the optimal trajectory (Heat Pump input power and Gas Boiler consumption) was obtained by minimizing the operational cost of the system, variable cost of the system must be implemented into the model. As introduced before, the heat pump operates with variable efficiency. Hence, its coefficient of performance must be determined for every hour with respect to source temperature, sink temperature and heat pump parameters. The model also provides an hourly value of COP for both space heating and hot water demand with respect to discretized level of temperature storage.

As the optimal trajectory is determined by the operational cost, cost of heat pump operation at each hour should be added to dataset. That must take into account the variable COP of heat pump with respect to ambient temperature and discretized level of temperature of the storage (sink temperature). The current model only implements one price of electricity for the whole year and two datasets with various electricity cost are implemented.

In comparison to space heating demand, demand for hot water only takes a small part of total heating demand. However, the hot water supply is demanded throughout the whole year. The current dataset is modeled with the hot water demand profile identical for every day in year. Therefore, the hot water demand for the time step *-23 hours* will predict hot water demand in next hour. When other profile for hot water would be used, model-determining patterns in behavior of the household would must have been determined in order to predict hot water demand.

When modelling the prediction for hot water demand, the level of storage must be also considered. The storage is continuously warmed up each hour in order to keep the level of demanded temperature in storage.

4.4.2 Configuration 2

Configuration 2 investigates the impact of photovoltaics installation on the operation of hybrid heat pump system. The electricity generated from the photovoltaics is primarily used to cover the electricity demand of the household. When residual electricity from PV is available, it is used for heat pump operation or feed back into grid as feed-in tariff respectively. The flow of the PV generation is depicted on Figure 17.

Table 6 Correlation table for configuration 2: T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating, D_hw: Demand hot water, PV_total: Photovoltaics generation, Storage_hw: Hot water storage level

	T_amb	Gas_consum	HP_consum	D_sh	D_hw	PV_total	Storage_hw
T_amb	1.0	-0.82	-0.75	-0.9	-0.019	0.35	0.39
Gas_consum	-0.82	1.0	0.58	0.89	0.061	-0.31	-0.33
HP_consum	-0.75	0.58	1.0	0.84	-0.0078	-0.23	-0.32
D_sh	-0.9	0.89	0.84	1.0	-0.013	-0.39	-0.31
D_hw	-0.019	0.061	-0.0078	-0.013	1.0	-0.061	0.1
PV_total	0.35	-0.31	-0.23	-0.39	-0.061	1.0	-0.18
Storage_hw	0.39	-0.33	-0.32	-0.31	0.1	-0.18	1.0

Comparing Table 6 with the correlation table of configuration 1 [Table 4], where PV was not implemented, there is a significant reduction of the correlation between heat pump input power and both ambient temperature and space heating demands. This implies that PV generation has an impact on the operation of heat pump for space heating generation at lower temperatures where the COP is low, and the operation of heat pump powered by grid electricity is hence not justifiable. On the contrary, the operation of heat pump for hot water trajectory does not increases proportionally with the PV generation.

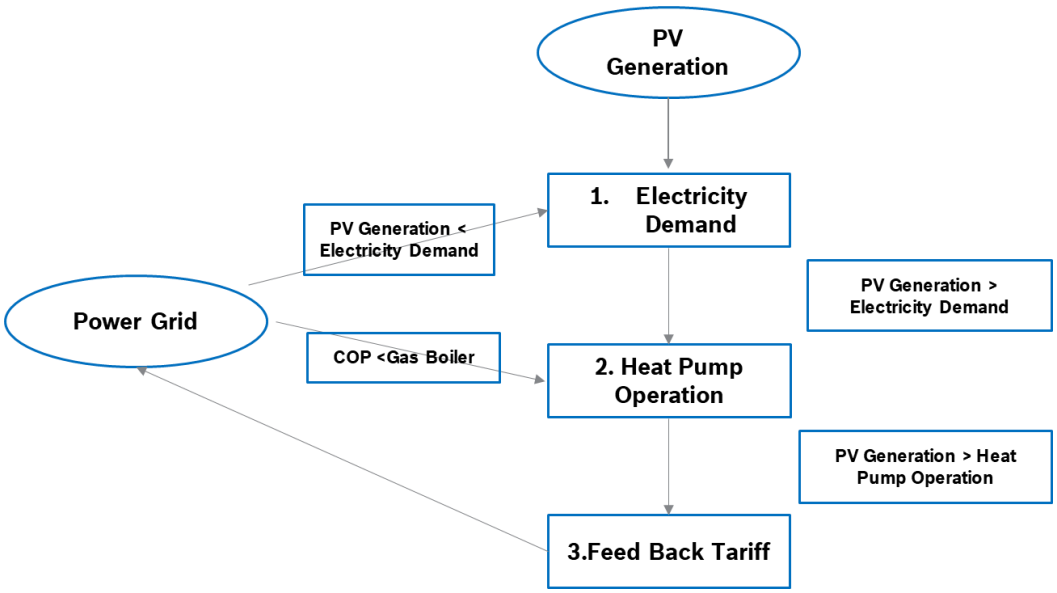


Figure 17 Priority chart of PV generation flow

The impact of photovoltaics on the operation of heat pump can be observed on Figure 18. Heat pump operation for the system with photovoltaics is more frequent (red curve) due to PV generation (orange curve) than for the system without PV (blue curve).

Besides the features used for configuration 1, configuration 2 will also take into account two more feature variables. The photovoltaics generation of the system (PV_total) and the electricity demand of the household (D_ele). Electricity demand must be considered in order to determine how much electricity would be available for other applications (HP, Feed-in).

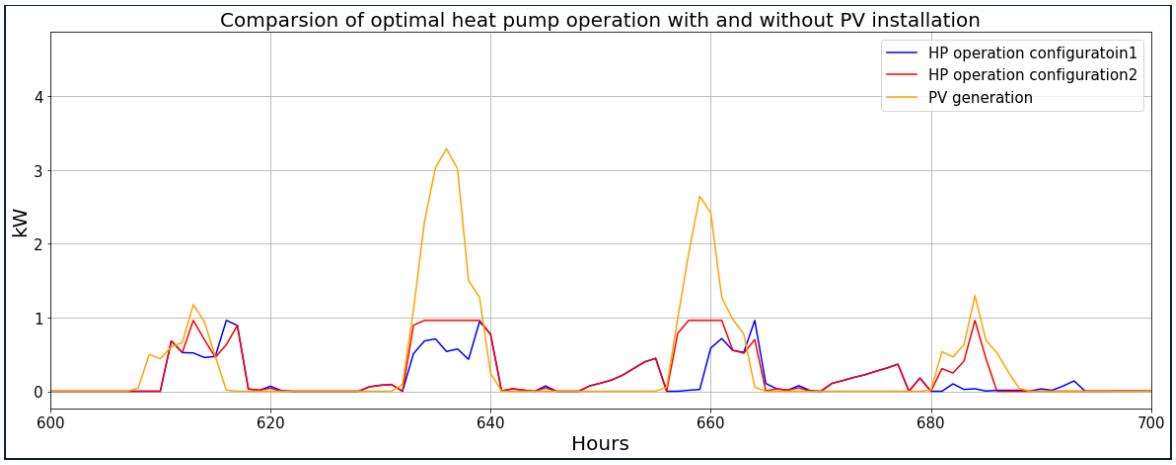


Figure 18 Operation of heat pump for configuration with and without PV installation with respect to PV generation.

4.4.3 Configuration 3

Configuration 3 explores the impact of battery on the operation of hybrid heat pump system with photovoltaics installation. Battery is used to better distribute the difference between the over production of electricity during the day and the household electricity demand during the evenings. Battery at this configuration can be only charged from the photovoltaics installation and the charging directly from the grid is not considered due to the constant price. Discharging of the battery is related to minimize supply of electricity from grid to cover electricity demand without taking into account the heat pump generation. Therefore, the installation of battery does not impact the operation of heat pump. The limitation of the battery is also the limited discharging energy rate which is 0.3 kW per hour. Observing the correlation matrix for configuration 3 [Table 7], it can be concluded that battery discharging has a very low correlation -0.23 with heat pump operation.

Table 7 Correlation table for configuration 3. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_hw: Demand hot water, D_sh: Demand space heating, Storage_hw: Hot water storage level, PV_total: Photovoltaics generation, Bat_dischar: Battery discharging.

	T_amb	Gas_consum	HP_consum	D_hw	D_sh	Storage_hw	PV_total	Bat_dischar
T_amb	1.0	-0.83	-0.75	-0.019	-0.9	0.4	0.35	0.17
Gas_consum	-0.83	1.0	0.58	0.062	0.89	-0.34	-0.31	-0.23
HP_consum	-0.75	0.58	1.0	-0.0087	0.84	-0.33	-0.23	-0.23
D_hw	-0.019	0.062	-0.0087	1.0	-0.013	0.097	-0.061	-0.043
D_sh	-0.9	0.89	0.84	-0.013	1.0	-0.32	-0.39	-0.2
Storage_hw	0.4	-0.34	-0.33	0.097	-0.32	1.0	-0.2	0.19
PV_total	0.35	-0.31	-0.23	-0.061	-0.39	-0.2	1.0	-0.28
Bat_dischar	0.17	-0.23	-0.23	-0.043	-0.2	0.19	-0.28	1.0

4.4.4 Configuration 4

Configuration 4 studies the impact of thermal energy storage on the optimal control trajectory of heat pump. Thermal storage accumulates heat when heat pump operation is optimal and discharges when the demand for space heating is requested. In SystemFinder, the thermal storage is modelled in such way that it is regarded as low-value and thus need to be paired with a certain portion of high-value heat from gas boiler. The limitation of SystemFinder is that it calculates only with energy flows, but it does not take into account temperature variable. Therefore, the model charges the storage when the ambient temperature is high and temperature at the inflow to space heating is low. And the heat is released when the efficiency of the system is low.

The correlation 0.3 between space heating demand (D_sh) and discharging of the thermal storage (PS_dischar) is low. No other correlations between the level of thermal storage, its charging and discharging present any significant correlation.

Table 8 Correlation table for configuration 4. T_amb: Ambient Temperature, Gas_consum: Gas consumption, HP_consum: Heat Pump input power, D_sh: Demand space heating, T_sh_flow: Temperature at the inflow of the space heating, PV_total: Photovoltaics generation, PS_level: Thermal storage level, PS_char: Thermal storage charging, PS_dischar: Thermal storage discharging.

	T_amb	Gas_consum	HP_consum	D_sh	T_sh_flow	PV_total	PS_level	PS_char	PS_dischar
T_amb	1.0	-0.79	-0.73	-0.9	-0.99	0.35	-0.043	-0.075	-0.24
Gas_consum	-0.79	1.0	0.52	0.81	0.79	-0.27	-0.18	0.2	0.025
HP_consum	-0.73	0.52	1.0	0.8	0.75	-0.18	0.18	0.2	0.22
D_sh	-0.9	0.81	0.8	1.0	0.91	-0.39	0.088	-0.01	0.3
T_sh_flow	-0.99	0.79	0.75	0.91	1.0	-0.35	0.043	0.073	0.25
PV_total	0.35	-0.27	-0.18	-0.39	-0.35	1.0	-0.11	0.12	-0.12
PS_level	-0.043	-0.18	0.18	0.088	0.043	-0.11	1.0	-0.0077	0.22
PS_char	-0.075	0.2	0.2	-0.01	0.073	0.12	-0.0077	1.0	0.13
PS_dischar	-0.24	0.025	0.22	0.3	0.25	-0.12	0.22	0.13	1.0

Plotting the thermal storage level [Figure 20] and thermal storage discharge [Figure 19], it can be observed that there are occurrences when the thermal storage operates even for summer months when space heating demand is not requested. Therefore, summer months, during which the space heating is not requested, will be treated as outliners and eliminated.

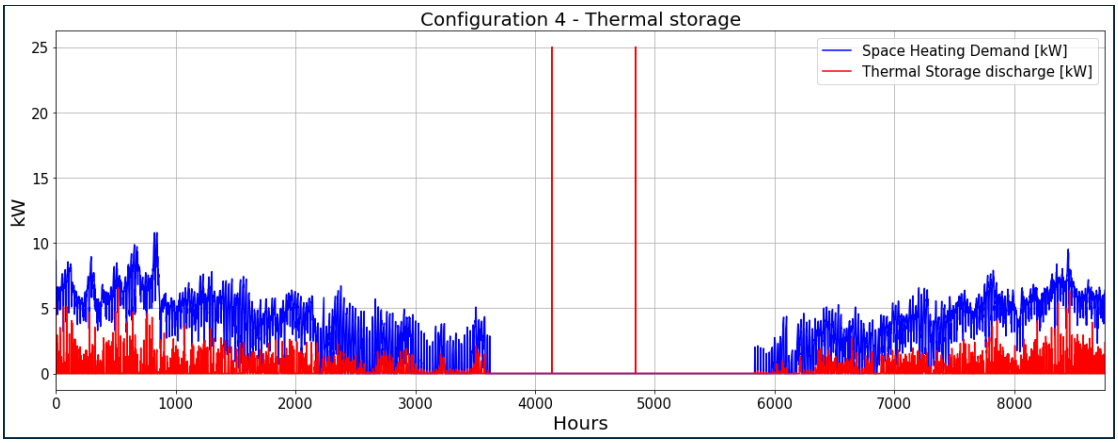


Figure 19 Annual distribution of space heating demand [kW] and thermal storage discharge [kW]. Hourly range between 3624-5832 corresponds to months June-August.

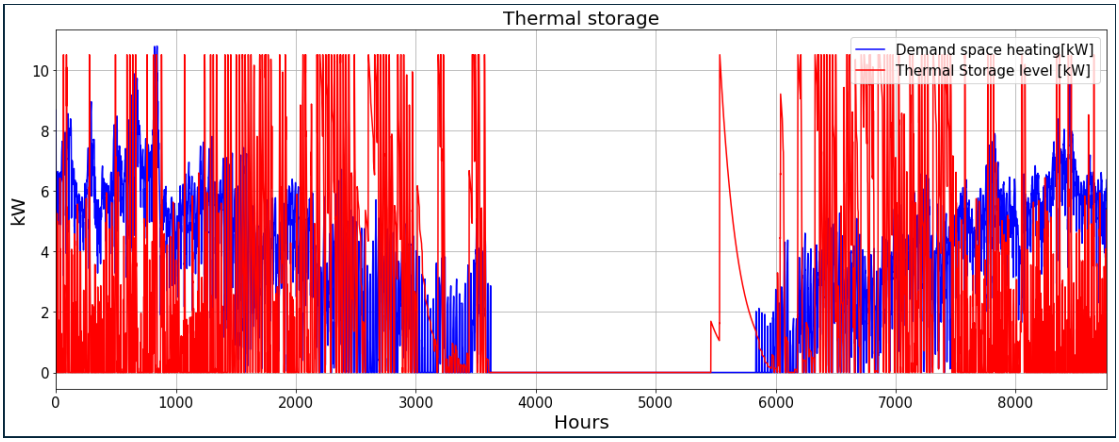


Figure 20 Level of thermal storage [kW] and space heating demand [kW] annual distribution.

4.5 Data Preparation

Features Construction

Timestamp conversion

The data is generated with hourly timestamp, which indicates the hour of the year. In order to handle better the data analysis, the timestamp is transformed into month of the year, day of the month and hour of the day.

Time of the day

The dataset is divided into hour of the day and a feature day/night is added in order to more precisely model the dependency of solar irradiation.

Operating Modes

Operating mode of each hour is added into dataset to determine current state of the system. This feature is mostly used for data visualization or classification. The modes are divided as listed in Table 9.

Table 9 Operating modes table of features

GB mode	Gas Boiler operation only
Bivalent mode	Gas Boiler and Heat Pump operation in parallel
HP mode	Heat Pump operation only
Off	No operation

Price

Price of the heat pump operation at each hour is added with respect to electricity price and current COP at each hour.

Rolling averages

It is used for several features in model 3, in order to represent the time sequence of the feature. It uses several prior time steps of a feature and creates a new feature which is an average of these prior step's values.

Space heating on/off

It is used to determine if the space heating is on or off.

Labels Construction

As the model objective is to predict operation of the hybrid system both gas boiler and heat pump operations are considered as labels.

Gas consumption – considers consumption for both space heating demand and hot water demand

Heat Pump input power – Input power of heat pump is considered to determine its operation. This label has been selected instead of delivered heat since it is not dependent on fluctuating COP.

Feature Scaling

Feature Scaling - Scaling of the features is necessary to implement on the gradient-based algorithms. The features were scaled using the standard scalar (chapter 2.4.4) for the application of NN and LSTM.

Division of the dataset into training and testing data

Division of the dataset into training and testing data. The ratio between training and testing data has been divided into ratio 75 to 25 percent. In case of RF and NN, shuffling of data has been applied contrary to LSTM, which has been used as time series.

Implementation of models

Based on the analysis of the configuration 1 three models have been designed in order to analyses how features impact the accuracy of the prediction.

Model 1: only considers ambient temperature as a feature. It is a baseline model used for benchmarking.

Model 2: uses features which are determined as relevant from the explanatory analysis for each configuration.

Model 3: uses features from *Model 2* and additionally it is extended for time sequence of selected features.

4.6 Parameter Optimization

Determining proper parameters of the algorithm is an essential task in order to obtain the best possible performance of the algorithm. Based on theoretical basis provided in Chapter 2.4.3 the following parameters are selected for grid parameter search with an aim of lowest mean absolute error.

Neural Network

Table 10 Parameters for grid search – Neural Network algorithm.

Neurons input layer	Number of features
Neurons hidden layer	2/3, 2x, 8/3 of neurons in input layer,
Neurons output layer	Number of labels
Activation function input layer	Tanh, linear,
Activation function hidden layers	ReLu
Optimizer	Adam
Learning rate	0,01, 0.001,
Batch size	Range 10-100 with step of 10
Epochs	40

Random Forest

Table 11 Parameters for grid search – Random Forest

Number of Trees	Range 1-500 with step of 20
Maximum Features	Sqrt, log2
Maximum Depth	Range 5-85 with step of 10
Maximum leaf nodes	Range 5-50 with step of 5
Minimum Samples Split	2,5,7,10
Minimum Samples Leaf	1, 2, 4
Resampling Data	True, False

Long Short-Term Memory Network

Table 12 Parameters for grid search – LSTM algorithm

Neurons input layer	Number of features
Neurons hidden layer	2/3, 2x, 8/3 of neurons in input layer,
Neurons output layer	Number of labels
Epochs	50
Time sequence	672 Hours
Batch size	2016

Final Parameters

The results of parameter grid search are set of optimized parameters achieving the lowest mean absolute error from given combinations. Table 13 presents the results of the grid search of the optimal parameters for each model and each algorithm. These parameters are used to construct the algorithms.

Table 13 Optimized parameters for algorithms development.

		Configuration 1			Configuration 2			Configuration 3			Configuration 4		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
RF	Number of estimators	394	158	158	394	158	158	394	158	158	368	158	158
	Max. Features	Sqrt	sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Sqrt	Log2	Sqrt	Sqrt
	Max leaf Nodes	30	45	45	30	45	45	30	45	45	25	45	45
	Min samples leaf	4	4	4	4	4	4	4	4	4	4	4	4
	Min samples split	5	2	2	5	2	2	5	2	2	10	2	2
	Bootstrap	True	False	False	True	False	False	True	False	False	True	False	False
NN	Hidden Layers	1	2	2	1	2	2	1	2	2	1	2	2
	Neurons Input L.	2	9	14	1	13	31	1	14	35	1	15	36
	Neurons L1	2	30	37	2	9	26	2	28	70	2	45	36
	Neurons L2	-	36	28	-	64	32	-	28	93	-	30	36
	Activation Function	Linear	tanh	Linear	Linear	Linear	Tanh	Linear	Linear	Tanh	Linear	Tanh	Tanh
	Learning Rate	0.001	0,001	0.001	0.01	0.01	0.01	0.01	0.001	0.01	0.01	0.001	0.01
	Batch	10	30	20	80	60	90	10	40	70	60	50	60
LSTM	Neurons Input L.	13			12			13			13		
	Neurons L1	18			16			18			18		
	Neurons L2	30			16			30			20		

5. Results

Results can be divided into three summary parts:

I. Overview of factors influencing the operation of the hybrid heat pump system.
II. Evaluation of models and importance of selected features.
III. Evaluation of machine learning algorithms for the purpose of predicting the operation of the hybrid heat pump system (Gas Boiler and Heat Pump).

I. Overview of factor influencing the operation of the hybrid system

From detailed analysis carried in Chapter 4.4 it can be concluded which parameters should be included or considered when building a model for predicting operation of the system. The list of all features used for each model is enclosed in Annex.

For the analysed system, space heating takes major part of the energy consumption of the household. Therefore, it is necessary to determine which features can improve accuracy of space heating consumption prediction. *Ambient temperature* and *inflow temperature of space heating* are the two essential variables considered when predicting the consumption for space heating demand. *Ambient temperature* is necessary for determining the proportion of generation between technologies in hybrid system. That is due to the optimal operation of heat pump being dependent on the temperature gradient between the sink and source. The inflow temperature at space heating can be used to model the approximate heating demand of the household with respect to ambient temperature.

Based on the analysis of Configuration 2, it can be deducted that photovoltaics installation has an impact on the operation of the heat pump in hybrid system. The occurrence of the operation becomes higher especially when compared to system with increasing electricity cost. PV generation also increases the input power resulting in higher heat pump production. *PV generation* and *electricity consumption* are considered as features in order to improve prediction of operation of such system. In real-time system forecasting of irradiation from third parties would be necessary in order to estimate the approximate generation in next hour. Alternatively, separated model for predicting the PV generation would be necessary to determine the potential production in next hour for an actual installed system based on the historical data. However, the accuracy of such model would be necessary to test due to many factors influencing the PV generation locally.

From the analysis of Configuration 3 it has indicated that the battery system does not influence the operation of the heat pump in the system. The operation exhibited similar tendencies as the system without battery included. However, battery decreases the amount of electricity feed back into grid and increases the self-consumption of the household.

The effect on the thermal storage with respect to the optimal operation of thermal storage is not very comprehensible due to simplifications on modelling temperature behaviour in SystemFinder.

Additionally, thermal storage is often charged with gas boiler generation and it is in operation for periods when heating demand is not requested.

Finally, determining COP and consequent price for generation of each technology are features that enable to improve prediction of bivalent heat pump systems.

II. Evaluation of Models

Based on the exploratory analysis conducted in prior steps, three models have been designed to evaluate the importance of the features on the algorithm performance. There are three models designed for each configuration (the list of features for each model is enclosed in annex). First model considers *ambient temperature* at current time step as the only variable used to predict the operation of the hybrid heat pump system. That feature is selected as it is used to determine the ratio of operation between the generation technologies in current hybrid heat pump systems. Second model operates with features at current time step which are selected as relevant from the exploratory analysis provided in Chapter 5.3. Third model is an extension of the second one with additional time series for selected relevant features. The final results and comparison of each model (Model 1 -M1-, Model 2 -M2- and Model 3 -M3-) and each configuration are presented on Figure 21for MAE, Figure 22 for MSE and Figure 23 for R^2.

Observing the results of Model 1 of each configuration, it can be concluded that ambient temperature is an important feature as it presents descent performance of R^2 above 0.67, and errors in prediction for MAE and MSE below 0.4 kW. Model 1 presents a good baseline for comparison to more advance models (Model 2 and 3). Comparing Model 1 for each configuration it can be observed that Model 1 gives the best performance for Configuration 1. That can be explained as the Configuration 1 consist of basic hybrid heat pump system only and no other factors such as PV or storage systems are considered. When other technologies are introduced, the accuracy of the model using ambient temperature decreases as the operation is influenced by these factors.

Examining the results of Model 2 for all configurations, it can be concluded that selected features improved the prediction of model for all configurations as expected. Observing the evaluation results for the first three configurations we could conclude that the accuracy of prediction of operation of the system is quite high. The coefficient of performance R^2 is above 0,91 and error of the MAE and MSE is below 0,17 kW and 0,07 kW respectively. The operation of storages in Configuration 3 and 4 adds extra complexities, which deteriorate the final results, especially in M1 of Configuration 3 and all models in Configuration 4. Model 2 of Configuration 4 does not perform as good comparing to previous three configurations. Possible reasons are the inconsistency of the data for that given configuration; inappropriate selection of features or parameters; or necessity to include time series for this configuration with thermal storage.

Model 3 presents the best results for all configurations, giving R^2 of approximately 95 for neural network algorithm (NN) and above 90 for random forest algorithm (RF). Its error of prediction of MAE are below 0,118 kW for NN and 0,137 kW for RF while MSE below 0.06 kW for NN and 0.07kW for RF (with an exception of Configuration 4).

It can be concluded that proper selection of the features has a significant impact on the performance and accuracy of the algorithm prediction. Adding past time series for selected features improves results. However, the difference in result is not that significant. When selecting the features for the algorithm, it must be always taken into account the availability and accuracy of such data in real time.

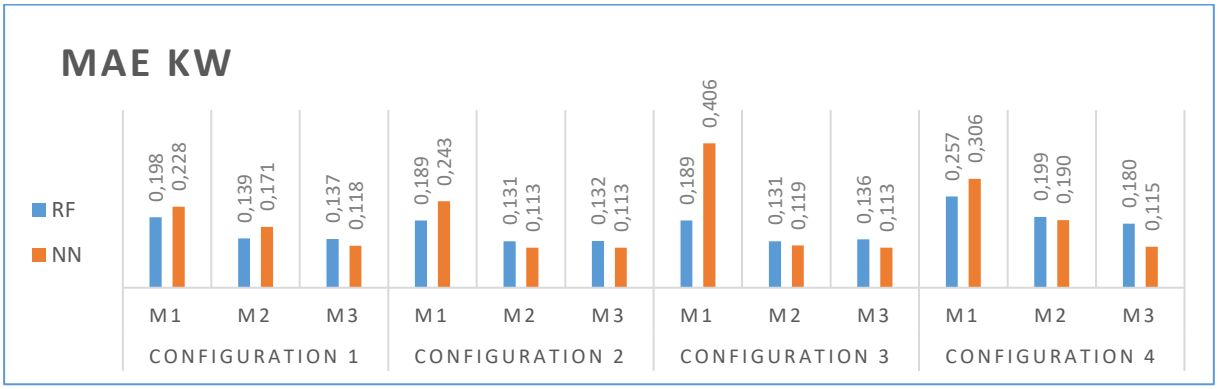


Figure 21 Result evaluation of Mean Absolute Error (MAE) for each model at each configuration.

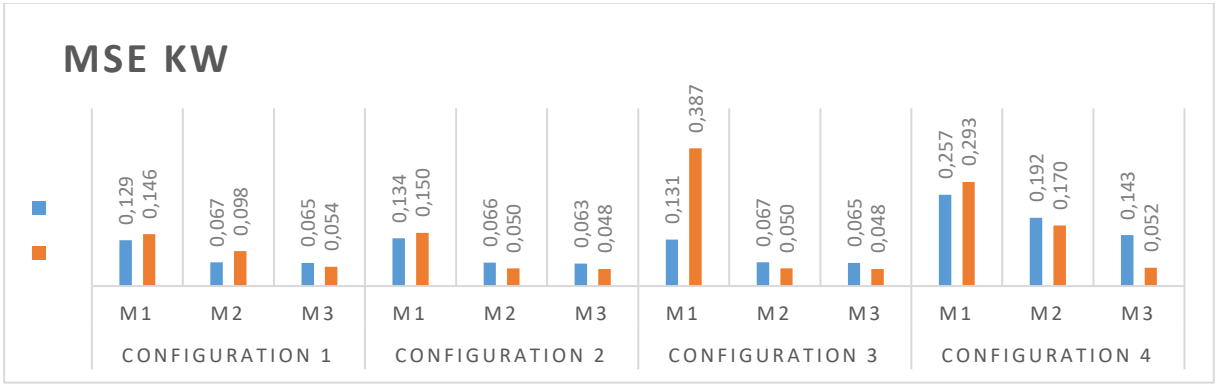


Figure 22 Result evaluation of Mean Square Error (MSE) for each model at each configuration.

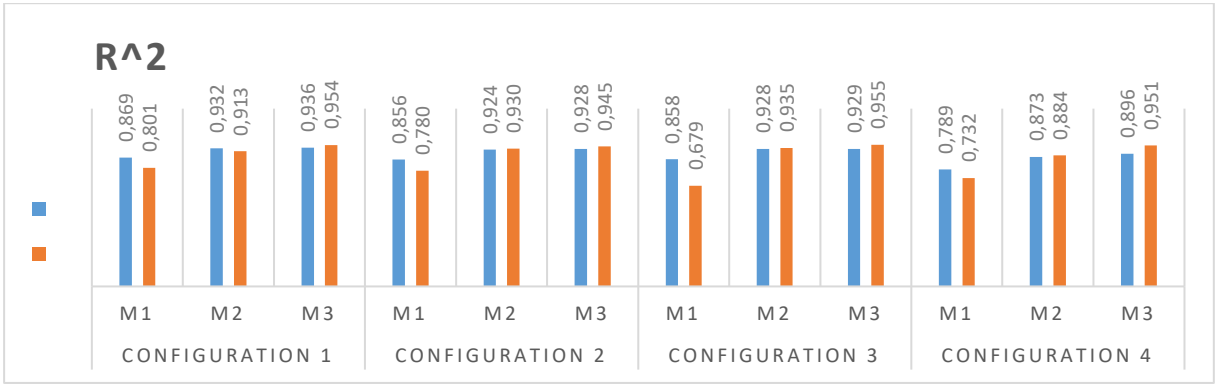


Figure 23 Result evaluation of coefficient of determination (R^2) for each model at each configuration.

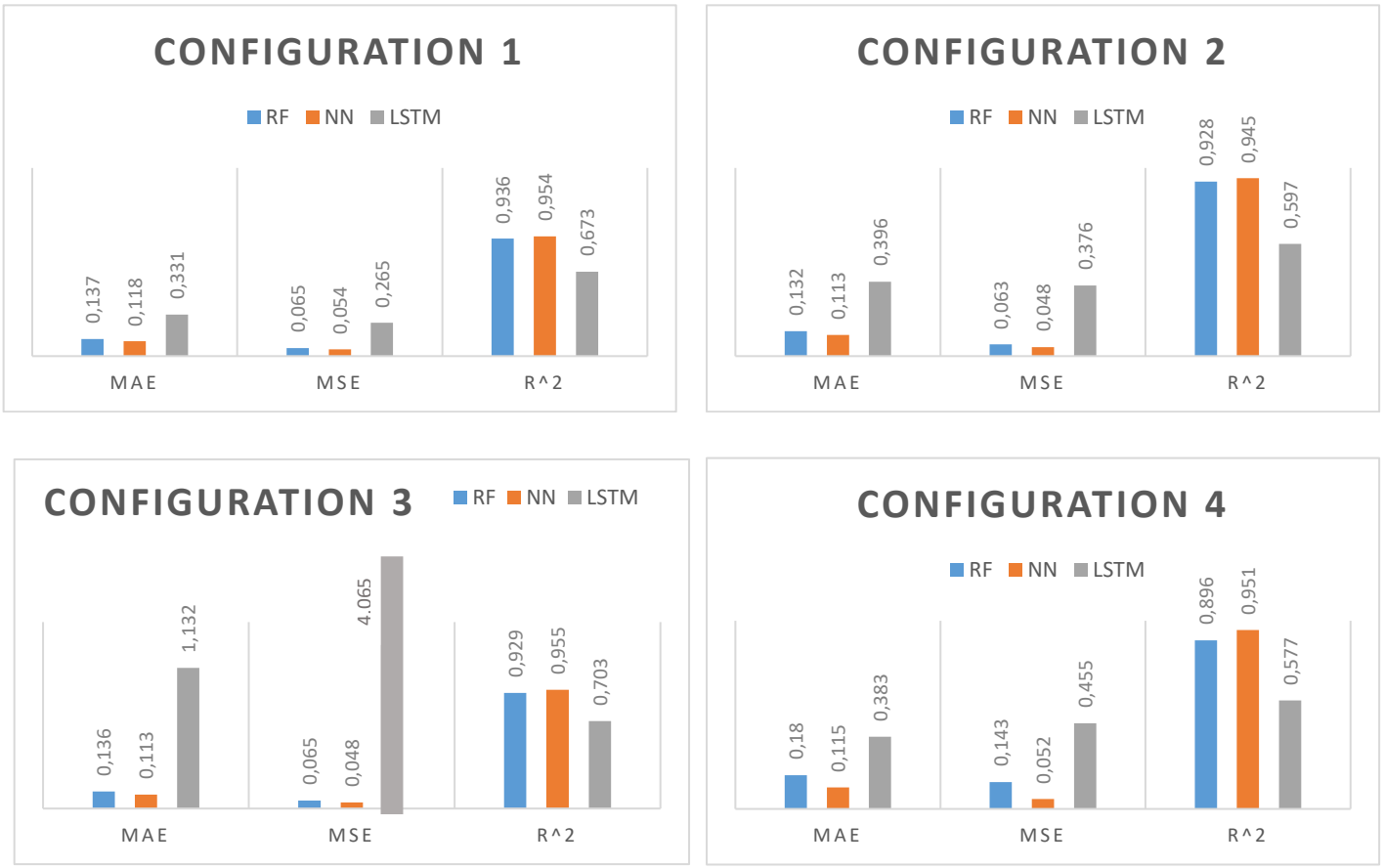


Figure 24 Comparison of results of each algorithm for each configuration

III. Evaluation of machine learning algorithms

Three ML algorithms have been selected in order to analyse its suitability for the problem of prediction of the operation of the hybrid system. Random Forest Regression (RF) as an ensemble method easy to implement on tabular data, Neural Networks (NN) for its capability of capture complex relations between the features, and Long-Short Term Memory network (LSTM), that can process the interdependencies in long time sequences.

RF and NN have been applied on all three models and configurations. The results can be observed on metrics charts depicted on Figure 21, Figure 22 and Figure 23. Additionally, the detailed results for both datasets are enclosed in Annex.

In order to compare the metrics of all three algorithms, Model 3 has been selected. That is due to the fact that model 3 applied on RF and NN includes time series for relevant features and LSTM processes all features as time series. The comparison of results for Model 3 of all three algorithms are depicted on Figure 24.

Random Forest (RF) Regression

The advantage of this algorithm is that it is very easy to implement as it does not require complex pre-processing, data scaling and works well on incoherent data. Also, it is very fast to optimize and to train. The disadvantage of the algorithm is that it is not able to capture complex relations between features as it works as an ensemble method.

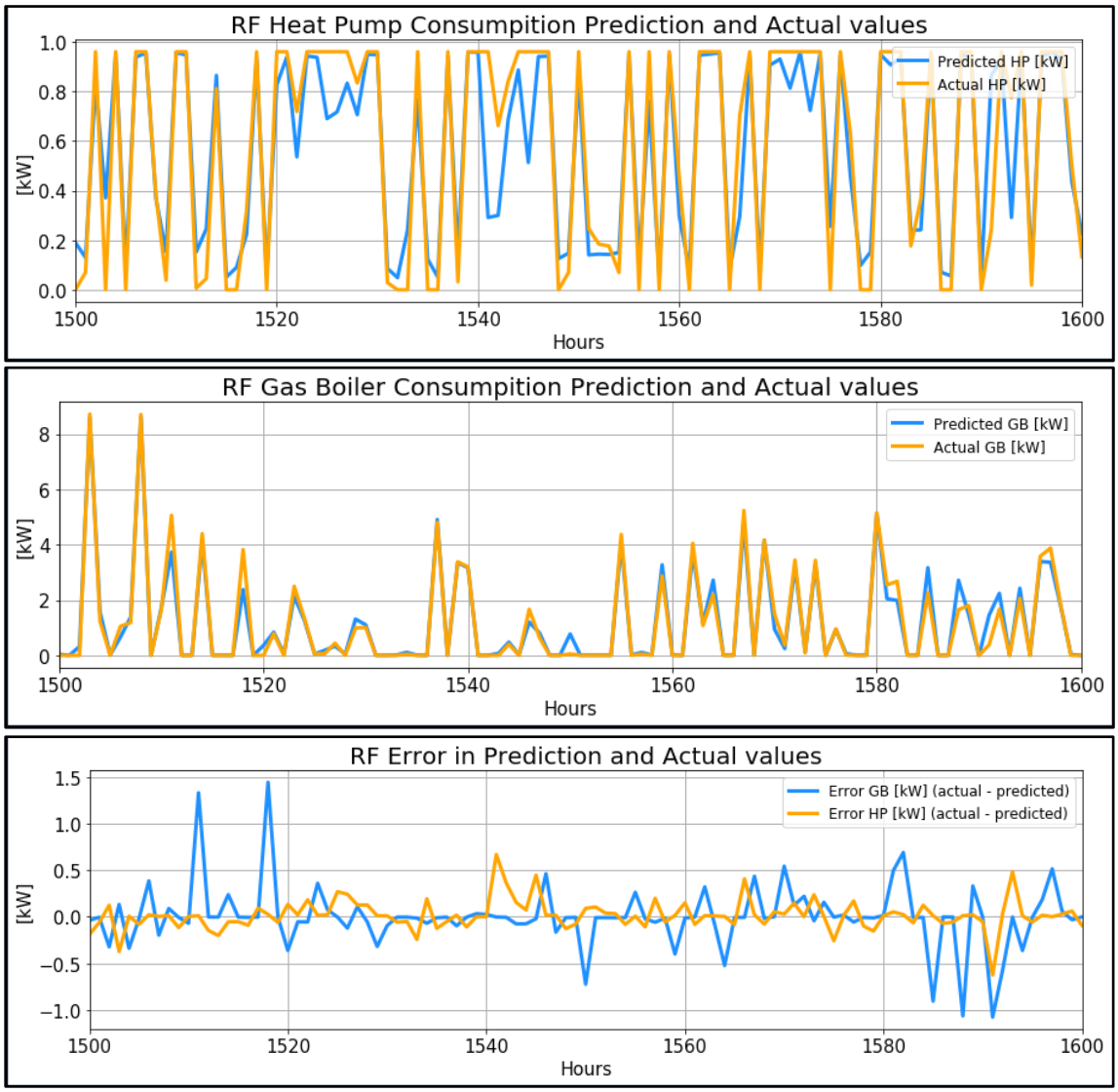


Figure 25 Random Forrest Regression Prediction Results. Configuration 2

For the Model 1, where only ambient temperature is considered, RF gives significantly better results comparing to NN results. The ensemble decision process behind the RF can capture the relation between one feature of ambient temperature and operation of the system in detail.

For Model 2, where more features are considered at current time step, RF can predict the operation of the generation technologies equally good as neural networks. In model 3, for which also time series of certain features are included, the performance of RF is slightly worse comparing to NN. However, the difference is not that significant. RF might not perform as good as NN for cases when the number of features increases and complexity of the relation between the features increases.

The predicted consumption and its prediction error using the RF algorithm are shown in Figure 25. Observing the results achieved, it can be concluded that RF gives consistent results. Interestingly, the random forest algorithm can predict the operation of the heat pump more accurately compared to neural networks. On the other hand, the prediction of gas boiler operation is better predicted with NN.

Neural Network (NN) Regression

The advantage of NN is the capability of capturing complex interdependencies between the features. On contrary, this is redeemed by longer training time and parameter optimization.

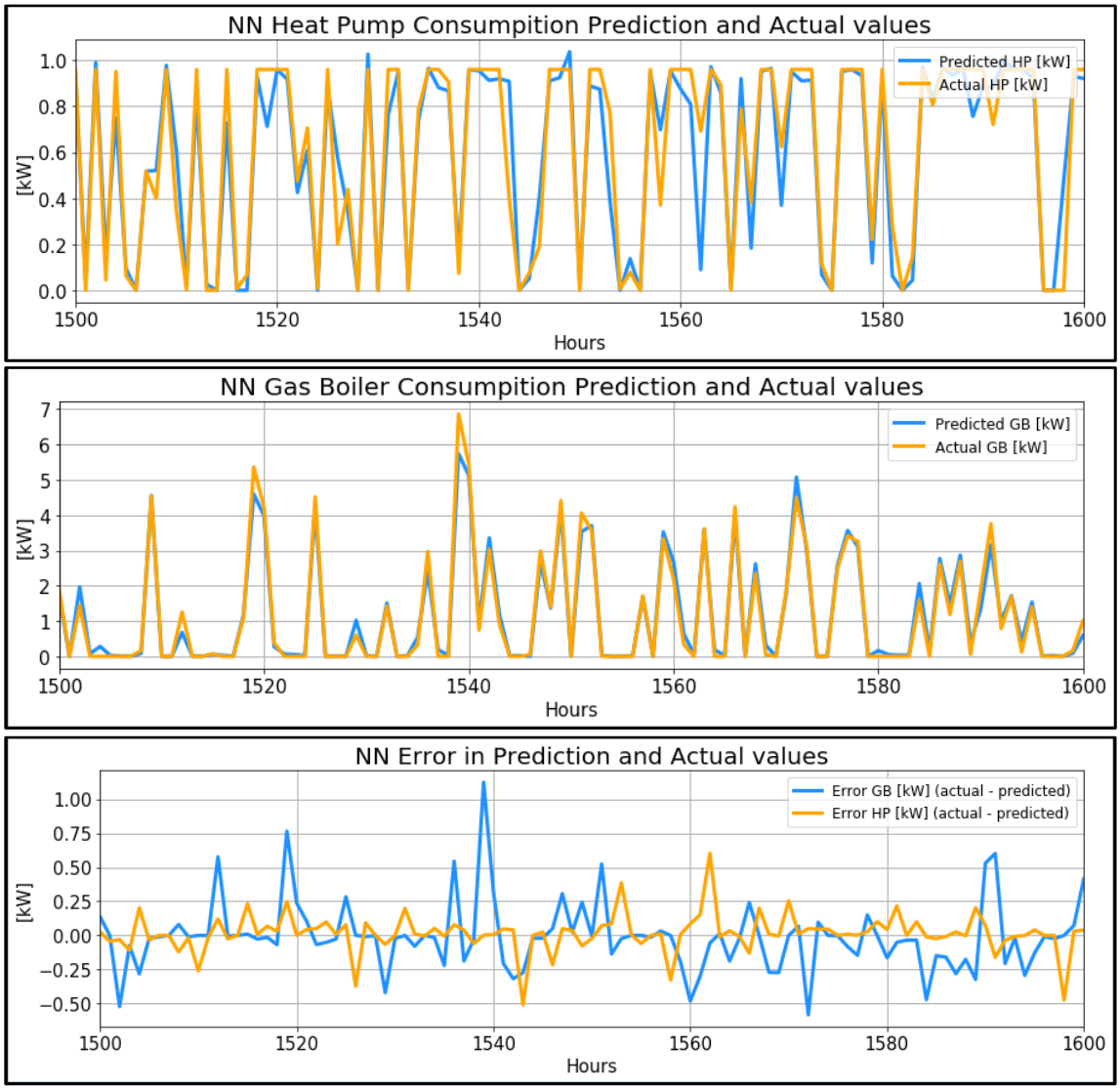


Figure 26 Neural Network Regression Prediction Results Configuration 2.

For model 1, where only one feature (*ambient temperature*) is considered, the NN algorithm underperforms compared to RF. For this model, NN uses one hidden layer in order to detect the functional relation between the feature (*ambient temperature*) and labels (*gas consumption* and *heat pump input power*). Adding more hidden layers worsens the performance of the algorithm. NN presents more inconsistences in performance comparing to RF. One particular case can be observed for Model 1

configuration 3, which demonstrates notably high error. However, the algorithm performs according to expectations when being tested with the dataset 2 of the same configuration.

NN improves its performance with respect to RF algorithm when increasing the number of features. That is due to the capability of NN to detect complex interdependencies between the features. Therefore, NN can predict operation of the system comparably well for Model 2. When the time series are taken account as for Model 3, NN present the best performance out of all three algorithms.

The prediction and error of predictions of NN for Model 3 Configuration 2 can be observed on Figure 26. Observing the table of results for individual prediction of [Annex] each gas boiler and heat pump, it can be concluded that NN can predict the operation of gas boiler better than operation of heat pump.

Long Short-Term Memory (LSTM) Network

Long short-term memory network processes each feature in the time series sequence. LSTM has been selected in order to determine if the model is able to capture the temporal aspect of configuration using storage elements. Based on the total one-year data provided, the model uses sequence of 28 days with the batch size of 4 months after. The results of LSTM prediction can be observed on Figure 27. LSTM does not shuffle data for training and testing as RF and NN do. Instead, it uses sequence of data for training the algorithm and following sequence of data for testing.

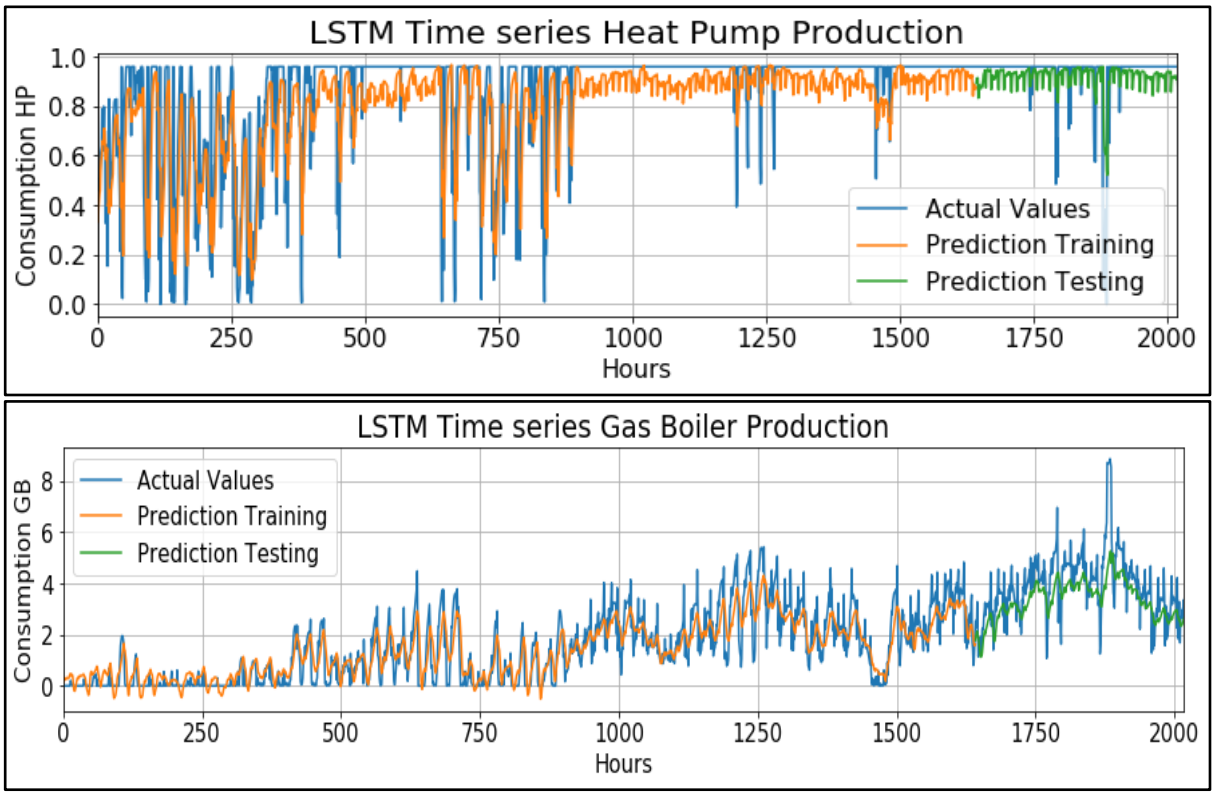


Figure 27 Long Short-term memory network training and testing results configuration 2.

The performance of LSTM is very inconsistent with R^2 between 0,577-0,703, error MAE between 0,331-1,132 and error of MSE between 0,206-4,065. This algorithm has not been proven as suitable for such problem primarily for two reasons. Firstly, the training of such model is lengthy as the number of features

is high. Secondly the performance of the LSTM is very low comparing to both RF and NN. LSTM might improve its performance by providing longer period of data in order to detect seasonal specificity.

6. Conclusion

The aim of the presented thesis was to analyze the optimal operation of the hybrid heat pump system based on various variables and pricing using selected machine learning algorithms. At the same time, the potential of the algorithms was evaluated in order to enable the prediction of the optimal operation of the hybrid systems.

From the analysis and results achieved, it can be concluded that regression machine learning models have the potential to be a significant part of the energy management activities for the future energy household management. It is necessary to always make sure what data is available in real time to be considered as a feature for the algorithm.

The results of the evaluation of the machine learning models show that both random forest and neural networks with optimized parameters are suitable algorithms to predict the operation of the hybrid heat pump systems. On the other hand, the LSTM algorithm performed poorly for all configurations. Implementation of random forest algorithm should be considered for cases when easier implementation, faster training and optimization times are preferred and for cases with fewer features. Neural network is more recommendable for cases with large number of features and when accuracy is being prioritized.

SystemFinder model could be used as a tool in order to obtain training data for prediction of an optimal operation of the system by inserting household demand profiles and technology dimensioning as an input. The data generated from SystemFinder would be suitable to predict the operation with accepted accuracy. In order to verify if SystemFinder could be utilized for generation of such data, an implementation and data collection in an actual physical house would be necessary to conduct. Such verification would consist of data collection from sensors in real system in order to implement them as an input variable to generate the optimal operation. Consequently, an analysis of the optimal operation and training of the algorithms would be carried and evaluated. Such implementation of the predictive algorithm into control of hybrid heat pump would be the next step.

Based on the results in this work, recommendations for the future research and implementation are listed as follows:

1. Data for longer period of time than one year for selected household should be provided in order to improve the accuracy of the algorithm and at the same time prevent algorithm from overfitting.
2. Additionally, to establish if the approach is adaptable, the algorithms should be tested on data for various households and location profiles.
3. Consider smaller time interval of prediction (20 minutes or 30 minutes interval) as well as larger time intervals to improve the accuracy of the model.
4. Adding indoor temperature as feature would open the approach also for other methods such as reinforcement learning.
5. As a next step would be recommendable to implement predictive model into controller which also takes into account the user preference as a respond value and test the prediction in real settings to verify its functions.

References

[1.] Eurostat. *Consumption of Energy, Statistics Explained Website.*[Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/Consumption_of_energy. [Accessed August 2019].

[2.] European Parliament and Council. *(2010).* Brussels. *Directive 2010/31/EU on energy performance of Buildings.* [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2010:153:0013:0035:EN:PDF. [Accessed August 2019].

[3.] European Parliament and Council. (2012). Brussels. *Directive 2012/27/EU on energy efficiency.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32012L0027. [Accessed August 2019].

[4.] Department for Bussiness, Energy and Industrial Strategy. (December 2017). *Hybrid Heat Pumps Final Report.* [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/700572/Hybrid_heat_pumps_Final_report-.pdf. [Accessed August 2019].

[5.] Bagarella, G. Lazzarin,R. and Noro,M. (25.4.2016). *Annual simulation, energy and economic analysis of hybrid heat pump systems for residential buildings.* Applied Thermal Engineering, v. 99, pp. 485-494.

[6.] Bagarella, G. Lazzarin, R. and Noro, M. (1.5.2016). *Sizing strategy of on–off and modulating heat pump systems based on annual energy analysis.* International Journal of Refrigeration*,* v. 65, pp. 183-193.

[7.] Rees,S. (1.1.2016). *An introduction to ground-source heat pump technology.* Advances in Ground-Source Heat Pump Systems, pp. 1-25.

[8.] Fischer,D. and Madani,H. (10/2017). *On heat pumps in smart grids: A review.* Renewable and Sustainable Energy Reviews*.* v. 70, pp. 342-357, 10/2017.

[9.] Rolando, D. and Madani,H. (2018). *Smart Control Strategies for Heat Pump Systems.* [Online]. Available: http://effsysexpand.se/wp-content/uploads/2018/09/P18_Project_Report_final_reviewed.pdf. [Accessed August 2019].

[10.] Bronwlee, J. (2016). *Machine Learning Mastery with Python*. Melbourne, Machine Learning Mastery.

[11.] Dietrich,C.F. (1973). *Uncertainty, calibration, and probability: The statistics of scientific and industrial measurement.* New York: Wiley.

[12.] Rodgers, J.L. and Nicewander,W.A. (1988). *Thirteen Ways to Look at the Correlation Coefficient.* The American Statistician, v. 42, n. 1, pp. 59-66.

[13.] Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems.* Sebastopol, CA: O'Reilly Media.

[14.] Heaton, J. (2008). *Introduction to Neural Networks for Java, Heaton Research.*

[15.] Amidi Afshine, A. S. (26.11.2018). *Recurrent Neural Networks Cheatsheet.* [Online]. Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks. [Accessed August 2019].

[16.] Pedregosa, F. Varoquaux, G. Gramfort,A. Michel,V. Thirion,B. Grisel,O. Blondel,M. Prettenhofer,P. Weiss,R. Dubourg,V. Vanderplas,J. Passos, A. Cournapeau, D. Brucher, M. Perrot, M. Duchesnay, E. and Louppe,G. (2012). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, v. 12, p.,10.

[17.] Demuth, H.B. Beale, M. H. De Jess, O. and Hagan, M. T. (2014). *Neural Network Design.* 2nd editor, USA: Martin Hagan.

[18.] Makridakis,S. (1.12.1993). *Accuracy measures: theoretical and practical concerns.* International Journal of Forecasting, v. 9, n. 4, pp. 527-529.

[19.] Drgoňa,J. Picard,D. Kvasnica,M. and Helsen,L. (15.5.2018). *Approximate model predictive building control via machine learning.* Applied Energy, v. 218, pp. 199-216.

[20.] Urieli,D. and Stone,P. *A Learning Agent for Heat-pump Thermostat Control.* Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems, Richland, SC, 2013.

[21.] Sun,M. Djapic,P. Aunedi,M. Pudjianto,D. and Strbac,G. (1.8.2019). *Benefits of smart control of hybrid heat pumps: An analysis of field trial data.* Applied Energy. v. 247, pp. 525-536.

[22.] Kato,K. Sakawa,M. Ishimaru,K. Ushiro,S. and Shibano,T. (2008). *Heat load prediction through recurrent neural network in district heating and cooling systems*. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.

[23.] Patyn,C. Ruelens,F. and Deconinck,G. (2018). *Comparing neural architectures for demand response through model-free reinforcement learning for heat pump control*. IEEE International Energy Conference (ENERGYCON), Limassol, pp. 1-6.

[24.] Dalipi,F. Yildirim Yayilgan,S. and Gebremedhin,A. (2016). *Data-driven Machine Learning Model in District Heating System for Heat Load Prediction.* Applied Computational Intelligence and Soft Computing v. 2016, p. , 10.

[25.] Ekici, B.B. and Aksoy, U.T. (1.5.2009). *Prediction of building energy consumption by using artificial neural networks.* Advances in Engineering Software. v. 40, n. 5, pp. 356-362.

[26.] Idowu,S. Saguna,S. Ahlund,C. and Schelén,O. (10/2015). *Forecasting heat load for smart district heating systems: A machine learning approach.* 2014 IEEE International Conference on Smart Grid Communications, SmartGridComm 2014, pp. 554-559.

[27.] Vázquez-Canteli,J. Kämpf,J. and Nagy,Z. (1.9.2017). *Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration*. Energy Procedia, v. 122, pp. 415-420.

[28.] Zhang,Q. Lv,N. Chen,S. Shi,H and Chen,Z. (1.1.2015). *Study on Operating and Control Strategies for Hybrid Ground Source Heat Pump System*. Procedia Engineering, v. 121, pp. 1894-1901.

[29.] D'Ettorre,F. Conti,P. Schito,E and Testi,D. (5.2.2019). *Model predictive control of a hybrid heat pump system and impact of the prediction horizon on cost-saving potential and optimal storage capacity.* Applied Thermal Engineering, v. 148, pp. 524-535.

[30.] W. M. P. D. Team, *pandas: powerful Python data analysis toolkit,* Release 0.25.2. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/pandas.pdf. [Accessed April 2019].

[31.] D. D. E. F. M. D. John Hunter. *Matplotlib.* Release 3.1.1. [Online]. Available: https://matplotlib.org/3.1.1/Matplotlib.pdf. [Accessed April 2019].

[32.] s.-l. developers, *scikit-learn user guide.* Release 0.21.3. [Online]. Available: https://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf. [Accessed April 2019].

[33.] K. Developers, *Keras: The Python Deep Learning library.* [Online]. Available: https://keras.io/. [Accessed April 2019].

[34.] T. Developers, *TensorFlow Core r2.0, API Documentation.* [Online]. Available: https://www.tensorflow.org/api_docs. [Accessed April 2019].

[35.] Beck,T. Kondziella,H. Huard,G. and Bruckner,T. (15.2.2017). *Optimal operation, configuration and sizing of generation and storage technologies for residential heat pump systems in the spotlight of self-consumption of photovoltaic electricity.* Applied Energy, v. 188, pp. 604-619.

[36.] Robert Bosch GmbH (2019) *Hybird Heat Pump presentation material.* Renningen: Corporate Research Unit.

[37.] Harp. *Harp Random Forest.* [Online]. Available: https://dsc-spidal.github.io/harp/docs/examples/rf/. [Accessed August 2019].

[38.] Super Data Science. (August 2019). *Artificial Neural Networks (ANN).* [Online]. Available: https://www.superdatascience.com/pages/deep-learning. [Accessed August 2019].

Annex

Table 14 The complete evaluations of all algorithms and models for both datasets.

Dataset1		Configuration 1						Configuration 2						Configuration 3						Configuration 4					
		RF			NN			RF			NN			RF			NN			RF			NN		
		Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP
M1	MAE	0,198	0,383	0,051	0,228	0,336	0,199	0,189	0,203	0,114	0,243	0,362	0,208	0,189	0,071	0,054	0,406	0,573	0,284	0,257	0,190	0,108	0,306	0,304	0,310
	MSE		0,129			0,146			0,134			0,150			0,131			0,387			0,275			0,293	
	R^2		0,869			0,801			0,856			0,780			0,858			0,679			0,789			0,732	
M2	MAE	0,139	0,293	0,110	0,171	0,044	0,306	0,131	0,044	0,034	0,113	0,003	0,050	0,131	0,041	0,044	0,119	0,043	0,071	0,199	0,149	0,101	0,190	0,032	0,144
	MSE		0,067			0,098			0,066			0,050			0,067			0,050			0,192			0,170	
	R^2		0,932			0,913			0,924			0,930			0,928			0,935			0,873			0,884	
M3	MAE	0,137	0,331	0,085	0,118	0,152	0,264	0,132	0,067	0,032	0,113	0,004	0,076	0,136	0,052	0,046	0,113	0,013	0,095	0,180	0,086	0,232	0,115	0,002	0,051
	MSE		0,065			0,054			0,063			0,048			0,065			0,048			0,143			0,052	
	R^2		0,936			0,954			0,928			0,945			0,929			0,955			0,896			0,951	

Dataset2		Configuration 1						Configuration 2						Configuration 3						Configuration 4					
		RF			NN			RF			NN			RF			NN			RF			NN		
		Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP	Total	GB	HP
M1	MAE	0,231	0,373	0,037	0,287	0,354	0,201	0,223	0,204	0,113	0,341	0,178	0,098	0,227	0,062	0,035	0,282	0,331	0,287	0,323	0,184	0,106	0,368	0,334	0,323
	MSE		0,165			0,222			0,185			0,246			0,180			0,235			0,415			0,444	
	R^2		0,827			0,738			0,800			0,506			0,804			0,701			0,710			0,636	
M2	MAE	0,179	0,283	0,115	0,171	0,044	0,306	0,166	0,051	0,045	0,174	0,030	0,174	0,167	0,043	0,036	0,166	0,093	0,166	0,268	0,129	0,104	0,190	0,032	0,144
	MSE		0,115			0,098			0,105			0,099			0,104			0,092			0,325			0,170	
	R^2		0,897			0,913			0,881			0,869			0,887			0,885			0,803			0,884	
M3	MAE	0,171	0,322	0,119	0,151	0,017	0,017	0,163	0,045	0,026	0,170	0,043	0,164	0,164	0,021	0,049	0,164	0,063	0,219	0,246	0,089	0,241	0,170	0,041	0,355
	MSE		0,104			0,077			0,094			0,096			0,092			0,095			0,236			0,100	
	R^2		0,908			0,918			0,889			0,891			0,895			0,908			0,836			0,910	

Table 15 The list of variables generated from the SystemFinder model.

Time		Hour of the year	
T_amb	°C	Ambient Temperature	
D_ele	kW	Demand Electricity	Demand Electricity profile for the full year (Hourly)
D_hw	kW	Demand Hot Water	Demand for Hot Water for full year (Hourly)
D_sh	kW	Demand Space Heating	
Gas_consum	kW	Gas consumption Total	Total Hourly Consumption of Gas
Gas_sh	kW	Gas Space Heating	Hourly Gas Boiler Production for Space Heating Demand
Gas_hw	kW	Gas Hot Water	Hourly Gas Production for Hot Water Demand
HP_consum	kW	Total Heat Pump Consumption	
HP_consum_il	kW	Heat Pump Consumption from Internal Load	PV or Battery
HP_consum_grid	kW	Heat Pump Consumption from Grid	
HP_sh	kW	Heat Pump Generation for Space Heating	
HP_hw	kW	Heat Pump Generation for Hot Water	
PV_total	kw	PV generation Total = Local Irradiation	
PV_total_2_load	kW	PV generation to Internal Load (Household consumption or storage)	
PV_feed_in	kW	PV generation to grid	
PV_prod	kW	Total PV production	
Ele_grid	kW	Electricity Grid	Electricity Consumption from Grid
Ele_il	kW	Electricity Internal Load	Electricity Consumption from Internal Load (Battery, PV)
Storage_hw	kW	Storage Hot Water	
Charging_hw	kW	Charging of Hot Water storage	
Discharing_hw	kW	Discharging of Hot Water storage	
T_sh_flow	°C	Temperature at the inlet of space heating	
T_sh_return	°C	Temperature at the outlet of space heating	
T_sh_hybrid	°C	Temperature at the inlet of space heating (from Heat Pump)	
Ratio	-	Ratio gas-electricity price	
COP	-		
[+1]	-	Time step +1	

Table 16 List of features for each model of Configuration 1.

Random Forest and Neural Network algorithms:	
Model 1.1	*Ambient Temperature (T_amb)*
Model 1.2	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Space Heating Operation (sh_on_off)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)*
Model 1.3	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Demand Hot Water (D_hw)* *Storage Hot Water (Storage_hw)* *Coefficient of Performance for hot water (COP1_hw)* *Coefficient of Performance for space heating (COP1_sh)* *Space Heating Operation (sh_on_off)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw_1)* *Demand Hot Water Pattern 'D_hw_-23h)* *Rolling means Storage Hot Water last three hours (Storage_hw_3h)* *Rolling means temperature at the inlet of space heating last three hours (T_sh_flow_3h)*
LSTM algorithm:	
	Ambient Temperature (T_amb) *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Gas Boiler Consumption (Gas_consum) – predicted value* *Heat Pump input power (HP_consum) – predicted value*

Table 17 List of features for each model of Configuration 2.

Random Forest and Neural Network algorithms:	
Model 2.1	*Ambient Temperature (T_amb)*
Model 2.2	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Space Heating Operation (sh_on_off)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Day Night (D/N)* *Level of battery charge (Battery_level)*
Model 2.3	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Demand Hot Water (D_hw)* *Storage Hot Water (Storage_hw)* *Coefficient of Performance for hot water (COP1_hw)* *Coefficient of Performance for space heating (COP1_sh)* *Space Heating Operation (sh_on_off)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw_1)* *Demand Hot Water Pattern 'D_hw_-23h)* *Rolling means Storage Hot Water last three hours (Storage_hw_3h)* *Rolling means temperature at the inlet of space heating last three hours (T_sh_flow_3h)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Day Night (D/N)* *Demand of electricity in last 6 hours (D_ele-1, D_ele-2, D_ele-3,D_ele-4, D_ele-5, D_ele-6)* *PV Generation in last 6 hours (PV_total-1, PV_total-2, PV_total -3, PV_total -4, PV_total -5, PV_total -6)* *PV Generation for the predicted hour in last 3 days (PV_total_23, PV_total_47, PV_total_71)*
LSTM algorithm:	
	Ambient Temperature (T_amb) *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Gas Boiler Consumption (Gas_consum) – predicted value* *Heat Pump input power (HP_consum) – predicted value*

Table 18 List of features for each model of Configuration 3.

Random Forest and Neural Network algorithms:	
Model 3.1	*Ambient Temperature (T_amb)*
Model 3.2	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Space Heating Operation (sh_on_off)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Day Night (D/N)* *Level of battery charge (Battery_level)*
Model 3.3	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Demand Hot Water (D_hw)* *Storage Hot Water (Storage_hw)* *Coefficient of Performance for hot water (COP1_hw)* *Coefficient of Performance for space heating (COP1_sh)* *Space Heating Operation (sh_on_off)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw_1)* *Demand Hot Water Pattern 'D_hw_-23h)* *Rolling means Storage Hot Water last three hours (Storage_hw_3h)* *Rolling means temperature at the inlet of space heating last three hours (T_sh_flow_3h)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Day Night (D/N)* *Level of battery charge (Battery_level)* *Demand of electricity in last 6 hours (D_ele-1, D_ele-2, D_ele-3,D_ele-4, D_ele-5, D_ele-6)* *PV Generation in last 6 hours (PV_total-1, PV_total-2, PV_total -3, PV_total -4, PV_total -5, PV_total -6)* *PV Generation for the predicted hour in last 3 days (PV_total_23, PV_total_47, PV_total_71)* *Level of battery charge in last three hours (Bat_char1, Bat_char2, Bat_char3)*
LSTM algorithm:	
	Ambient Temperature (T_amb) *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Level of battery charge (Battery_level)* *Gas Boiler Consumption (Gas_consum) – predicted value* *Heat Pump input power (HP_consum) – predicted value*

Table 19 List of features for each model of Configuration 4.

Random Forest and Neural Network algorithms:	
Model 4.1	*Ambient Temperature (T_amb)*
Model 4.2	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Space Heating Operation (sh_on_off)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Day Night (D/N)* *Level of thermal storage (PS_level)* *Thermal storage charging (PS_char)* *Thermal storage discharging (PS_dischar)*
Model 4.3	*Ambient Temperature (T_amb)* *Temperature at the inlet of space heating (T_sh_flow)* *Demand Hot Water (D_hw)* *Storage Hot Water (Storage_hw)* *Coefficient of Performance for hot water (COP1_hw)* *Coefficient of Performance for space heating (COP1_sh)* *Space Heating Operation (sh_on_off)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw_1)* *Demand Hot Water Pattern 'D_hw_-23h)* *Rolling means Storage Hot Water last three hours (Storage_hw_3h)* *Rolling means temperature at the inlet of space heating last three hours (T_sh_flow_3h)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Hour of the day (Hour)* *Level of thermal storage (PS_level)* *Thermal storage charging (PS_char)* *Thermal storage discharging (PS_dischar)* *Demand of electricity in last 6 hours (D_ele-1, D_ele-2, D_ele-3,D_ele-4, D_ele-5, D_ele-6)* *PV Generation in last 6 hours (PV_total-1, PV_total-2, PV_total -3, PV_total -4, PV_total -5, PV_total -6)* *PV Generation for the predicted hour in last 3 days (PV_total_23, PV_total_47, PV_total_71)* *Level of thermal storage in last three hours (PS_level1, PS_level2, PS_level3)*
LSTM algorithm:	
	Ambient Temperature (T_amb) *Temperature at the inlet of space heating (T_sh_flow)* *Coefficient of Performance for space heating (COP1_sh)* *Coefficient of Performance for hot water (COP1_hw)* *Cost of HP operation for space heating (costsh_1)* *Cost of HP operation for hot water (costhw1)* *Storage Hot Water (Storage_hw)* *Demand Hot Water (D_hw)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Demand Electricity (D_ele)* *PV generation (PV_total)* *Level of thermal storage (PS_level)* *Gas Boiler Consumption (Gas_consum) – predicted value* *Heat Pump input power (HP_consum) – predicted value*