NILM with Low Carbon Technologies

Jichao Yang

Master of Science

School of Informatics

University of Edinburgh

2023

Abstract

Non-intrusive load monitoring (NILM) is a popular home energy monitoring solution that uses smart meter data to predict the operation of the target appliances in households. The deep neural network-based NILM approach has better performance, but there is no evidence that it is effective for households with low-carbon appliances installed. In this project, we evaluated and improved the applicability of existing NILM techniques to heat pumps, a type of low-carbon appliance. Firstly, we extended the IDEAL dataset using the DECC RHPP dataset to obtain a dataset of simulated households that included virtual heat pumps. Secondly, we evaluated the performance of the NILM model on the new dataset and compared the results with the baseline to observe the impact of the new data on the model predictions. Finally, we retrained the model with the new data and evaluated it again to observe the impact of the new data on model training. The results show that the original model has poor prediction performance on the heat pump data. After retraining, this prediction performance can be improved to some extent, but it still cannot reach the level of handling the heat pump-free dataset. However, the addition of heat pump data did not contaminate the model training process.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jichao Yang*)

Acknowledgements

I would like to thank the project supervisor, Dr. Nigel Goddard, for his supervision and guidance throughout the project. I would like to thank the researchers Jonathan Kilgour and Elaine Farrow for providing me with the materials needed for my experiments and related help. In addition, I would like to thank my family and friends for the support they provided me during this period.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Nowadays, environmental protection and energy conservation and emission reduction are topics of widespread concern. For the short term, saving energy can bring economic benefits to individuals or enterprises; for the long term, paying attention to conservation and protection while developing natural resources can promote the sustainable development of the ecosphere. In practice, energy conservation requires everyone to start with their daily lives, e.g. individuals and families can endeavour to reduce energy consumption in their households. To achieve this, efficient and accurate energy monitoring is in demand.

Non-intrusive load monitoring (NILM) [1] is a popular energy monitoring solution. Typically, a NILM system is deployed in a user's home and connected to the home's smart electric meter to obtain the home's electric mains energy consumption data. The system uses this information to predict the energy consumption and operating status of specific appliances in the home in real time. In contrast to traditional energy monitoring solutions, it does not need to monitor each appliance individually, but only needs to obtain the total smart meter data at regular intervals, which is why it is called "non-intrusive". Therefore, it is easy to implement and cost-effective for the user. To summarise, the spread of this technology makes it easier for people to monitor energy in the homes.

However, the single-channel blind source separation (BBS) task [2] faced by NILM technically is more challenging than expected. Nowadays, popular solutions to the NILM problem introduce machine learning methods. Specifically, developers input smart data and energy consumption data of several target appliances to the machine

learning system in order for the model to perform supervised learning from these data. Although current NILM systems using machine learning methods have demonstrated good generalisation performance [3], there is still no evidence that the generalisation performance of current models will not be affected when more new appliances are introduced. For example, no deep neural network-based NILM method has yet been shown to maintain good performance with the addition of devices that adopt low-carbon technologies.

It is worth noting that research in NILM technology has been driven by the need for applications. One of its key applications is to monitor the electricity consumption of the elderly and vulnerable groups in order to provide assistance at the right time. The reliability of NILM technology in deployment is particularly important as it relates to the security of users' health and property. Therefore, we have sufficient motivation to evaluate the effects of equipment employing low carbon technologies, represented by heat pumps, on existing NILM technologies in this project.

1.2 Problem Statement

As a new appliance, heat pumps may pose the following potential pitfalls to existing NILM technologies:

1. Heat pumps have a different pattern of energy consumption than traditional appliances. When this unknown pattern appears in new data, the prediction of the machine learning system may be misguided as the pattern may be similar or have a similar trend to existing appliances.

2. Heat pumps may have a more flexible control strategy. This control strategy may lead to frequent changes in its energy consumption pattern, for example when the operating state is divided into multiple gears, or when there is frequent feedback regulation. Such variations may increase the complexity of the energy data, thus making the task of distinguishing between different loads challenging for machine learning systems.

In summary, the introduction of these low-carbon devices with unique and flexible energy consumption patterns may have a misleading effect on the system's prediction of the target appliances, which could seriously affect the reliability and accuracy of the NILM technology. Therefore, we would like to answer the following questions in our study:

1. Whether the introduction of heat pump data affects the performance of existing NILM models and whether this impact is significant.

2. How the predictive performance of the model for the target device will change after retraining with the new data.

1.3 Achieved Result

First, in this work, we augmented the home energy records in the IDEAL dataset with heat pump data from the DECC RHPP dataset. After adapting the data to the needs of the NILM task, we created two datasets for training and evaluation: the heat pump-free energy dataset and the simulated energy dataset with virtual heat pumps.

Next, we evaluated the two datasets separately using the heat pump-free model. The evaluation results of the two datasets were compared to investigate the impact of the newly added heat pump data on the original model. We find that the performance of the decomposition task performed with the new data is disturbed to varying degrees for different target appliances.

Finally, we retrained the NILM model using the new simulation dataset. We evaluated the performance improvement of the new model for the decomposition task, as well as the impact of training with the new data on the performance of the original model.

1.4 Dissertation Outline

In the rest of the dissertation, the content is divided into four main sections: In Chapter 2, we provided a brief introduction to the background of the NILM methodology and low carbon devices for which machine learning has been used, and described in detail the two main datasets that we have used in our experiments. In Chapter 3, we explain the overall design and technical details of this experiment, with the two main focuses being our approach for processing the data and the methodology for evaluating the results. In Chapter 4, we provide a detailed description of the entire experimental process, which includes a chronological record of the procedure, the results obtained from the experiment, a critical analysis of the results, and further exploration based on the results. In Chapter 5, we summarise the findings of the study and present the limitations of the work and future directions.

Chapter 2

Background

2.1 Related Work

As mentioned above, Non-intrusive load monitoring (NILM) is an energy monitoring solution that identifies the energy consumption of multiple appliances by disaggregating the readings from a single mains meter in a household, dating back to the early 1980s [4]. The application of this technology in society gained momentum after smart meters were popularised. Research on this technology was further accelerated after machine learning methods were found to provide excellent solutions to the BBS problem [2]. In November 2022, the 6th International Workshop on Non-Intrusive Load Monitoring was organised, where the latest trends in the technology and the industry were widely discussed by the participants [5].

Researchers have tried to solve the problem using methods such as Hidden Markov Models [6]. These methods are characterised by learning from features, so features such as state changes in energy consumption or duration of states need to be artificially organised and input to the machine. This limitation hinders convenient implementations.

After the development of deep neural networks [7], it became a common approach to solve NILM tasks because it does not require feature engineering and enables the model to learn directly from the data and achieve good performance. Depending on the inputs and outputs of the model, these solutions are classified into ”sequence to sequence” methods (S2S) [8] and ”sequence to point” (S2P) [2] methods. Due to the translational invariance and temporal order of energy consuming data, the introduction of convolutional neural networks [9] and recurrent neural networks [10] added an infinitely strong prior to the features of the data, which made the networks significantly better at handling every type of data that matched these features. On top of this,

researchers have further progressively improved model performance through various complementary methods. In 2018, the NILM network architecture proposed by Nigel Goddard's team significantly improved computational efficiency [3]. By now, deep neural networks have become a powerful support for the practical application of NILM technology.

2.2 Low Carbon Technologies and Heat Pumps

Low-carbon (LC) technologies are technologies that aim to reduce emissions of greenhouse gases or other pollutants, and their development history can be traced back to the early 1970s [11]. Electric vehicles, equipment that reduces industrial emissions, and home appliances that conserve electricity all fall into this category. Usually, household appliances using LC technology introduce innovative energy regulation methods, such as automatic load sensing, in order to cut energy consumption as much as possible.

Heat pumps are heat transfer devices that perform both heating and cooling functions, often replacing gas boiler-based temperature regulation systems in low-carbon homes. The technology was first invented in the mid-19th century [12] and has been widely used for temperature regulation in buildings since the 20th century [13]. Unlike traditional thermoregulation devices, heat pumps work by absorbing heat from a heat source and releasing it to a target location with the help of refrigerant fluid circulation [14]. Air-source heat pumps use outdoor air as their heat source, while ground-source heat pumps use buried pipes to absorb heat from the ground. Typically, the energy consumption of a heat pump is measured by the ratio of its heat output to its electrical input, which is known as the coefficient of performance (COP) of the heat pump [15].

2.3 Datasets

In this study, the datasets we use are the IDEAL dataset [16] and the DECC RHPP dataset [17]. The IDEAL dataset, which contains the time series of the total energy readings in the target households without heat pumps, and the time series of the energy readings for each of the individual target appliances, plays the role of the basic training data for the NILM method. We use this data to train the baseline model for electricity disaggregation and use this dataset as the basis for creating the simulation dataset containing heat pumps. The DECC RHPP dataset, which contains the energy consumption

time series data for the heat pumps, our target low-carbon technology appliances, will be used as an expansion data source for building the simulation dataset.

2.3.1 IDEAL Dataset

The IDEAL Household Energy Dataset [16] is a multifunctional dataset containing a large amount of real household data on electricity, gas and other energy sources. The data related to household energy consumption in this dataset was collected from 255 UK households over a 23-month period ending June 2018. As the dataset was created with the aim of broadly examining household energy use patterns and motivations, and with a balance of versatility to provide as universal support as possible for subsequent research, data were collected as comprehensively and exhaustively as possible for each of the target households within the collection period. Specifically, with the consent of the subjects, various types of sensors were installed in the homes to collect data. The sensors were able to capture the true energy consumption of the homes for the duration of the experiment, as the subjects maintained a normal life throughout the test period and the installation of the sensors had no feedback effect on the operating status of the appliances in the homes. The energy information collected by these sensors includes not only electricity data, but also gas, temperature and other indicators. All data were recorded as time series. In addition to sensor readings, metadata and survey data on household information are also important components of the dataset. Since our study only deals with electricity disaggregation, we only need to use the subset of sensor data in this dataset that deals with electricity use by appliances.

In addition, the 39 households in this dataset form a data enhanced group, which was specifically prepared for potential follow-up NILM studies and is of particular interest to us. The main difference between the enhanced group data and the regular data is that these households additionally recorded data on the electricity consumption of a wide range of appliances in the home, in addition to the total power consumption. These data are measured by a large number of additional specialised sensors, which require a higher level of fitness and technological environment in the target homes, and therefore the number of homes able to provide these data is significantly reduced compared to the total number of homes participating in the experiment. These device-specific data are not necessary for a machine learning task where the goal of the task is to make predictions about the potential future direction of a sequence based on energy consumption patterns or regularities. However, these data are crucial for our study because in the NILM task

we have to use the energy consumption of individual appliances as training labels for supervised learning, i.e., the output of the readings disaggregation. Therefore, we must use a subset of the augmented set of this dataset.

2.3.2 DECC RHPP Dataset

The DECC RHPP dataset [17] records the operation of 700 heat pumps during the period of data collection from October 2013 to March 2015. The project to build this dataset was initiated by the Department for Energy and Climate Change (DECC) and the Renewable Heat Premium Payment (RHPP) grant scheme to support a research programme of 14,000 trial heat pumps installed between 2009 and 2024 to investigate and improve the actual deployment performance of these heat pumps. For low carbon equipment, the energy consumption during the operating hours is an important part of the performance evaluation and is the main subject of recording in this dataset, therefore the information provided by this dataset fits well with our project needs.

Unlike the more versatile IDEAL data, the DECC RHPP dataset focuses on recording the operation of a single piece of equipment, the heat pump, including energy consumption information. Furthermore, the data is recorded around the metrics of the target heat pump, rather than the target household, and thus the data can be easily used in our project as a simulation extension to the IDEAL dataset. In the simulation dataset, the heat pump can be considered as one of the individual appliances recorded in the IDEAL enhanced set.

Another positive aspect of the DECC RHPP data is the existence of a data-cleaned version, B2. For raw data acquired directly by the sensors, RAPID-HPC was commissioned to perform sensible completions and pruning of missing and duplicated time-series readings, as well as systematic removal of defective measurements caused by the equipment. The systematic removal of defective measurements due to equipment was carried out. After filtering, 418 heat pumps that met the requirements were exported as cleaned data. These processes allowed us to apply the dataset more directly.

2.4 Network Model

As described in 2.1, deep neural network models have now demonstrated superior performance and strong applicability in NILM tasks. Currently, the dominant deep neural network models being used for NILM tasks all adopt similar problem modelling

and infrastructure. In the NILM task, our goal is to acquire a mapping to predict the energy consumption time series of individual target appliances in a house, provided the total power consumption time series of the house. Unlike the traditional idea of single-channal blind source separation (BBS), in the understanding of the task objective, the neural network treats it as a mapping of multiple sets of single input sequences to single output sequences, rather than a mapping of one set of single input sequences to multiple sets of single output sequences, i.e., the so-called ”separation”. Specifically, for each target appliance, a separate model will be learned. At this point, for each individual point in the input time series, the total power consumption is considered to be the sum of that target appliance's energy consumption and the energy consumption of other appliances (including noise). The model learns a direct mapping of this aggregate sequence to the target sequence from the data, without taking into account the joint effects or interactions of multiple appliances.

After the neural networks were introduced to the BBS problem, the importance of the principle problem of power disaggregation decreased, since the learning of such mappings was always based on black-box systems. Instead, the performance of predictive models on applications became the focus. In this project, we use the Fully Convolutional Network (FCN) model proposed by Cillian Brewitt and Nigel Goddard for solving the NILM problem [3]. As one of the most advanced models for the NILM problem, it is able to combine high computational efficiency with high performance for the prediction task, making it well suited for real-world deployments of NILM systems.

The superiority of the model can be summarised by its removal of redundant processing of sliding windows and the use of dilated convolution. The focus of this project was on the evaluation of the dataset rather than on model improvement, so the use of a high-performance existing model was able to satisfy our needs for training and prediction. For more detailed information about the technical principles of the model, please read the original paper [3]. For the new dataset, we used the same network architecture as the baseline model to maintain consistency of evaluation. In order to work with the model, the preprocessing of the dataset can be divided into two parts:

1. The format of the data is modified in order to make it match the input and output formats required by the model.

2. The data that is fed into the model is normalised as well as windowed in order to enhance the learning performance of the neural network.

The former is one of our main tasks in the data processing phase.

Chapter 3

Methodogy

3.1 Experiment Design

The experiments of this project can be divided into three phases.

The first phase is the dataset preparation phase. Within this phase, the main work is the processing of the IDEAL dataset and the heat pump dataset. Our goal is to evaluate the predictive performance of the NILM model in homes with heat pumps, so we need a dataset consisting of a collection of homes without heat pumps as a baseline for the evaluation, as well as a dataset consisting of homes with heat pumps.The IDEAL data can fulfil the former requirement, but we do not have a complete data case from homes with heat pump installations, and therefore we need to create that simulation dataset. The specific way to do this is to augment the IDEAL dataset with information on heat pump installations from the DECC RHPP data to simulate the case where heat pumps are installed in these homes. Firstly, the two datasets will be processed separately into a form suitable for network modelling, which involves extracting the subset we need from the huge dataset and adapting its format. Secondly, the forms of the two datasets will be unified, the key aspect of which is the unification of the temporal resolution. Finally, the two datasets are merged, i.e., the IDEAL dataset is expanded with the heat pump dataset. Each dataset consists of several households. The data for each household contains time series of electrical energy consumption for several target appliances during the data collection period, one of which is a sequence of electric-mains readings. In the dataset containing the heat pump data, an additional data sequence for a virtual heat pump was added to the collection. After this stage, we created two datasets prepared for the NILM task, namely the real home energy dataset without heat pump data and the simulated home energy dataset with heat pump data.

However, there are prerequisites for our heat pump simulation through the method described above. A detail worth noting is that when we assume that the simulated households use heat pumps instead of the pre-existing heating appliances, we not only need to add the heat pump data to the energy consumption of the households, but we also have to remove the energy consumption of the pre-existing heating appliances from the sum, which can be a complex issue. For this work, the households in the IDEAL dataset we selected were previously heated with gas boilers and there was no electricity supply involved, so the energy consumption that needs to be removed is zero, and therefore we can avoid this problem.

The second stage is to train and evaluate the model on the unmodified IDEAL dataset. Within this phase, we train with data sets that do not contain heat pump data. The homes were divided into a training set, a validation set and a test set. The use of homes rather than specific data for the division is to ensure that the evaluation is consistent across the two datasets. An early stopping strategy based on the performance evaluation of the validation set is adopted during model training, which shrinks the training duration while preventing the risk of overfitting. For each target appliance, one predictive model will be generated. We tested the models on each of the two datasets and compared the performance of the models. If, for a particular appliance, the model's prediction performance in the home with the simulated heat pump installed is inferior to the other dataset, the model is poorly adapted to the new data, i.e., the installation of the heat pump in the home interferes with the model's ability to predict the energy consumption of the target appliances from the total energy consumption of the home.

If the model obtained in the absence of the heat pump training data proves to be inapplicable to homes with heat pumps, the experiment will proceed to the third phase. Within this phase we will attempt to train a new model on the simulated dataset containing the heat pump data. Again, we will evaluate the predictive performance of the model on both datasets. In this phase we will look at how the introduction of new training data affects the model's ability to generalise. Specifically, if training the model on new data improves its performance, what may be happening is that more information in the training phase improves the model's adaptability to data with new features; if the model's performance decreases instead, what may be happening is that the introduction of new data makes it more difficult for the model to learn the correct features, i.e., the introduction of the heat pump " pollutes" the training data.

3.2 Processing of the IDEAL Dataset

Data processing plays an important role in this project, and the IDEAL dataset is the base dataset for subsequent training as well as merging work. Its processing will be divided into three main segments:

1. Dataset-level processing. In this session, we filter the subset of data that we need for our experiments from the huge dataset. As mentioned above, we first filter out the data subset of the augmented group of families and take only the electricity data records.

2. Sequence-level processing. Both the input and output of the model are one-dimensional time series of electricity data. Therefore, we need to extract the data sequence that matches the format of the model. The information we need is the total electricity consumption of the target households over time and the electricity consumption of each target appliance. It is important to note that the former is not equivalent to the sum of the latter, as the total electricity consumption also includes any power-consuming devices present in the household other than the target appliances, as well as noise due to factors such as measurements. Therefore, we keep only these one-dimensional data series from the dataset, which are the power data with the energy consumption of individual appliances.

3. Data point level processing. In this session, we clean and fix the data series to make the raw data complete. This includes resampling of the time series and filling of missing data.

3.3 Processing of Heat Pump Dataset

The role of the heat pump data in this project is to expand the IDEAL dataset to obtain a home simulation dataset containing heat pump data. Therefore, the processing of the heat pump dataset can be regarded as a process of creating virtual heat pumps. The complete portrayal of the properties of these virtual heat pumps is the sequence of their power during the measurement time. The same three aspects of heat pump data processing apply to the IDEAL dataset. Firstly, the heat pump dataset has a relatively simple structure, all of which are stored as parallel tables indexed by heat pumps. We chose dummy heat pumps equal to the number of target households selected in the IDEAL dataset. These heat pumps were matched one-to-one with the target household

in preparation for playing the role of the virtual heat pump that was ported to that household. Secondly, of all the statistics for that heat pump, only the sequence of electricity readings was retained. In this way, the data for each heat pump would obtain the same one-dimensional time series form as the IDEAL electrical readings. Finally, the modification of data points is greatly simplified because that original dataset has already been cleaned. Our task is to perform unit conversion and resampling of the cleaned data to match the temporal resolution we require.

It is worth noting that the design of the simulation data is a relatively open problem. Our needs are to evaluate the performance of the model in homes with heat pumps and to improve that performance by retraining it again, so we make some modifications to the heat pump time series to meet our needs while retaining the characteristics of the data. Here I present five approaches to create new simulations based on the original heat pump data.

3.3.1 Realistic Heat Pump Simulation

The first approach is that we do not make any changes to the original sequence of electricity readings from the heat pump and use it directly for merging with the corresponding household data. This is the easiest and most straightforward option to implement, and it is also completely feasible because the actual data of the virtual heat pump is completely independent from the real household energy consumption data. The transplantation of the virtual heat pump would not affect the energy consumption of other independent appliances, but would only change the total power consumption.

However, I soon discovered a potential problem. Due to the seasonal nature of the operation of the heat pump unit, the time period in which the heat pump operates is very sparse throughout the data series. This leads to the fact that in most cases the heat pumps in the households remain inactive in the new household dataset. Due to the translational invariance of the prediction model window, the data set is in fact not altered during these time periods and therefore its predictions do not change. In addition to this, there are more potential low activity operating states for the heat pump. In this case, it is likely that the model is insensitive to changes in the dataset, which will result in a performance evaluation that does not truly reflect the impact of the heat pump installation on the model.

3.3.2 High-intensity Operation Simulation

Based on the first approach, the second attempts to address its shortcomings. Here we will drastically modify the original sequence of heat pump data to weaken its sparsity. The basic idea is that we set a specific threshold for the operating energy consumption of the heat pump to filter the states in which it operates at high intensity. We extract fragments of these states as sub-sequences of the virtual heat pump's "high intensity operation". We select and assemble (or repeat) such sub-sequences to form a simulation sequence of a target length, in which the heat pump is continuously operating at high intensity in the output simulation sequence.

However, we soon realised that this destroys some of the realism of the simulation data, as such filtering amounts to " cutting off" specific parts of the heat pump's natural behaviour, leading to the creation of unnatural jumps. As neural network learning is likely to learn the pattern of the appliance's operation from changes in state, it is likely that such changes would prevent the network from learning correctly from the data.

3.3.3 Intensive Operation Simulation

The third approach tries to think in a different way to avoid the potential pitfalls of option two. Now, we have to figure out how to preserve all the useful patterns and structures present in the data while trying to reduce sparsity. Due to the black-box nature of neural networks, it's difficult to assert what patterns will play an important role, so preserving the realism of the data would be a good starting point. In the new data series, I tried to ensure that the following three occurrences could be present in the data window intercepted by the model's sensory field:

1. The heat pump is in operation.
2. The heat pump gradually starts from a silent state.
3. The heat pump is gradually switched off from the running state.

In terms of specifics, I removed only the longer periods of time when the heat pump was continuously off (i.e., when the reading was 0) and spliced and expanded the data segments in a similar manner as in the second approach. In this way, although the data still retains some sparsity, the simulation data has made the original sequence denser while retaining all the full operating cycles of the heat pump.

3.3.4 Average Operation Simulation

We soon discovered that Method 3 also posed a subsequent problem: we managed to generate a dataset that could be used to assess the impact of heat pumps on the baseline model, but it was clear that it was unsuitable to be used to train a new model because it lacked information about the data in the home when the heat pumps were not working. In a real-world deployment, heat pumps in the home would not work at such an unnaturally high intensity. Therefore, if whether the heat pump is working or not is indeed an influencing factor in the predictive performance of the model, training with this dataset is likely to result in overfitting of the model. That is, the performance of the trained new model on the new simulation dataset may be dramatically improved, but the performance on the original dataset may be reduced sharply. This is not what we expect to see in training. Therefore, we propose a more eclectic solution based on approach 3: we record the duration of heat pump operation in each subperiod, and then fill in a silent time period of the same length corresponding to the gaps between heat pumps' operation, which ultimately ensures that the times when the heat pumps are working and not working are averaged over the simulation sequence.

3.3.5 Data Enhancement Simulation

In order to address the above issues, a fifth scenario was proposed as we wanted to utilise more information for training. We realised that we do not need to keep the size of the dataset constant if the problem involves actual deployment rather than just performance evaluation, so we could introduce both the real dataset without heat pumps and the simulated dataset with intensive heat pump operation during training. At this point, the number of homes used for training becomes twice as large, including the IDEAL dataset augmented with homes and the same number of "virtual homes" augmented with virtual heat pumps. If the model is able to correctly learn features from the data when the heat pumps are operating, then training with this enhanced dataset is expected to maximise the performance of the model in real-world deployments, even when the introduction of the heat pumps has an impact on the performance of the base model.

3.4 Combining of Datasets

As mentioned above, the combining of the IDEAL dataset with the heat pump dataset means adding the heat pump as a virtual appliance to the IDEAL dataset's household

electricity records. The process involves two main actions:

1. Placing the heat pump data into the catalogue in a form consistent with the individual appliances in the IDEAL enhanced dataset. This means that we simulate the presence of heat pump appliances in the homes.

2. Using the sum of the heat pump energy consumption sequence and the electric-mains sequence to replace the latter. This means that the energy consumption when the heat pump is working becomes part of the total energy consumption at the corresponding point in time, i.e. it is overlaid with the sum of the electricity consumption of all other appliances. Such a change does not affect the real energy consumption of any other appliance individually.

In the specific operation of the merging, two key points were noted:

1. The units of measure and time resolution of the IDEAL data and the heat pump data need to be unified. In terms of temporal resolution, the electric-mains readings and appliance readings in the IDEAL dataset and were recorded at a maximum of once per second, and there were cases of omitted recordings where readings varied slowly; the appliance readings for the heat pumps were sampled at a very standard two-minute interval. For the input data to the network model, we expected a sampling interval of 10 seconds, so the IDEAL sequence was uniformly padded first to 1 Hz and then downsampled to 10 Hz; the data for the heat pumps was upsampled, i.e., padded backwards to 10 Hz.

2. Before merging, the timestamps of the two datasets should be unified. In the case of the real data simulation scenario for the heat pump data, for example, although for a set of IDEAL homes and heat pumps, both of them have data collection lengths of approximately up to 1-2 years, they cannot be directly aligned and merged. A better approach here would be to extract only the sequence of readings for the heat pump data without extracting the timestamps, since we assume that the operating state of the virtual heat pump has a translational invariance. For other simulation scenarios, we can also manually control the length of the sequences to coincide with the acquisition duration of the IDEAL data to create convenience for the experiment.

3.5 Selection of Target Appliances

There are a large number of specific appliances that were measured in the IDEAL dataset. In the experimental phase, it was not possible to train separate models for all appliances, so we chose four of these common household appliances as our target appliances, which are cooker, shower, kettle, and washingmachine. Our criteria for selecting the target appliances are as follows:

1. These appliances should be common in households and recorded a high number of times in the dataset. With a relatively small number of households in the augmented group, choosing appliances that are commonly installed in a larger number of households can provide a larger amount of data and a higher confidence level of evaluation for our model training. Ensuring the performance stability of common appliances also benefits the system being deployed in practice.

2. These appliances have functionally similar units to heat pumps. In heat pumps, the main energy-consuming units are a booster heater and a circulation pump. In the target appliances, the cooker, the shower and the kettle all have a heating unit as the main part; and the washingmachine has both a heating unit and a motor unit.

3. These appliances operate at similar power levels to the heat pumps, so their evaluation allows for maximum testing of the interference of the heat pumps with the model. Heat pumps operate at peak power readings of thousands of watts. In comparison, the target appliances all operate in the range of at least several hundred watts.

In subsequent experiments, we will train the corresponding prediction model separately for each target appliance, and our performance evaluation will always be for these appliances, even after the heat pump has been introduced. This is because what we are concerned with is whether the predictive performance of the model for the target appliances is affected by the heat pump data.

3.6 Selection of Target Households

Of the 255 families recorded in the IDEAL dataset, only 34 families in the enhanced data subset with relatively complete power data records could be used in our experiments.

We divide these households into training, validation and test sets. The distribution of target appliances is different in different households. Therefore, an important criterion for the division is to ensure that the appearances of each target appliance in the three subsets are relatively even. We ensure that at least each target appliance appears at least twice in each of the three sets to ensure the balance of the data and the generalisation performance of the model. Here, we adopt a division consistent with the baseline paper [3], as shown in table 3.1. The number of occurrences of each target appliance in each of the three subsets is shown in table 3.2. 18 Household data from the training set is used for model training. The optimisation target of the deep neural network is the error between its prediction and the real appliance energy consumption. 8 validation sets are used to evaluate the generalisation performance of the model in real time during the training process. The early stop mechanism is triggered to prevent overfitting when the model's performance improves on the training set but no longer increases on the validation set. 8 test sets are used to generate evaluation metrics for the model during the evaluation phase. The prediction errors of the model on these electric-mains data were aggregated and counted to measure the generalisation performance of the final model for each appliance.

	Homes
Training	62, 65, 96, 105, 106, 128, 136, 145, 162, 168, 169, 175, 228, 231, 238, 255, 263, 328
Validation	61, 63, 139, 140, 146, 208, 225, 268
Test	73, 171, 212, 227, 242, 249, 264, 266

Table 3.1: Division of the data set

Appliance Type	Training	Validation	Test	Total
Cooker	5	1	2	8
Shower	6	2	3	11
Kettle	8	3	4	15
Washingmachine	13	3	5	21

Table 3.2: The count of each appliance type

Here, the use of cross-validation methods can lead to stronger evaluation results. Specifically, we can choose a variety of different ways of dividing the training, validation, and test sets, and then train/evaluate each set and average the evaluation results.

The advantage of this method is that it can substantially reduce the randomness associated with the choice of datasets, thus improving the confidence of the results. The disadvantage of this method is that when the amount of data is small, it is difficult to find a variety of divisions that can balance the number of appliances, which leads to a potential degradation of training quality.

3.7 Evaluation Metrics

The evaluation of the model's predictive performance is an important part of this project. The electric-mains data from the target dataset is fed into the model we wish to evaluate. The predictions obtained from forward propagation are compared to real data for a specific appliance. These evaluation metrics are categorised as energy metrics and activation metrics. The quantified energy sequences obtained from the model predictions can be used directly in the calculation of the energy metrics. The difference is that, before the calculation of the activation metrics, the energy sequences need to be converted into binary activation sequences, i.e., labels that determine whether the appliance is in operation or not, by means of activation detection.

An important application of the product is to monitor the electricity consumption of elderly and vulnerable people in order to decide when to provide help. Therefore the need for model deployment is to determine the working status of individual appliances in the home (e.g., whether cookers are turned on abnormally), rather than how much power they are specifically consuming. That is, we are more concerned with activation metrics than energy metrics.

3.7.1 Energy Metrics

The energy evaluation metrics are aggregated in different forms to calculate the prediction error between the predicted energy sequence and the true energy sequence. Data points in the true energy sequence are denoted by x_i; data points in the predicted energy sequence are denoted by $\hat{x}_i$; and the length of the sequence of readings is denoted by n. There are four energy metrics for prediction error: (1) Mean Absolute Error (MAE), as in equation 3.1. (2) Signal Aggregate Error (SAE), as in equation 3.2. (3) Match Rate (MR), as in equation 3.3. (4) Normalised Disaggregated Error (NDE), as in equation 3.4.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.1}$$

$$SAE = \frac{|\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \hat{x}_i|}{\sum_{i=1}^{n} x_i} \tag{3.2}$$

$$MatchRate = \frac{\sum_{i=1}^{n} \min(x_i, \hat{x}_i)}{\sum_{i=1}^{n} \max(x_i, \hat{x}_i)} \tag{3.3}$$

$$NDE = \frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{\sum_{i=1}^{n} x_i^2} \tag{3.4}$$

Since the timestamps of the real energy series do not exactly overlap with the electric-mains data, whereas the timestamps of the appliance prediction energy series are the same as the electric-mains data, data points with no real values for that time period were excluded from the evaluation.

Each sequence is derived from the readings of a specific appliance in a specific household, so the above predictors also correspond to specific appliances in a specific household. Therefore in order to measure the total predictive performance of the target appliance across all households, we also compute two aggregated energy metrics:

1. Average statistics: first obtain statistics for each household and then average these statistics.

2. Aggregate statistics: first the series of all different households are aggregated and then the evaluation results are computed for this collection of series.

3.7.2 Activation Detections

The activation of an appliance is defined as the period of time during which it performs its major function. The purpose of activation detection is to convert the sequence of values obtained from the regression into a sequence of activations that determines whether the appliance is operating or not. This conversion is achieved through a series of thresholds that are predefined for the appliance:

1. Minimum activation power. This is a common metric, and when the appliance power reading exceeds this threshold, the appliance is determined to be in operation.

2. Minimum and maximum duration of activation. When the duration of an activation is below the minimum threshold, the activation is removed; above the maximum threshold, it is replaced with a shorter activation period.

3. Minimum time interval between adjacent activations. When the time interval between two activations is below the threshold, remove that too short offs and subsequent activations.

The specific rules for determining that a target appliance is active are shown in table 3.3, where the parts except for the heat pump are provided by the IDEAL code base.

Appliance Type	Minimum Activation	Minimum Duration	Maximum Duration	Minimum Interval
Cooker	200	10	10000	120
Shower	300	20	5000	30
Kettle	1000	10	600	10
Washingmachine	50	1200	30000	600
Heatpump	100	10	5000	10

Table 3.3: Specific appliances activation rules

When we train and evaluate heat pump models in augmented data, we also need to develop appropriate activation rules for heat pumps. For this purpose, we conducted an exploratory data analysis of the heat pump data.

Firstly, we counted the power distribution of the data points in the heat pump sequence, and the results are shown in figure 3.1, where logarithmic coordinates were used for the vertical axis. The results showed that 100 Watt is a significant cut-off point for the power statistics, and 200 Watt is another cut-off point, but not as significant as the former. Here we consider the characteristics of heat pumps. When conventional appliances are in operation, energy consumption is generally maintained at a relatively constant level. Unlike them, heat pumps are different because their energy consumption changes frequently and is in a large range of instability when they are in operation, depending on the need for heat transfer. Therefore, in order to ensure that the normal operation of the heat pump is not neglected while filtering out the noise, we set the minimum activation power threshold to 100.

Secondly, we aggregated the statistics of the features related to the length of the heat pump data sequence. When the activation power is set to 100, the activation

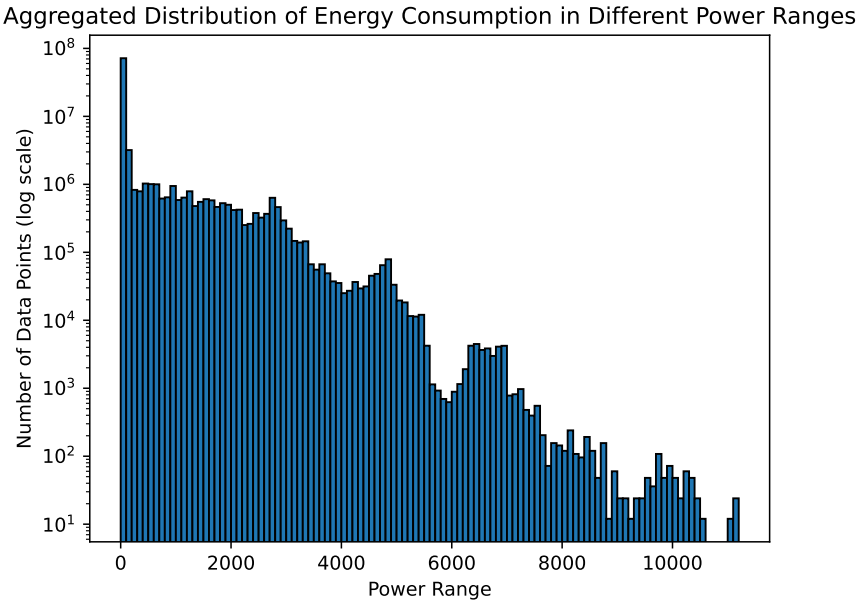

Figure 3.1: The Power Distribution of the Data Points in the Heat Pump Sequences

segment length of the sequence is a minimum of 12 and a maximum of 45504. When the activation power is set to 200, the activation segment length of the sequence is a minimum of 12 and a maximum of 45492. Here, we again consider the characteristics of heat pumps. When a heat pump is used to maintain a constant indoor temperature, it tends to be turned on or off frequently in response to heating or cooling demands, and the duration of this operation is very flexible. Therefore, we should try to relax the restrictions on the operating cycles of heat pumps. Therefore, the minimum and maximum duration thresholds are set to 10 and 50,000, respectively; the minimum interval is set to 10.

3.7.3 Activation Metrics

After activation detection, true and predicted sequences described by state change labels (on or off) are used to compute the evaluation results. Activation evaluation replaces the error metric in energy evaluation with a correctness metric. The activation evaluation metrics include Recall, Precision, and F1 score. Similar to the evaluation of a general binary labelled machine learning model, Recall indicates how much of all the data in the activated state is correctly predicted, and Precision indicates how much of all the data predicted to be activated matches the true state. F1 score is used to combine the performance in Recall and Precision. They are formally expressed as equation 3.5,

equation 3.6 and equation 3.7.

$$precision = \frac{truepositives}{truepositives + falsepositives} \tag{3.5}$$

$$recall = \frac{truepositives}{truepositives + falsenegatives} \tag{3.6}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.7}$$

Chapter 4

Experiment Description and Result Analysis

4.1 Experiment Environment

The dataset processing and evaluation part of the experiment was run on a Windows PC. The Python version was 3.10. We used the pandas library to perform the reading, processing, and writing of all the csv and hdf files from the datasets and the cached data.

The training and prediction part of the experiment was realised by remote access to a Linux PC on the DICE system provided by Informatics of the University of Edinburgh. The host provided hardware computing support with four NVIDIA Titan X GPUs. The code used to train the deep neural network models and obtain predictions used the deep learning library keras2.4.3 based on the framework tensorflow2.4.1. The Python version was 3.9.

4.2 Creation of the Heat Pump-free Dataset

The first task of the project was to build a heat pump-free dataset based on the IDEAL dataset. This process mainly involved screening, pre-processing and format conversion of the data. During data collection, the time series of readings from each sensor in the IDEAL dataset were stored in separate csv files. The sensor information can be indexed by its file name, which includes the household the sensor belongs to, the numerical identification of the location, the sensor category, and the appliance function being measured.

After initially filtering out the broad categories of electricity data, we filtered the sensors by their household serial numbers and functional identifiers to get a subset of data for all target households and target appliances. This is done by extracting the target household fields and target appliance fields from the file names. The target households include the 34 augmented group households that were discussed in 3.6. The variable "install_type" in the household list indicates whether the household has a standard or enhanced system installed. The target appliances include the electric-mains readings named "electric-combined" and the data for the four target appliances discussed in 3.5: cooker, shower, kettle, and washingmachine. The size of the dataset is significantly reduced after filtering.

Inside each csv file, the raw data format presents concise two columns. The first column records the detailed sampling timestamp. The raw sampling interval is 1 s. The second column records the sensor reading at the corresponding time, which is in watts. When reading the file, we use the time data in the first column as row labels in the dataframe. Since then, the data format is transformed into a 1-dimensional sequence of readings. For each sequence of sensor readings, we perform the following data preprocessing:

1. Remove unreasonably spiky data. For electric-mains, readings over 20,000 are set to 20,000; for appliances, readings over 4,000 are set to 0.

2. Filling of missing values. The electric-mains and appliance sampling intervals in the raw data were not strictly 1 s. Therefore, we resampled the raw data by padding forward to make the sampling intervals strictly 1 s. Specifically, the range of forward padding for the electric-mains and appliance readings were 1 min and 1 h, respectively.

3. Downsampling the data sequences at sampling intervals of 10 s. The 10 s is the uniform sampling interval that we set for the IDEAL dataset and heat pump dataset.

In practice, preprocessing of the dataset is used as part of the training process. The script first tries to read data from the pre-specified raw data directory or cached data directory. If there is no data in the cache directory, the preprocessing of the data is performed preferentially and then the preprocessed data is saved in the cache directory in the form of an hdf file. Meanwhile, the information being recorded in the filename was simplified to the id of the residence, the appliance name and the sampling interval

(default 10s). These filenames were used as a uniform index format in subsequent experiments. By setting the ”prepare-only” parameter in the training script, we can make the programme stop after caching the data without subsequent training. Since then, we have obtained these heat-pump-free datasets in the form of cached files, reaching a milestone in the dataset creation phase of the project.

4.3 Training and Evaluation of Heat Pump-free Households

After the heat pump-free dataset is ready, the next step is to train using that dataset. The training of the model is done individually for each target appliance. After running the script that was used to train the model with the target appliance as a parameter, the program takes priority to read from the cache directory all the corresponding appliance sequences of the households belonging to the training and validation sets.

4.3.1 The Second Data Preprocessing

The data used for training will undergo a second preprocessing after being fed into the model:

1. The sequence is normalised. The data points of the electric-mains are subtracted from the mean of the sequence and divided by the standard deviation of the sequence. The data points for each appliance are divided by its maximum value.

2. The sequence was windowed. The processed sequence is overlaid and intercepted into a data window that conforms to the input dimensions of the NILM model.

3. Data windows with missing appliance readings were recorded or discarded.

The main purpose of the above preprocessing is to make the data suitable for the model, to improve model performance, or to increase training speed.

4.3.2 Model Training

The FCN neural network model which was discussed in 2.4 is used for training. The number of dilation layers of the network is 9. The dilation filter size is 3. The initial filter size is 9. The number of filters is 128. The loss measure for training is the mean

square error and the optimiser used is adam. The initial learning rate is set to 0.001. The batchsize of the training data is 256. The data of each appliance is trained for a maximum of 200 epochs.

After each training epoch, the model of the specified appliance is always saved. An optimal model is additionally saved and is continuously updated. The early stop mechanism is triggered when the training error continues to decrease but the validation error no longer does. After training, the best models corresponding to the four target appliances are saved as the results of the training phase.

4.3.3 Prediction and Evaluation

After the training of the model is complete, we get the predicted readings on the test set to prepare for the evaluation. First, the prediction script is loaded with a sequence of electric-mains readings from the homes in the test set. Secondly, the sequence is pre-processed and fed into the model. The output sequence obtained from the model feed-forward is the prediction result of the target appliance. Finally, the prediction results are saved as a three-column hdf file of ” Mains Readings - Predicted Appliance Readings - True Appliance Readings”.

Subsequently, the models for each target appliance were evaluated individually on the test set. The evaluation includes the energy metrics and activation metrics discussed in 3.7. It is worth noting that prior to the activation evaluation, the appliance readings including the predicted sequences are subjected to activation detection to generate activation labelled sequences. For the model obtained by training using the heat pump-free dataset, the activation evaluation results on its own test set are shown in table 4.1. The aggregated evaluation results for multiple households take the aggregated statistics presented in 3.7.1. The complete energy evaluation statistics can be found in Appendix B. The results of this evaluation are used as a baseline for the model performance on the different datasets.

4.4 Creation and Evaluation of the Simulated Heat Pump Dataset

The next task of the project was to augment the IDEAL dataset with the DECC RHPP dataset to create a simulation dataset with heat pump households. The cleaned DECC RHPP dataset contains the operating records of 418 heat pumps during the data col-

Appliance Type	Recall	Precison	F1 Score
Cooker	0.77	0.10	0.18
Shower	0.99	0.60	0.75
Kettle	0.85	0.88	0.87
Washingmachine	0.78	0.83	0.81

Table 4.1: The aggregated activation statistics of the heat pump-free model on the heat pump-free dataset

lection period. Each heat pump's is saved as a csv file. Unlike the independence of the sequences in the IDEAL dataset, multiple sequences of readings are combined in a single heat pump file, which includes records of energy consumption, heat production, and relevant details of energy conversion.

4.4.1 Processing of Heat Pump Data

We extracted only the sequence of electrical readings with the field "E_hp" and the line labels with the time in the format "YYYY-MM-DD hh:mm:ss". Thus, a one-dimensional sequence having a consistent format with the IDEAL specific electrical reading sequence is initially obtained. Further, the data was subjected to the following processing in preparation for merging with IDEAL data:

1. Unit conversion.The DECC RHPP dataset records the energy consumed by the heat pump in Wh over a 2-minute period. We converted it to units of Watts consistent with the IDEAL dataset.

2. Upsampling. The heat pump data was sampled at 2 min intervals. we backfilled the data to get a new sequence with a target sampling interval of 10 s.

4.4.2 Data Combining

As indicated in 3.4, we selected the first 34 files in the heat pump dataset to augment the IDEAL dataset in order to simulate the installation of heat pumps in the homes.The correspondence between the IDEAL homes and the virtual heat pumps can be found in Appendix A. In order to avoid the hassle of unifying the timelines, the electric-mains sequences of each home were read as the basis for the two new sequences. Therefore, the heat pump data will be used directly for merging with the pre-processed IDEAL

data instead of the original IDEAL data. The result of the data combining will be the cached dataset with the heat pump household version mentioned in 4.2.

For the virtual heat pumps, we take the realistic data simulation approach that was discussed in 3.3.1, which is the most basic simulation approach. When using this method, we do not make any changes to the content of the heat pump readings. Firstly, we modify the length of the heat pump reading sequence by the length of the energy reading sequence. Insufficient parts are made up with zeros, while excesses are discarded. Second, the energy sequence in the base dataframe is replaced by the heat pump sequence. The result is saved as a heat pump cache hdf file. Finally, the energy sequence in the base dataframe is replaced by the sum of itself and the heat pump sequence. The result is saved as a new electric-mains cache dhf file. After the data of all households are updated, the complete new simulation dataset is obtained.

4.4.3 Evaluation and Discussion

Evaluating the performance of the original model for the heat pump dataset is an important goal of this project. As operated in 4.3.3, the previously obtained model was used to test the model on the dataset with heat pumps that we just obtained. The results of the evaluation are shown in table 4.2 and Appendix B. This result is compared to the performance of the model on the data without heat pumps as shown in figure 4.1, 4.2, 4.3.

Appliance Type	Recall	Precison	F1 Score
Cooker	0.81	0.10	0.18
Shower	0.98	0.08	0.15
Kettle	0.78	0.53	0.63
Washingmachine	0.80	0.31	0.44

Table 4.2: The aggregated activation statistics of the heat pump-free model on the heat pump dataset(Realistic heat pump simulation)

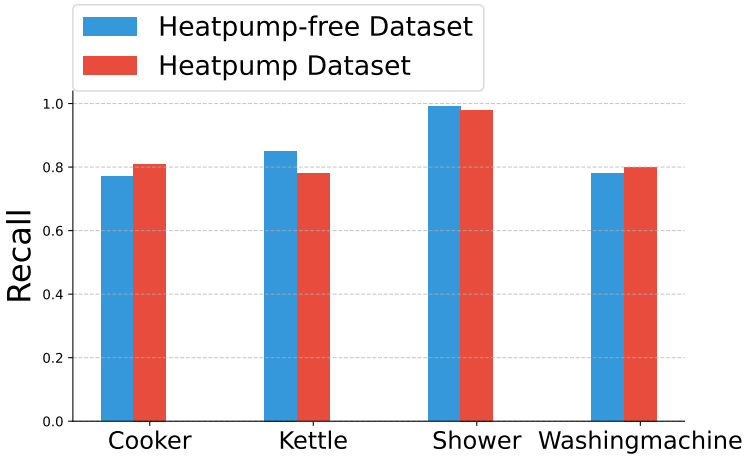

Figure 4.1: Recall: comparing the difference in heatpump-free model performance between heatpump-free dataset and heatpump dataset

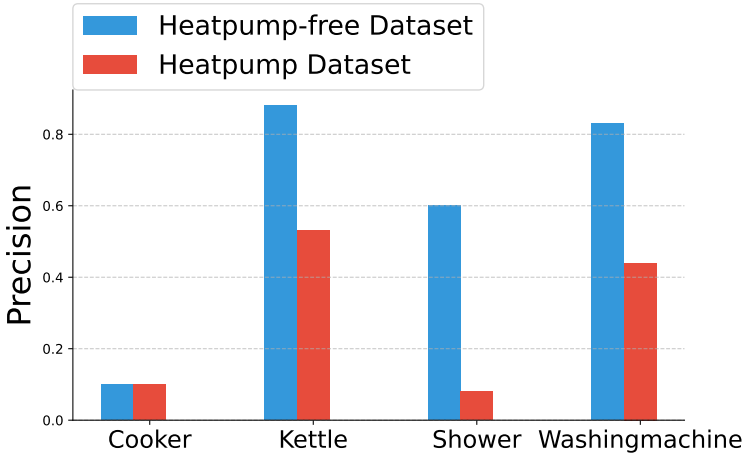

Figure 4.2: Precision: comparing the difference in heatpump-free model performance between heatpump-free dataset and heatpump dataset

The results of the comparisons show that for the four target appliances evaluated, the performance of the models is affected to varying degrees by the heat pump data. For the cooker, the difference in evaluation before and after the change in the dataset was the smallest. Even though its precision is already very low, in this project we focus only on the change in performance and do not discuss the prediction performance itself. For kettle, shower, and washingmachine, the predicted precisions all experienced significant decreases, suggesting that the addition of heat pumps is often incorrectly recognised by the model as the switching on of the target device. For kettle, the predicted recall also showed a small decrease, suggesting that when kettle is operating normally, the model

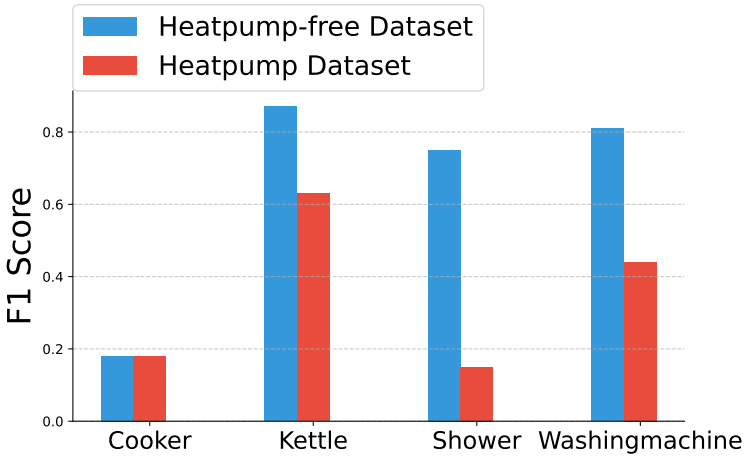

Figure 4.3: F1 Score: comparing the difference in heatpump-free model performance between heatpump-free dataset and heatpump dataset

is misled by the power fluctuations brought on by the heat pump and thus misjudge kettle to be off.

Meanwhile, there is a pitfall here. As we discussed in 3.3.1, when using real data simulation methods for simulation dataset creation, we may suffer from data sparsity. Thus, we conducted an investigation into the sparsity of the heat pump sequence. In the case of heat pump 5104, for example, the length of the sampled sequence is 3153589, of which there are 2297149 data points with the value of 0, accounting for 72.84%. This further supports the validity of our concern. Therefore, we suspect that perhaps this better evaluation result does not reflect the actual situation. Specifically, we are concerned that the smaller performance difference may come from the fact that the actual change in the dataset is smaller than we expected. When the simulated heat pump in the heat pump dataset is left inactive for long periods of time, the model is actually still demonstrating its performance for the no heat pump data.

4.4.4 Creating Low-sparsity Datasets

In order to solve the above mentioned problems and to give a more reliable evaluation, we chose the alternative method of building a virtual heat pump sequence that has been proposed in 3.3.3 - intensive working simulation - to be tested again. For the evaluation only, on the one hand, the use of this method does not cause problems of overfitting, and on the other hand, this allows to test to a greater extent the reflection of the model on the newly introduced heat pump data, as it increases as significantly as possible the

proportion of time that the heat pump is in operation in the sequence. Moreover, it is not unrealistic to create a virtual heat pump like this, as it still simulates the operating modes of the appliance as they might exist in reality.

After 4.4.1, for heat pump sequences that have been downsampled, we additionally made the following modifications in accordance with the intensive working simulation methodology:

1. Detecting and replacing consecutive zeros in the sequences in order to avoid excessively long silences of the heat pump. Specifically, all sequences of consecutive zeros with a length of more than 12 were retained up to 12, i.e., the maximum length of complete silence of the heat pump was set to 2 minutes.

2. After the sequence length was significantly shortened, the sequence was repeated until the length of the sequence reached the requirement.

We investigated the sparsity of the sequence again after modification. In the case of heat pump 5104, for example, there are 109,008 data points with the value of 0, accounting for 3.34%, which shows a significant decrease in the sparsity of the sequence compared to the previous one.

4.4.5 Re-evaluation and Discussion

Using the modified dataset, we ran the evaluation of the model again, and the results are shown in table 4.3 and Appendix B. The results are compared with the previous two evaluations as shown in figure 4.4, 4.5 and 4.6.

Appliance Type	Recall	Precison	F1 Score
Cooker	0.79	0.08	0.15
Shower	0.97	0.06	0.11
Kettle	0.65	0.54	0.59
Washingmachine	0.84	0.15	0.26

Table 4.3: The aggregated activation statistics of the heat pump-free model on the heat pump dataset(Intensive operation simulation)

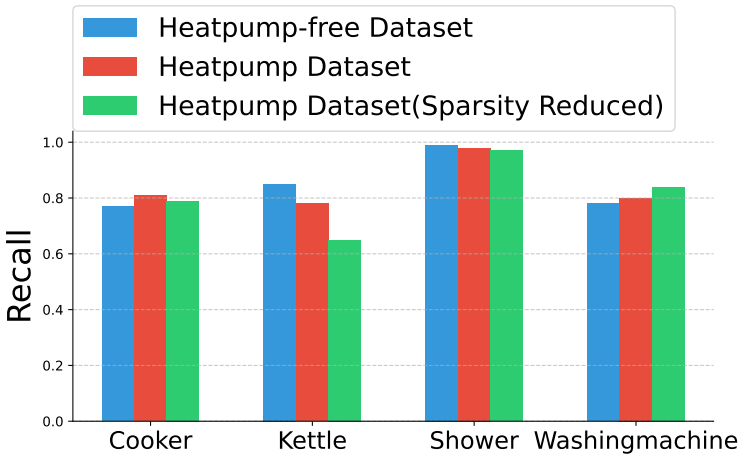

Figure 4.4: Recall: comparing the difference in heatpump-free model performance among heatpump-free dataset, heatpump dataset and heatpump dataset(sparsity reduced)

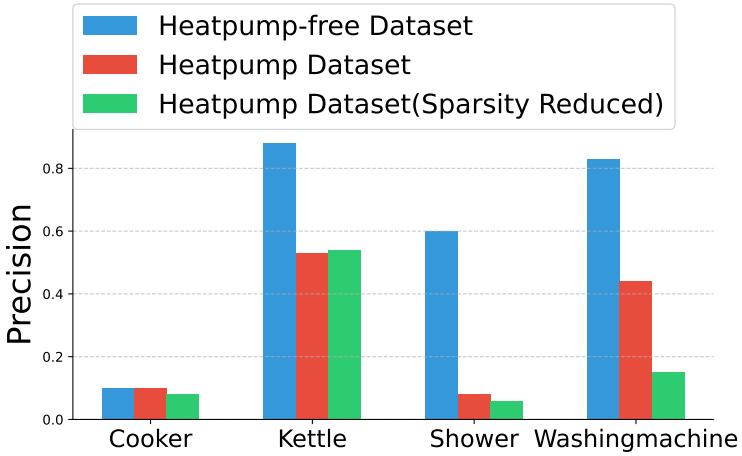

Figure 4.5: Precision: comparing the difference in heatpump-free model performance between heatpump-free dataset, heatpump dataset and heatpump dataset(sparsity reduced)

The results show a further decrease in the predictive performance of the model on the four target devices, almost maintaining the decreasing trend shown in the first comparison. For recall, a more significant decrease in the prediction performance of kettle can be observed, while no significant change is found for the other devices. For precision, washingmachine's performance shows a more drastic downward trend compared to the previous comparison. The combined F1 score shows that the prediction performance of all four devices is lower on the new heat pump dataset compared to

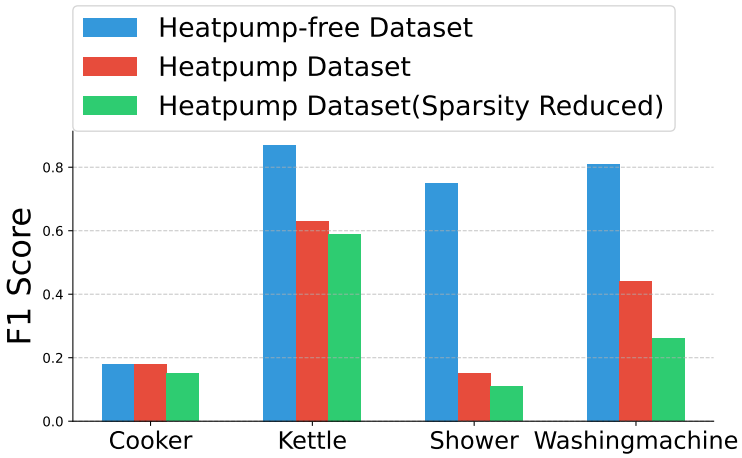

Figure 4.6: F1 Score: comparing the difference in heatpump-free model performance between heatpump-free dataset, heatpump dataset and heatpump dataset(sparsity reduced)

on the sparse heat pump dataset. Such a result supports the weakening of the model brought about by the heat pump data that was observed in the first comparison. At the same time, it suggests that using sparse data for the evaluation does to some extent make the model less affected than it actually appears to be, although it is also able to capture the trend of this effect. The model is more heavily influenced by the heat pump data than is reflected in the evaluation results.

4.5 Training and Evaluation of the New Model

The above analysis shows that it is necessary to retrain the model using an updated dataset to try to solve the problem of the original model not adapting to the heat pump data. Here, we use the heat pump dataset created from the real simulation data instead of the sparsity-reduced heat pump dataset, as the latter suffers from imbalance in the training data, as mentioned in 3.3.4. After training, we obtained a new model trained on the energy consumption records of homes with virtual heat pumps installed. We evaluate the performance of this model on the dataset with and without heat pumps and compare it with the evaluation of the model without heat pumps on the same dataset.

4.5.1 Evaluation of the New Model on Heatpump Data

Using the updated dataset, we run the evaluation of the new model and the results are shown in table A.1 and Appendix B. The results are compared with the evaluation of the heat pump-free model on this dataset as shown in figure 4.7, 4.8 and 4.9.

Appliance Type	Recall	Precison	F1 Score
Cooker	0.71	0.12	0.20
Shower	0.98	0.08	0.15
Kettle	0.74	0.66	0.70
Washingmachine	0.80	0.74	0.77

Table 4.4: The aggregated activation statistics of the new model on the heat pump dataset

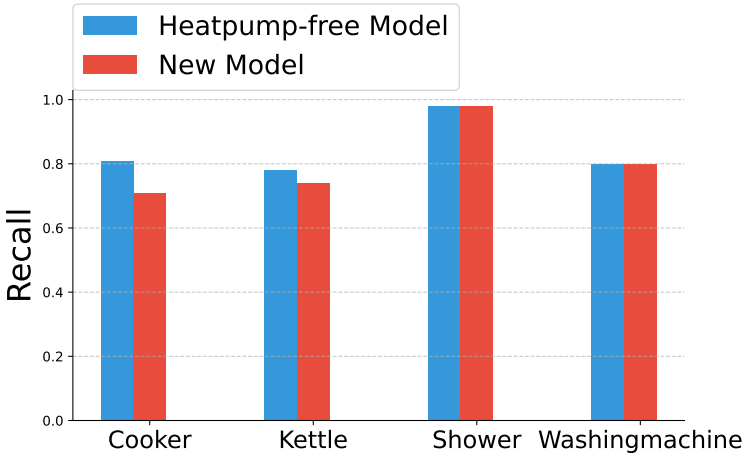

Figure 4.7: Recall: comparing the performance difference in new model and heatpump-free model on heatpump dataset

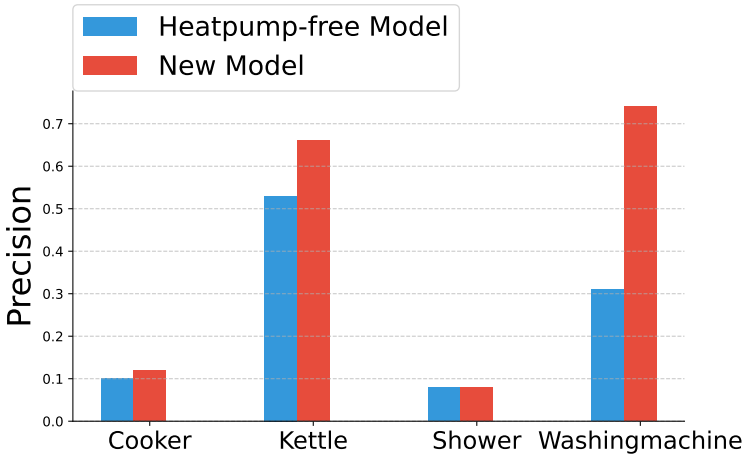

Figure 4.8: Precision: comparing the performance difference in new model and heatpump-free model on heatpump dataset

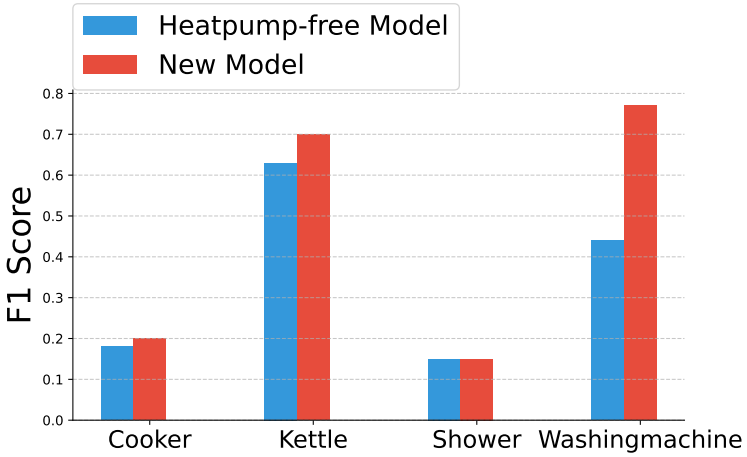

Figure 4.9: F1 Score: comparing the performance difference in new model and heatpump-free model on heatpump dataset

Comparison of the results for the two models shows that retraining improves the predictive performance of the models to some extent on the dataset with heat pumps. For cooker and kettle, the recall of the new model shows small fluctuations. However, for kettle and washingmachine, the training of the new model significantly improves the precision performance, which is the main disadvantage of the model without heat pump. In terms of combined F1 score performance, for cooker, kettle and washingmachine, the new model can be observed to improve prediction performance on the heat pump dataset compared to the previous model. However, retraining does not seem to work for shower, as it fails to produce any improvement in its low precisions.

4.5.2 Evaluation of the New Model on Heatpump-free Data

At the end of the experiment, we again ran the evaluation of the new model on the no-heat-pump dataset, and the results are shown in 4.5 and Appendix B. On this basis, we can observe whether training on the heat-pump dataset has an impact on the prediction performance of the other devices. Also, we can compare the performance of the new model on the two datasets. The comparison results are shown in figure 4.10, 4.11 and 4.12.

Appliance Type	Recall	Precison	F1 Score
Cooker	0.71	0.11	0.19
Shower	0.99	0.68	0.81
Kettle	0.81	0.91	0.85
Washingmachine	0.77	0.82	0.80

Table 4.5: The aggregated activation statistics of the new model on the heatpump-free dataset

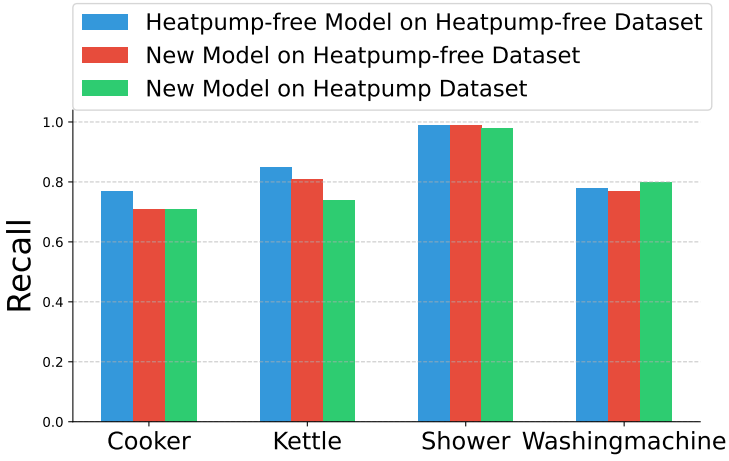

Figure 4.10: Recall: comparing the performance difference in new model and heatpump-free model on heatpump-free dataset

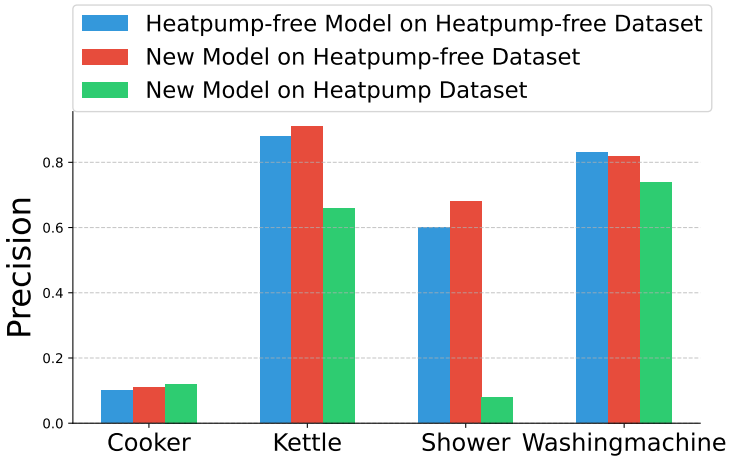

Figure 4.11: Precision: comparing the performance difference in new model and heatpump-free model on heatpump-free dataset

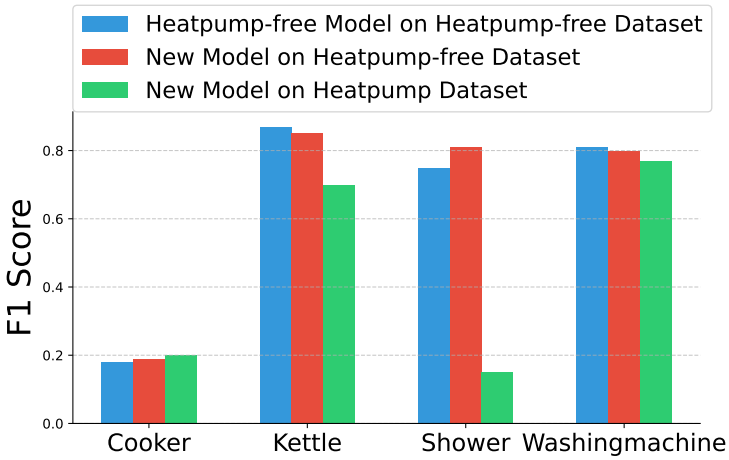

Figure 4.12: F1 Score: comparing the performance difference in new model and heatpump-free model on heatpump dataset

First, we compare the performance of the new model and the heat pump-free model on the heat pump-free dataset (Observe columns 1 and 2). The results show that for all four target devices, the predicted recall, precision, and F1 scores fluctuate only slightly before and after the model replacement, without significant differences. This suggests that the inclusion of the heat pump data did not make it more difficult for the new model to learn the true characteristics of the energy consumption patterns of the target devices during the training phase.

Then, we compare the performance of the new model on two datasets (observation column 2 and column 3). The results show that there is no significant difference in

recall for all four devices. However, for kettle and shower, the performance of the new model on the heat pump-free data is significantly higher than that on the heat pump data. This shows a similar trend to the difference in performance of the heat pump-free model on the two datasets. Combined with the comparison results in 4.5.1, we can say that model retraining has some mitigating effect on the model adaptation problem on the new data, but does not fundamentally address the low model decomposition performance for the target device in the case of the heat pump operation (compared to the case of the no-heat-pump operation).

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this work, we created a simulation dataset of households with low-carbon technology appliances - heat pumps - based on the IDEAL dataset and DECC RHPP data. We first trained the NILM power disaggregation model using the dataset of households without heat pumps and compared the predictive performance of the model on the dataset with heat pumps versus the dataset without heat pumps. To reliably verify that the model was not affected by the heat pump data that was introduced, we used two different virtual heat pump simulation datasets.We then retrained the model and evaluated it again to analyse the performance and impact of the new model.

The evaluation results show that for all three target appliances (except for the cooker, which already has low predictive performance), the addition of heat pump data significantly reduces the predictive performance of the original models, especially the precision, i.e., the original models are not able to adapt to the new dataset. For kettle and washingmachine, retraining significantly improves model performance on the new dataset, but this does not work for shower. Despite these improvements, the performance of the new model on the heat pump dataset is still significantly lower than on the heat pump-free dataset, suggesting that the misleading nature of the heat pump data for prediction has not been fundamentally addressed in this way. At the same time, it is encouraging that training on the new data does not affect the model's ability to disaggregate the appliances themselves.

5.2 Limitations and Future Work

The project has the following limitations, so we have more room for improvement and exploration in the future.

5.2.1 Reliability of Results

Although we have avoided the evaluation results being affected by the sparse heat pump data by reducing the sparsity, the reliability of the results of this evaluation remains to be tested. For the IDEAL dataset, the number of augmented-group households that can be used for the experiment only accounts for a relatively small portion of it. To make things worse, many more data were discarded during the experiment for reasons such as insufficient data in the corresponding window, which further reduced the amount of data actually used for the experiment. Therefore, in the future, we may be able to obtain more reliable validation results by collecting more dedicated data. In addition, more rigorous statistical testing of the data is also an important task.

5.2.2 Generalisation of Results

Our work is currently limited to discussing the impact of the introduction of heat pump data on the four target appliances we have chosen, and cannot be simply generalised to a wider range of low carbon technologies and target appliances. Until we gain more understanding of the models and the appliances themselves, we can only examine the effects of new appliances on a case-by-case basis in the future.

5.2.3 Interpretability

As mentioned above, we are currently unable to understand how this black-box system works internally, and therefore it is difficult for us to work on the principle level to improve the performance of the system.. If its predictive principles cannot be thoroughly investigated, then this work will never be able to be widely promoted, especially when it is applied to such a sensitive area as personal safety and security. Indeed, this is the challenge for the entire field of ”applying machine learning to NILM problems”. When breakthroughs are made in this area, we will also have a deeper understanding of the blind source decomposition problem itself.

Bibliography

[1] George William Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[2] Chaoyun Zhang, Mingjun Zhong, Zongzuo Wang, Nigel Goddard, and Charles Sutton. Sequence-to-point learning with neural networks for non-intrusive load monitoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[3] Cillian Brewitt and Nigel Goddard. Non-intrusive load monitoring with fully convolutional networks, 2018.

[4] George W Hart. Residential energy monitoring and computerized surveillance via utility power flows. *IEEE Technology and Society Magazine*, 8(2):12–16, 1989.

[5] Nilm workshop 2022. http://nilmworkshop.org/2022/. Accessed: April 20, 2023.

[6] Mingjun Zhong, Nigel Goddard, and Charles Sutton. Interleaved factorial non-homogeneous hidden markov models for energy disaggregation. *arXiv preprint arXiv:1406.7665*, 2014.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[8] Jack Kelly and William Knottenbelt. Neural nilm: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM international conference on embedded systems for energy-efficient built environments*, pages 55–64, 2015.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[10] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[11] Marcel Elswijk and Henk Kaan. European embedding of passive houses. *Final report of the IEE-SAVE project Promotion of European Passive Houses (PEP)*, 2008.

[12] David Banks. *An introduction to thermogeology: ground source heating and cooling*. John Wiley & Sons, 2012.

[13] Martin Zogg. History of heat pumps-swiss contributions and international milestones. 2008.

[14] Harry J Sauer and Ronald Hunter Howell. Heat pump systems. 1983.

[15] Shenghua Zou and Xiaokai Xie. Simplified model for coefficient of performance calculation of surface water source heat pump. *Applied Thermal Engineering*, 112:201–207, 2017.

[16] Martin Pullinger, Jonathan Kilgour, Nigel Goddard, Niklas Berliner, Lynda Webb, Myroslava Dzikovska, Heather Lovell, Janek Mann, Charles Sutton, Janette Webb, et al. The ideal household energy dataset, electricity, gas, contextual sensor data and survey data for 255 uk homes. *Scientific Data*, 8(1):146, 2021.

[17] R Lowe et al. Renewable heat premium payment scheme: heat pump monitoring: cleaned data, 2013-2015. 2022.

Appendix A

Correspondence of Data Combining

The correspondence of IDEAL households(34) and virtue heat pumps(34) in data combining is in table A.1

IDEAL Household	**Virtue Heat Pump**
home-105	processed_rhpp5104
home-106	processed_rhpp5105
home-128	processed_rhpp5106
home-136	processed_rhpp5107
home-139	processed_rhpp5108
home-140	processed_rhpp5111
home-145	processed_rhpp5112
home-146	processed_rhpp5113
home-162	processed_rhpp5114
home-168	processed_rhpp5117
home-169	processed_rhpp5118
home-171	processed_rhpp5119
home-175	processed_rhpp5121
home-208	processed_rhpp5123
home-212	processed_rhpp5124
home-225	processed_rhpp5126
home-227	processed_rhpp5128
home-228	processed_rhpp5129
home-231	processed_rhpp5132
home-238	processed_rhpp5136
home-242	processed_rhpp5141

home-249	processed_rhpp5144
home-255	processed_rhpp5145
home-264	processed_rhpp5149
home-266	processed_rhpp5150
home-268	processed_rhpp5152
home-328	processed_rhpp5154
home-61	processed_rhpp5155
home-62	processed_rhpp5156
home-63	processed_rhpp5157
home-65	processed_rhpp5159
home-73	processed_rhpp5160
home-96	processed_rhpp5161

Table A.1: The correspondence of IDEAL households and virtue heat pumps in data combining

Appendix B

Detailed Result of Evaluations

B.1 Evaluation 1 (4.3.3)

This section shows the complete results of predicting and evaluating the test set in the heatpump-free using the heatpump-free model. See 3.5 for the selection of target appliances and 3.7 for details of the evaluation metrics.

homeid	NDE	MAE	MR	SAE
242	0.93	55.09	0.06	0.90
264	0.21	38.04	0.43	0.68
Aggregated	0.57	48.06	0.22	0.79
Averaged	0.57	46.56	0.25	0.79

Table B.1: Cooker: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
242	99.0	1150.0	43.0	0.70	0.08	0.14
264	68.0	320.0	6.0	0.92	0.18	0.29
Aggregated	167.0	1470.0	49.0	0.77	0.10	0.18
Averaged	-	-	-	0.81	0.13	0.22

Table B.2: Cooker: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.09	1.26	0.68	0.04
227	0.27	15.59	0.47	0.18
249	0.30	24.64	0.47	0.34
264	0.20	8.95	0.52	0.20
Aggregated	0.22	5.56	0.53	0.17
Averaged	0.22	12.61	0.53	0.19

Table B.3: Kettle: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	384.0	10.0	39.0	0.91	0.97	0.94
227	322.0	46.0	113.0	0.74	0.88	0.80
249	256.0	84.0	24.0	0.91	0.75	0.83
264	348.0	33.0	59.0	0.86	0.91	0.88
Aggregated	1310.0	173.0	235.0	0.85	0.88	0.87
Averaged	-	-	-	0.85	0.88	0.86

Table B.4: Kettle: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.23	13.45	0.46	0.31
227	0.47	61.89	0.29	0.22
249	0.22	21.00	0.48	0.33
Aggregated	0.34	27.96	0.38	0.27
Averaged	0.31	32.11	0.41	0.29

Table B.5: Shower: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	112.0	6.0	3.0	0.97	0.95	0.96
227	167.0	166.0	1.0	0.99	0.50	0.67
249	68.0	59.0	1.0	0.99	0.54	0.69
Aggregated	347.0	231.0	5.0	0.99	0.60	0.75
Averaged	-	-	-	0.98	0.66	0.77

Table B.6: Shower: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.10	4.06	0.65	0.17
212	0.20	7.32	0.61	0.28
171	0.39	30.63	0.46	0.42
227	0.48	17.56	0.40	0.12
242	0.47	27.22	0.36	0.54
264	0.08	6.13	0.72	0.10
Aggregated	0.30	10.83	0.51	0.30
Averaged	0.29	15.49	0.53	0.27

Table B.7: Washingmachine: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	137.0	2.0	0.0	1.00	0.99	0.99
212	69.0	3.0	27.0	0.72	0.96	0.82
171	48.0	5.0	15.0	0.76	0.91	0.83
227	89.0	84.0	17.0	0.84	0.51	0.64
242	94.0	4.0	75.0	0.56	0.96	0.70
264	51.0	2.0	0.0	1.00	0.96	0.98
Aggregated	488.0	100.0	134.0	0.78	0.83	0.81
Averaged	-	-	-	0.81	0.88	0.83

Table B.8: Washingmachine: Activation Metrics

B.2 Evaluation 2 (4.4.3)

This section shows the complete results of predicting and evaluating the test set in the heatpump dataset using the heatpump-free model. See 3.5 for the selection of target appliances and 3.7 for details of the evaluation metrics.

homeid	NDE	MAE	MR	SAE
242	0.92	58.89	0.08	1.22
264	0.24	42.13	0.41	0.81
Aggregated	0.58	51.98	0.22	1.00
Averaged	0.58	50.51	0.24	1.01

Table B.9: Cooker: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
242	107.0	1154.0	35.0	0.75	0.08	0.15
264	69.0	415.0	5.0	0.93	0.14	0.25
Aggregated	176.0	1569.0	40.0	0.81	0.10	0.18
Averaged	-	-	-	0.84	0.11	0.20

Table B.10: Cooker: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.29	3.46	0.42	0.64
227	1.02	54.13	0.15	1.11
249	0.33	25.99	0.44	0.35
264	0.21	9.39	0.51	0.19
Aggregated	0.50	12.30	0.31	0.37
Averaged	0.46	23.24	0.38	0.57

Table B.11: Kettle: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	373.0	147.0	50.0	0.88	0.72	0.79
227	238.0	818.0	197.0	0.55	0.23	0.32
249	250.0	89.0	30.0	0.89	0.74	0.81
264	345.0	37.0	62.0	0.85	0.90	0.87
Aggregated	1206.0	1091.0	339.0	0.78	0.53	0.63
Averaged	-	-	-	0.79	0.65	0.70

Table B.12: Kettle: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.24	14.99	0.43	0.26
227	0.73	159.15	0.12	1.21
249	0.22	21.85	0.47	0.30
Aggregated	0.47	55.43	0.22	0.45
Averaged	0.39	65.33	0.34	0.59

Table B.13: Shower: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	112.0	39.0	3.0	0.97	0.74	0.84
227	164.0	3707.0	4.0	0.98	0.04	0.08
249	68.0	75.0	1.0	0.99	0.48	0.64
Aggregated	344.0	3821.0	8.0	0.98	0.08	0.15
Averaged	-	-	-	0.98	0.42	0.52

Table B.14: Shower: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.52	13.37	0.35	0.68
212	0.33	14.86	0.42	0.08
171	0.60	56.55	0.30	0.01
227	1.27	48.99	0.15	1.01
242	0.77	44.37	0.23	0.22
264	0.18	10.56	0.59	0.10
Aggregated	0.63	23.98	0.30	0.26
Averaged	0.61	31.45	0.34	0.35

Table B.15: Washingmachine: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	135.0	338.0	2.0	0.99	0.29	0.44
212	73.0	231.0	23.0	0.76	0.24	0.36
171	48.0	54.0	15.0	0.76	0.47	0.58
227	98.0	393.0	8.0	0.92	0.20	0.33
242	94.0	62.0	75.0	0.56	0.60	0.58
264	51.0	47.0	0.0	1.00	0.52	0.68
Aggregated	499.0	1125.0	123.0	0.80	0.31	0.44
Averaged	-	-	-	0.83	0.39	0.50

Table B.16: Washingmachine: Activation Metrics

B.3 Evaluation 3 (4.4.5)

This section shows the complete results of predicting and evaluating the test set in the heatpump dataset(sparsity reduced) using the heatpump-free model. See 3.5 for the selection of target appliances and 3.7 for details of the evaluation metrics.

homeid	NDE	MAE	MR	SAE
242	0.94	58.21	0.07	1.12
264	0.40	61.76	0.32	1.40
Aggregated	0.67	59.67	0.20	1.27
Averaged	0.67	59.99	0.20	1.26

Table B.17: Cooker: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
242	102.0	1158.0	40.0	0.72	0.08	0.15
264	68.0	792.0	6.0	0.92	0.08	0.15
Aggregated	170.0	1950.0	46.0	0.79	0.08	0.15
Averaged	-	-	-	0.82	0.08	0.15

Table B.18: Cooker: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.35	3.89	0.39	0.71
227	1.12	76.44	0.04	1.50
249	0.33	26.54	0.43	0.36
264	0.30	15.05	0.36	0.02
Aggregated	0.57	16.48	0.21	0.56
Averaged	0.53	30.48	0.30	0.65

Table B.19: Kettle: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	371.0	192.0	52.0	0.88	0.66	0.75
227	55.0	493.0	380.0	0.13	0.10	0.11
249	249.0	81.0	31.0	0.89	0.75	0.82
264	328.0	104.0	79.0	0.81	0.76	0.78
Aggregated	1003.0	870.0	542.0	0.65	0.54	0.59
Averaged	-	-	-	0.67	0.57	0.62

Table B.20: Kettle: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.23	14.94	0.44	0.25
227	0.65	154.28	0.11	1.04
249	0.22	22.09	0.47	0.30
Aggregated	0.43	54.12	0.22	0.37
Averaged	0.37	63.77	0.34	0.53

Table B.21: Shower: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	112.0	48.0	3.0	0.97	0.70	0.81
227	163.0	5417.0	5.0	0.97	0.03	0.06
249	68.0	68.0	1.0	0.99	0.50	0.66
Aggregated	343.0	5533.0	9.0	0.97	0.06	0.11
Averaged	-	-	-	0.98	0.41	0.51

Table B.22: Shower: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.58	14.91	0.32	0.80
212	0.67	31.27	0.21	0.77
171	1.60	175.76	0.08	1.95
227	3.91	146.07	0.03	5.07
242	0.77	43.59	0.23	0.25
264	0.77	46.64	0.21	1.64
Aggregated	1.29	48.40	0.15	1.34
Averaged	1.38	76.37	0.18	1.75

Table B.23: Washingmachine: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	136.0	400.0	1.0	0.99	0.25	0.40
212	82.0	653.0	14.0	0.85	0.11	0.20
171	59.0	231.0	4.0	0.94	0.20	0.33
227	99.0	1004.0	7.0	0.93	0.09	0.16
242	97.0	46.0	72.0	0.57	0.68	0.62
264	51.0	544.0	0.0	1.00	0.09	0.16
Aggregated	524.0	2878.0	98.0	0.84	0.15	0.26
Averaged	-	-	-	0.88	0.24	0.31

Table B.24: Washingmachine: Activation Metrics

B.4 Evaluation 4 (4.5.1)

This section shows the complete results of predicting and evaluating the test set in the heatpump-free dataset using the new model. See 3.5 for the selection of target appliances and 3.7 for details of the evaluation metrics.

homeid	NDE	MAE	MR	SAE
242	0.90	48.39	0.05	0.49
264	0.20	35.94	0.41	0.40
Aggregated	0.56	43.26	0.21	0.44
Averaged	0.55	42.16	0.23	0.45

Table B.25: Cooker: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
242	86.0	852.0	56.0	0.61	0.09	0.16
264	67.0	276.0	7.0	0.91	0.20	0.32
Aggregated	153.0	1128.0	63.0	0.71	0.12	0.20
Averaged	-	-	-	0.76	0.14	0.24

Table B.26: Cooker: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.20	1.72	0.57	0.07
227	0.70	32.41	0.20	0.08
249	0.37	26.64	0.40	0.48
264	0.26	8.84	0.49	0.36
Aggregated	0.41	8.19	0.37	0.18
Averaged	0.38	17.40	0.42	0.25

Table B.27: Kettle: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	373.0	37.0	50.0	0.88	0.91	0.90
227	216.0	454.0	219.0	0.50	0.32	0.39
249	237.0	70.0	43.0	0.85	0.77	0.81
264	323.0	26.0	84.0	0.79	0.93	0.85
Aggregated	1149.0	587.0	396.0	0.74	0.66	0.70
Averaged	-	-	-	0.75	0.73	0.74

Table B.28: Kettle: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.29	18.65	0.36	0.20
227	0.73	170.97	0.12	1.41
249	0.27	26.43	0.40	0.29
Aggregated	0.49	61.47	0.20	0.57
Averaged	0.43	72.02	0.29	0.63

Table B.29: Shower: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	112.0	18.0	3.0	0.97	0.86	0.91
227	165.0	3822.0	3.0	0.98	0.04	0.08
249	68.0	25.0	1.0	0.99	0.73	0.84
Aggregated	345.0	3865.0	7.0	0.98	0.08	0.15
Averaged	-	-	-	0.98	0.54	0.61

Table B.30: Shower: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.15	6.24	0.56	0.11
212	0.71	17.65	0.16	0.60
171	0.48	40.62	0.36	0.34
227	0.74	28.38	0.23	0.05
242	0.50	31.11	0.31	0.50
264	0.09	9.97	0.60	0.04
Aggregated	0.46	16.34	0.36	0.23
Averaged	0.45	22.33	0.37	0.27

Table B.31: Washingmachine: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	137.0	3.0	0.0	1.00	0.98	0.99
212	47.0	4.0	49.0	0.49	0.92	0.64
171	56.0	11.0	7.0	0.89	0.84	0.86
227	95.0	138.0	11.0	0.90	0.41	0.56
242	109.0	16.0	60.0	0.64	0.87	0.74
264	51.0	5.0	0.0	1.00	0.91	0.95
Aggregated	495.0	177.0	127.0	0.80	0.74	0.77
Averaged	-	-	-	0.82	0.82	0.79

Table B.32: Washingmachine: Activation Metrics

B.5 Evaluation 5 (4.5.2)

This section shows the complete results of predicting and evaluating the test set in the heatpump-free dataset using the new model. See 3.5 for the selection of target appliances and 3.7 for details of the evaluation metrics.

homeid	NDE	MAE	MR	SAE
242	0.98	49.69	0.04	0.54
264	0.20	37.03	0.40	0.43
Aggregated	0.60	44.48	0.21	0.48
Averaged	0.59	43.36	0.22	0.49

Table B.33: Cooker: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
242	87.0	947.0	55.0	0.61	0.08	0.15
264	67.0	267.0	7.0	0.91	0.20	0.33
Aggregated	154.0	1214.0	62.0	0.71	0.11	0.19
Averaged	-	-	-	0.76	0.14	0.24

Table B.34: Cooker: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.14	1.11	0.68	0.20
227	0.34	14.63	0.44	0.40
249	0.35	25.73	0.42	0.47
264	0.26	8.66	0.50	0.37
Aggregated	0.28	5.34	0.50	0.36
Averaged	0.27	12.54	0.51	0.36

Table B.35: Kettle: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	383.0	1.0	40.0	0.91	1.00	0.95
227	293.0	31.0	142.0	0.67	0.90	0.77
249	245.0	72.0	35.0	0.88	0.77	0.82
264	327.0	25.0	80.0	0.80	0.93	0.86
Aggregated	1248.0	129.0	297.0	0.81	0.91	0.85
Averaged	-	-	-	0.81	0.90	0.85

Table B.36: Kettle: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.29	19.20	0.34	0.19
227	0.49	62.80	0.27	0.28
249	0.29	26.64	0.39	0.31
Aggregated	0.39	32.38	0.32	0.25
Averaged	0.36	36.21	0.34	0.26

Table B.37: Shower: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	112.0	3.0	3.0	0.97	0.97	0.97
227	167.0	140.0	1.0	0.99	0.54	0.70
249	68.0	17.0	1.0	0.99	0.80	0.88
Aggregated	347.0	160.0	5.0	0.99	0.68	0.81
Averaged	-	-	-	0.98	0.77	0.85

Table B.38: Shower: Activation Metrics

homeid	NDE	MAE	MR	SAE
73	0.10	5.45	0.59	0.06
212	0.71	17.30	0.16	0.62
171	0.39	34.11	0.43	0.35
227	0.43	19.42	0.39	0.00
242	0.45	28.57	0.36	0.47
264	0.08	9.32	0.62	0.02
Aggregated	0.38	14.10	0.42	0.25
Averaged	0.36	19.03	0.43	0.25

Table B.39: Washingmachine: Energy Metrics

homeid	TP	FP	FN	Recall	Precision	F1
73	137.0	2.0	0.0	1.00	0.99	0.99
212	40.0	1.0	56.0	0.42	0.98	0.58
171	57.0	7.0	6.0	0.90	0.89	0.90
227	96.0	86.0	10.0	0.91	0.53	0.67
242	101.0	8.0	68.0	0.60	0.93	0.73
264	51.0	3.0	0.0	1.00	0.94	0.97
Aggregated	482.0	107.0	140.0	0.77	0.82	0.80
Averaged	-	-	-	0.80	0.88	0.81

Table B.40: Washingmachine: Activation Metrics