Proceedings of Machine Learning Research vol 242:222–234, 2024

Interpretable Data-Driven Model Predictive Control of Building Energy Systems using SHAP

Patrick Henkel PATRICK.HENKEL@EONERC.RWTH-AACHEN.DE
Tobias Kasperski TOBIAS.KASPERSKI@RWTH-AACHEN.DE
Phillip Stoffel PHILLIP.STOFFEL@EONERC.RWTH-AACHEN.DE
Dirk Müller DMUELLER@EONERC.RWTH-AACHEN.DE
RWTH Aachen University, Aachen, Germany.

Abstract

Advanced building energy system controls, such as model predictive control, rely on accurate system models. To reduce the modelling effort in the building sector, data-driven models are becoming increasingly popular in research. Despite their promising performance, data-driven models are considered black boxes. This black box nature is an obstacle to widespread application, as it is difficult for building operators to understand how predictions are made. Concepts known as Explainable Artificial Intelligence are being developed to improve the interpretability of black box models. This work combines the popular Explainable Artificial Intelligence method Shapley Additive Explanations (SHAP) with data-driven model predictive control to increase the interpretability of artificial neural networks used as process models during model creation. Using a standardised residual building energy system for controller testing, an in-depth analysis of how the models make predictions is carried out. In addition, the influence of different model setups on the control performance is evaluated. The results show that the different control performances can be justified by analysing the underlying models with SHAP. SHAP shows how the characteristics of a feature affect the prediction and reveals weaknesses in the model. In addition, the features can be sorted according to their influence on the prediction, which is utilized for feature selection.

Keywords: interpretable machine learning, explainable AI, XAI, MPC, DDMPC, ANN.

1. Introduction

The building sector is a significant contributor to climate change, accounting for about 30 % of global final energy demand (United Nations Environment Programme, 2022). Advanced control strategies, such as model predictive control (MPC), offer a promising approach to reduce CO_2 emissions during building operation (Drgoňa et al., 2020).

MPC strategies use a mathematical model of the controlled system to determine an optimal sequence of control variables. MPC can consider additional information such as weather forecasts and dynamic electricity profiles, and exploits the inertia and storage masses of the building. However, a significant barrier to widespread MPC implementation is the effort required to create a sufficiently accurate system model (Sturzenegger et al., 2016). This is a particular problem for the building sector due to the heterogeneous nature of the building stock.

Due to the modelling effort involved, data-driven model predictive control (DDMPC) of building energy systems is increasingly becoming the focus of research (Kathirgamanathan et al., 2021). In data-driven modelling, system behaviour is not described by physical equations, but is learned directly from measured training data. Data-driven process models can even outperform physics-based models (Krzysztof Arendt et al., 2018).

© 2024 P. Henkel, T. Kasperski, P. Stoffel & D. Müller.

A wide range of data-driven models can be used for DDMPC. Among other things, they differ in their accuracy, interpretability, and implementation effort. Bünning et al. compare Autoregressive-Moving-Average with Exogenous Inputs (ARMAX) models identified through linear regression with random forests and input convex neural networks (Bünning et al., 2022). The authors demonstrate that the resulting MPCs achieve savings of between 26 % and 49 % of heating and cooling energy. Other researchers also use non-linear approaches such as Gaussian process regression (Jain et al., 2018) and artificial neural networks (ANNs) (Stoffel et al., 2023b).

Despite their widespread use in research, data-driven models are considered black boxes. It is difficult for model developers and building operators to understand how such models make predictions. The black box nature hinders large-scale implementation in practice (Machlev et al., 2022). The engineering community has traditionally favoured transparent methods (Naser, 2021). Due to this challenge, concepts, known as Explainable Artificial Intelligence (XAI), are developed to increase the interpretability of data-driven models (Molnar, 2019).

In this work, DDMPC is combined with the popular XAI method Shapley Additive Explanations (SHAP) to increase the interpretability of ANNs during model creation. The contributions of this work are as follows:

- Combination of DDMPC and the XAI method SHAP to increase the interpretability of ANNs used as process models for MPC during the model creation process.
- In-depth analysis of how ANNs make predictions, modelling the standardised BESTEST Hydronic Heat Pump test case of the *Building Optimization Testing Framework* (BOPTEST).
- Evaluation of the influence of different ANN setups on the control performance of the DDMPC.

Section 2 summarises the state of the art of XAI methods. In section 3 the SHAP methodology and in section 4 the considered use case are introduced. The results and a conclusion are presented in sections 5 and 6.

2. State of the Art

In recent years, XAI methods have received increasing attention in the energy sector (Machlev et al., 2022). XAI methods are designed to increase the interpretability of black box models. They can be categorised according to their model dependency, application stage, and interpretability score, as shown in figure 1. The two categories of model dependency are model-specific and model-agnostic. Model-specific methods are specialised for certain model types, such as ANNs, and cannot be easily applied to other model types. Model-agnostic methods have the advantage that they can be applied to different model types, but may not provide explanations as well as specialised methods. In the review by Chen et al., which focuses on interpretable machine learning for building energy management, about 56 % of the papers reviewed use model-specific methods (Chen et al., 2023).

The two application stages are ante-hoc and post-hoc. Ante-hoc methods are applied during model creation, whereas post-hoc methods are applied to finished models. In the articles reviewed by Chen et al., about 43 % of the methods used are ante-hoc (Chen et al., 2023). In the literature, most ante-hoc methods are model-specific, whereas most post-hoc methods are model-agnostic. Ante-hoc methods usually improve interpretability by modifying some characteristics of the model, whereas post-hoc methods usually do not rely on model characteristics.

The two interpretability scopes are global and local. Local methods explain individual predictions, whereas global methods explain the general characteristics of a model. Global methods often

evaluate the importance of input features and can assist model developers with feature selection. Local methods focus on individual input samples and their contribution to a prediction and can therefore help building operators to understand the output of a model. In (Chen et al., 2023), about 60 % of the methods reviewed are global.

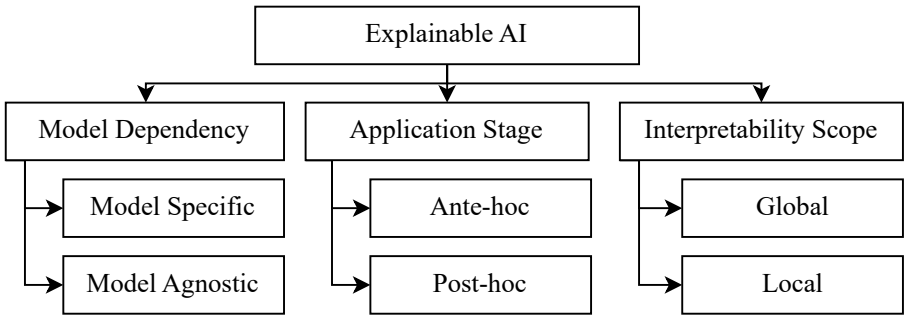

Figure 1: Categorisation of XAI methods.

In the literature, only about 9 % of the publications that apply XAI methods for building energy management focus on control. The majority of publications (about 62 %) focus on load or power prediction. (Chen et al., 2023) Therefore, one contribution of this paper is the combination of DDMPC and XAI methods to increase the interpretability of ANNs used as process models for MPC. In the literature, ANNs, which belong to the class of deep learning models, are considered to have a high model accuracy but low model interpretability (Barredo Arrieta et al., 2020). Therefore, this model type is an ideal application example for the combination of DDMPC and XAI. The following sections summarise typical methods that explain ANNs and are used in the control of building energy systems.

Typical ante-hoc methods are modified neural networks (Drgoňa et al., 2021; Di Natale et al., 2022) and the attention mechanism (Gangopadhyay et al., 2020). Although these methods achieve promising results, this paper focuses on post-hoc methods because ante-hoc methods are often model-specific and limit the choice of possible process model types. Details for the focus on post-hoc methods are presented in section 3.

In general, the two most popular post-hoc XAI methods are SHAP (Lundberg and Lee, 2017) and Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016). SHAP determines the individual contribution of each feature to a given model prediction. An example of SHAP applied to building energy management is given by Białek et al. (Białek et al., 2022). The authors train an ANN to predict the heat demand of a district heating network and interpret the influence of different features on the model output. LIME is a local method that generates an interpretable local substitute model for a given sample. LIME is used in building energy management for example to interpret electricity demand predictions and to support the model selection process (Grzeszczyk and Grzeszczyk, 2022).

Only three papers were found that use post-hoc XAI methods for the control of building energy systems. Mao et al. (Mao et al., 2023) explore the use of interpretable machine learning techniques such as LIME and SHAP for the purpose of Heating, Ventilation and Air Conditioning (HVAC) predictive control. However, the focus is on data analysis and modelling and the controller design is not discussed. Kotevska et al. (Kotevska et al., 2020) use LIME, partial dependence plots and individual conditional expectations for interpretable reinforcement learning (RL) of a heating, ven-

tilation and air conditioning control use case. The main difference with this paper is the focus on RL instead of DDMPC. Yu and Pavlak (Yu and Pavlak, 2022) generate interpretable building control rules from MPC data sets using rule extraction. The authors represent the simplified rules in the form of interpretable decision trees. The main difference with this paper is that the interpretable models are extracted from MPC data sets, whereas in this work the process models of a DDMPC are interpreted. In summary, to the best of the author's knowledge, there is no publication that combines DDMPC and XAI for building energy management.

3. Methodology

In this work, the interpretability of ANNs used as process models for MPC is increased by using the XAI method SHAP. First, the selection process of the XAI method used is explained, and then SHAP is presented in detail.

There exists a wide range of possible data-driven models that can be used for DDMPC. In order not to limit the choice of models by the choice of XAI method, it is advantageous if the XAI method used is model-agnostic. Most model-agnostic methods are also applied post-hoc. Regarding the interpretability scope, it is favourable, if the XAI method used can provide global and local explanations. Global explanations are helpful during model creation and feature selection. Local explanations are useful for analysing how the different features influence the output of specific predictions during DDMPC runtime.

The two most commonly used model-agnostic XAI methods are SHAP and LIME. Both generate local explanations. With SHAP, the individual local explanations can be aggregated globally. In (Chika E. Ugwuanyi, 2021), all the tests surveyed stated that SHAP generates more readable explanations than LIME. Therefore, SHAP is used in this work to increase the interpretability of ANNs used as process models for MPC.

SHAP belongs to the class of additive feature attribution methods (Lundberg and Lee, 2017). These methods approximate the prediction $f(x)$, of the original model f, based on the input x with m features, with a local substitute model g, where g is a linear function of binary variables:

$$g(z^{'}) = \phi_0 + \sum_{i=1}^{m} \phi_i z_i^{'}, \;\; z^{'} \in \{0,1\}^m \tag{1}$$

Substitute models often use simplified inputs $x^{'}$ that map to the original inputs through a mapping function $x = h_x(x^{'})$. Hence, $g(z^{'})$ should be an approximation of $f(x)$ $[g(z^{'}) \approx f(h_x(z^{'}))]$, whenever $z^{'} \approx x^{'}$. ϕ_i is the Shapley value of feature i. The Shapley values originate in cooperative game theory and determine the individual contribution of a player to the coalition outcome. The contribution of feature i is the marginal contribution of this feature to each coalition in which it is not included:

$$\phi_i(f,x) = \sum_{z^{'} \subseteq x^{'}} \frac{(m - |z^{'}| - 1)!|z^{'}|!}{m!} (f(h_x(z^{'})) - f(h_x(z^{'} \setminus \{i\})) \tag{2}$$

The marginal contribution is the difference between $f(h_x(z^{'}))$ and $f(h_x(z^{'} \setminus \{i\})$. The contribution of a feature i is the weighted average of these marginal contributions. The term $\frac{(m-|z^{'}|-1)!|z^{'}|!}{m!}$ is

used to weight the marginal contributions.

In a data-driven model, each coalition is a subset of the input features. Most models cannot handle missing features. Therefore, the absent features are replaced by the feature input values of background data, for example, the training data. ϕ_0 is the expected value of $f(x)$ over this background data. The Shapley values ϕ_i increase the interpretability of black box models by showing the individual contribution of features to the model output. As absent features are replaced by the feature input values of the background data, a major drawback of SHAP is the assumption that features are independent. To compare different models, the same background data must be used, as this influences the SHAP values.

In this work, the regression-based KernelSHAP (Lundberg and Lee, 2017) is used to calculate the SHAP values. A preliminary comparison with DeepSHAP, which was developed for deep learning models, shows that DeepSHAP can significantly reduce the calculation time. However, the SHAP values are sometimes very different. Therefore, the more accurate KernelSHAP is used.

4. Use Case

The DDMPC use case studied is based on BOPTEST (Blum et al., 2021). BOPTEST provides use cases for benchmarking building control strategies. In this work, the BESTEST Hydronic Heat Pump case is used. The test case is based on the BESTEST case 900 building with $192\,\mathrm{m}^2$ (R. Judkoff and J. Neymark, 1995) extended by an underfloor heating system and an air-to-water modulating heat pump. The residential building is located in Brussels and is inhabited by a family of five. BOPTEST defines reference periods for controller testing. The peak heating period is from January 17th to 31st and the typical heating period is from April 19th to May 3rd, respectively.

The control task of this test case is to keep the zone temperature T_{zone} within comfort constraints while minimising the electric costs C_{el}. The heat pump's modulation $u_{\mathrm{hp}} \in [0, 1]$ is used as the control signal. The DDMPC used in this work is mainly based on our previous work and a detailed description can be found in (Stoffel et al., 2023a). Two ANNs are used to model the quantities of interest. One ANN predicts the change in zone temperature $\Delta T_{\mathrm{zone,k}}$. Therefore, the zone temperature at the next time step $T_{\mathrm{zone,k+1}}$ can be expressed as:

$$T_{\mathrm{zone,k+1}} = T_{\mathrm{zone,k}} + \Delta T_{\mathrm{zone,k}} \tag{3}$$

The second ANN is used to model the electricity consumption $P_{\mathrm{el,k}}$. The predictors use not only the current (k) but also the lagged (past) values (k - M) of the input features to account for the thermal inertia of the system. The input features are listed in table 1. The features are selected manually and are based on system analysis and initial closed-loop experiments (Stoffel et al., 2023a). We construct the logistic modulation $u_{\mathrm{hp,log}}$ as an additional feature to support the learning of the heat pump's minimal power consumption. The logistic function continuously approximates a step that outputs 0 if $u_{\mathrm{hp}} = 0$ and 1 if $u_{\mathrm{hp}} > 0$. The time of day t_{day} and time of the week t_{week} are encoded as *sin* and *cos*.

Two different data sets are considered as training data. Both data sets are generated by simulating the BOPTEST framework during the first two weeks of January. The data set 'Base' is generated using the reference controller of the framework. When using the reference controller, the zone temperature varies little, making model identification difficult. To provide a data set with more

Table 1: Features and lags considered to predict the quantities of interest in the DDMPC. A lag of one means that the time steps k and $k-1$ are considered. (Stoffel et al., 2023a)

	ΔT_{zone}		$P_{\mathrm{el,hp}}$	
Feature	Considered	Lag	Considered	Lag
Zone temperature T_{zone}	x	2	x	0
Ambient temperature T_{amb}	x	1	x	0
Heat Pump modulation u_{hp}	x	2	x	0
Heat Pump modulation (logistic) $u_{\mathrm{hp,log}}$	-	-	x	0
Specific direct solar radiation $\dot{q}_{\mathrm{sol,dir}}$	x	0	-	-
Time of the Day (*sin* and *cos*)	x	0	-	-
Day of the Week (*sin* and *cos*)	x	0	-	-

operating points, the 'Explo' data set is introduced. This data set uses random zone temperature set points within the comfort bounds to excite the system. We show in our previous work (Stoffel et al., 2023a), that the base ANNs fail to generalise beyond the training data, while the explorative ANNs achieve better control performance. In addition, to show how SHAP can help with feature selection, a third ANN 'Feature selection' is introduced. It is based on the data set 'Explo', but unimportant features with low SHAP values are removed to investigate if the model can generalise better with a reduced feature space. In the following, the training data of the data set 'Explo' will be used as background data (see section 3) for the SHAP methodology.

5. Results

First, the influence of the different data sets on the control performance of the DDMPC is evaluated. Afterwards, the ANNs used are interpreted in-depth using SHAP.

Figure 2 shows the discomfort and operational cost of the BOPTEST scenarios for the investigated setups. In the peak heating period, both the 'Base' and 'Explo' DDMPCs can significantly improve the operational cost with only a small increase in discomfort compared to the reference controller. It is important to note that the increase in discomfort is negligible compared to the savings. The DDMPC 'Explo' outperforms the DDMPC 'Base' due to its more informative data set. In the typical heating period, the DDMPC 'Base' is not able to improve the control performance compared to the reference controller, while the DDMPC 'Explo' can improve the operational cost. However, compared to the peak heating period, the increase in discomfort is higher, but still negligible compared to the savings. As shown in our previous work (Stoffel et al., 2023a), the poorer performance of the DDMPC in the typical heating period can be explained by the fact that in this period the ANNs have to extrapolate more often beyond the known training data.

In the following, the ANNs are analysed in-depth using SHAP in order to find differences that explain the different control performances. First, the ANNs using the data set 'Explo' are analysed. Then the differences with the ANNs using the data set 'Base' are examined. Figure 3 shows the distribution of all calculated SHAP values for the ANN -$P_{\mathrm{el,hp,Explo}}$. The colour of a point indicates whether the feature input value for the calculated SHAP value is high or low. For a better understanding, the point which is furthest to the left and belongs to the feature u_{hp} is explained:

- The feature input value of u_{hp} is low, indicated by the blue colouring.
- The SHAP value is approximately -1600, which means that in this prediction, feature u_{hp} has reduced the prediction of the electrical power from the base value by 1600 W.

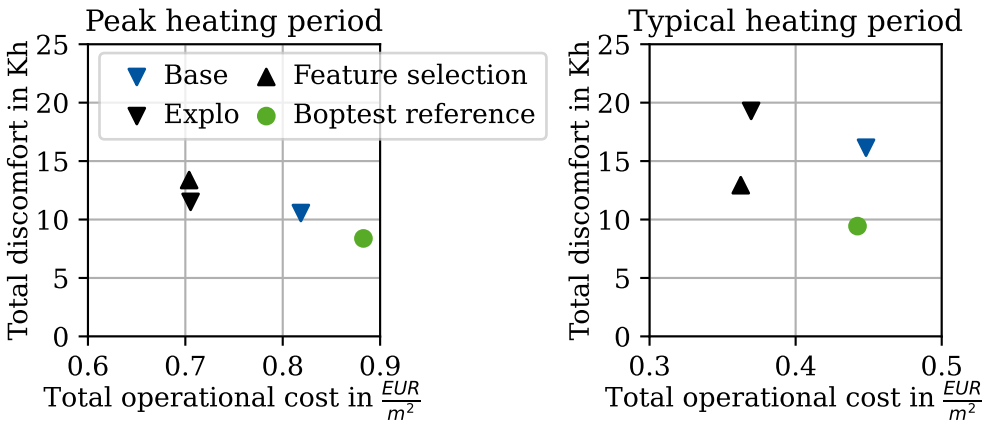

Figure 2: Control performance results of the BOPTEST scenarios for the examined setups.

The figure shows that the relative modulation of the heat pump u_{hp} is the feature with the strongest influence on the model prediction. Furthermore, high feature input values of u_{hp} increase the prediction of the electrical output, which is a physically meaningful correlation. All other features have only a small influence on the prediction.

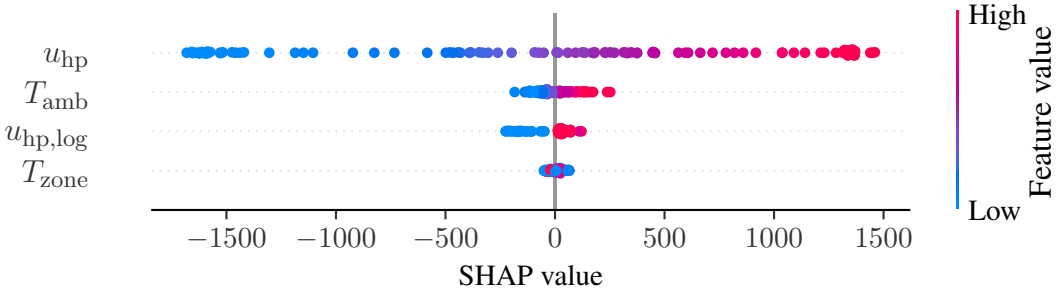

Figure 3: SHAP values for ANN-$P_{\mathrm{el,hp,Explo}}$.

Figure 4 shows the summary of the SHAP values for the ANN-$\Delta T_{\mathrm{zone,Explo}}$. It can be seen that the features T_{zone} with no lag and with a lag of 1 have the greatest influence on the model prediction. A high value of T_{zone} with no lag increases the model's prediction and a low value decreases it. This correlation is reversed for T_{zone} with a lag of 1. This phenomenon is also visible in the features T_{amb}. As a result, the model learns to predict the change in zone temperature not only from the feature input values but also by calculating the derivatives of some of the features.

In contrast, u_{hp} has only a small influence on the prediction of ΔT_{zone}. This is problematic, as this feature is the control variable and therefore should have a noticeable influence. However, the direction of the influence of u_{hp} is correctly learned by the model. In addition, it becomes apparent that u_{hp} with a lag of 1 has a greater influence than the current value of u_{hp}. Given the slow system dynamics of building energy systems, this can be explained physically.

Furthermore, the figure shows that in the learned model the four time features and the solar radiation have only a small influence on the prediction. To validate the assumption that the time features are not important, a new ANN called 'Feature selection' is trained. It also uses the data set 'Explo' with all the features shown in table 1 without the time features. The control performance is shown in

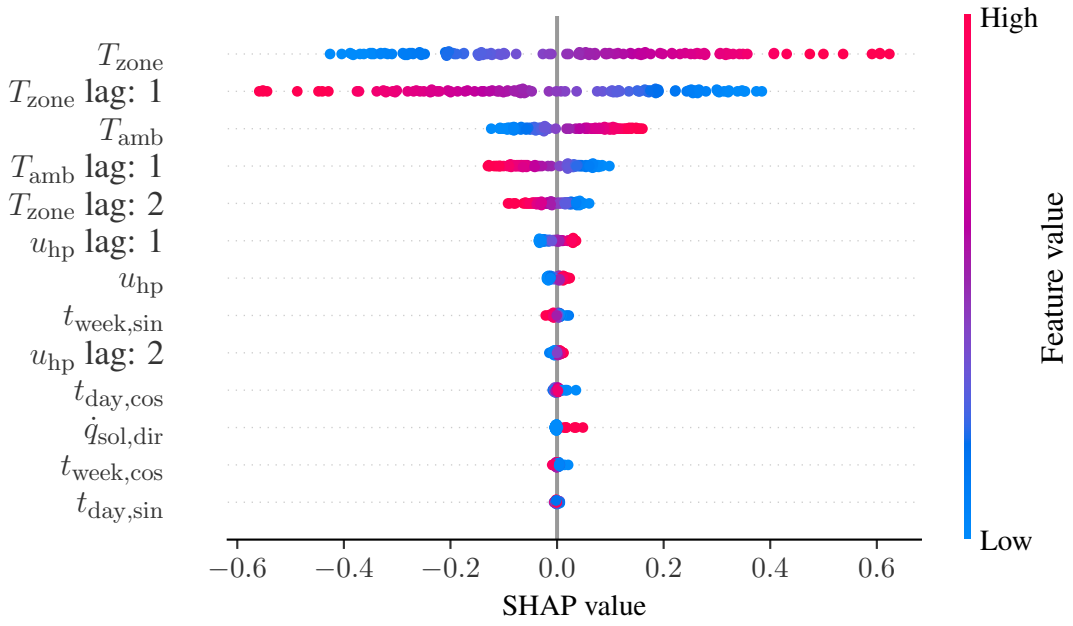

Figure 4: SHAP values for ANN-$\Delta T_{\text{zone,Explo}}$.

figure 2. It can be seen that removing the time features results in a comparatively small difference in the peak heating period. However, the control performance could be improved in the typical heating period. This illustrates that it can be advantageous to eliminate unimportant features, which allows the model to generalise better due to the reduced feature space. In addition this demonstrates the strength of the SHAP methodology in the feature selection and model creation process.

Figure 5 shows an overview of the SHAP values for the ANN-$P_{\text{el,hp,Base}}$. In comparison to the ANN-$P_{\text{el,hp,Explo}}$, noticeable differences are recognizable. The most important feature is still u_{hp}. However, the SHAP values for this feature are lower, while the other features have higher influences. The ANN-$P_{\text{el,hp,Explo}}$ already recognises whether the heat pump is switched on or off by the feature input value of u_{hp}. The ANN-$P_{\text{el,hp,Base}}$, on the other hand, also requires the feature $u_{\text{hp,log}}$, which reduces the prediction of the electrical output when the heat pump is switched off.

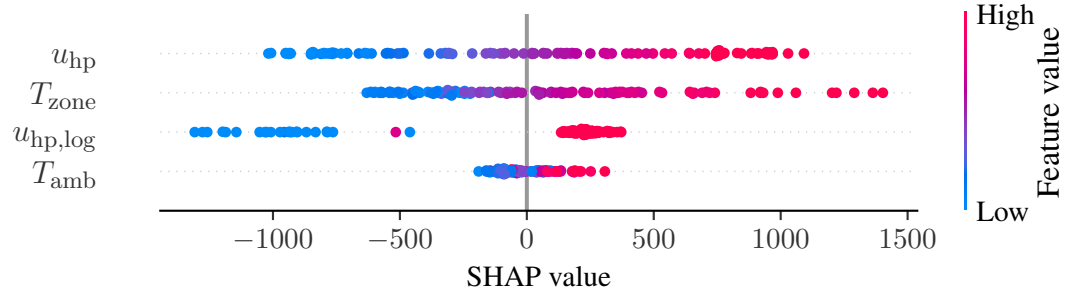

Figure 5: SHAP values for ANN-$P_{\text{el,hp,Base}}$.

Another difference between the two ANNs can be seen in T_{zone}. In the ANN-$P_{\text{el,hp,Base}}$, T_{zone} has a large influence on the prediction, whereas in the model ANN-$P_{\text{el,hp,Explo}}$ it has almost no rec-

ognizable influence. A possible reason for this is that the setpoint temperature for the base controller is reduced when the occupants of the house are not present. As a result, little electrical power is required for heating at low zone temperatures. This means that the ANN-$P_{\mathrm{el,hp,Base}}$ does not learn a physical relationship, but rather imitates the behaviour of the base controller.

The distribution of the SHAP values for the ANN-$\Delta T_{\mathrm{zone,Base}}$ is shown in figure 6. It can be seen that the time of day in cosine format has a significantly greater influence than in the ANN-$\Delta T_{\mathrm{zone,Explo}}$ shown in figure 4. The SHAP values for this feature are negative between 7 am and 5 pm and positive between 5 pm and 7 am. This roughly corresponds to the periods in which the occupants are present or absent during the week. The presence of occupants in the building results in internal heat gains. These have a positive influence on ΔT_{zone}. The ANN-$\Delta T_{\mathrm{zone,Base}}$ has presumably learned the influence of occupant behaviour on the change in zone temperature. However, this influence seems to be overestimated, as the feature has the fifth largest influence on the model prediction and therefore a greater influence than the system control variable.

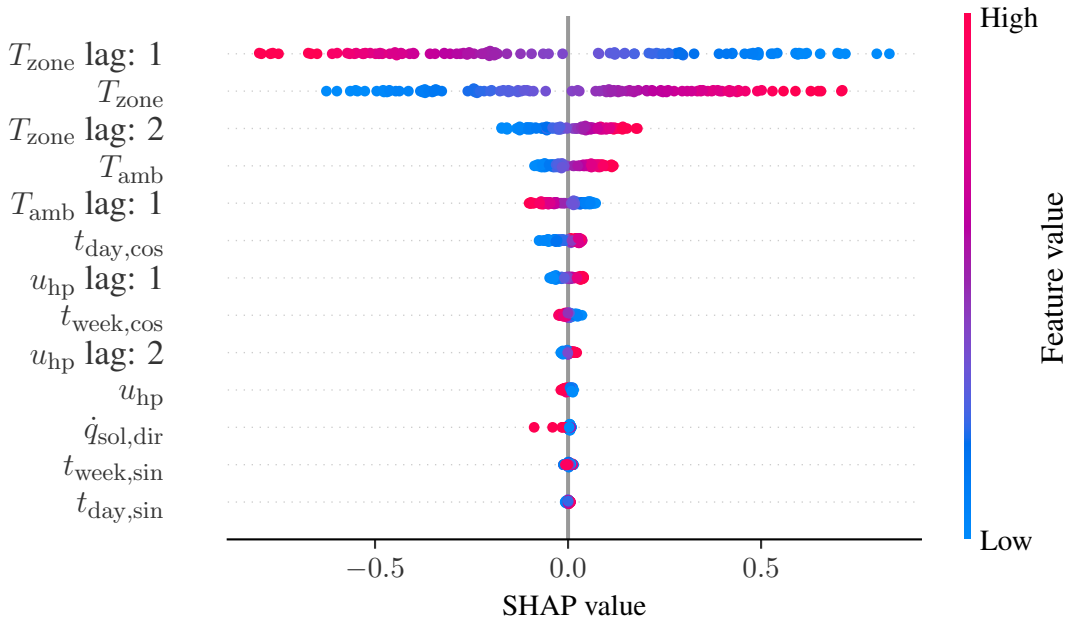

Figure 6: SHAP values for ANN-$\Delta T_{\mathrm{zone,Base}}$.

Another difference between the ANN-$\Delta T_{\mathrm{zone,Base}}$ and the ANN-$\Delta T_{\mathrm{zone,Explo}}$ concerns the feature u_{hp}. While with the explorative ANN, an increase in the control variable increases ΔT_{zone}, the opposite is true for the base ANN, which cannot be explained physically. Thus, u_{hp} is a feature for which the SHAP approach can show that the explorative ANN has learned the influence more reliably. The same can be seen for the solar radiation.

In summary, the previous section shows the differences between the explorative ANNs and the base ANNs by analysing the SHAP values. At first sight, the distribution of the SHAP values appear similar. However, a deeper analysis reveals some differences. Looking at the two ANNs modelling $P_{\mathrm{el,hp}}$, we can see that they learn in different ways whether the heat pump is switched on or off.

Looking at the feature T_{zone}, it is suspected that the base ANN imitates the base controller rather than learning actual correlations. Natale et al. (Di Natale et al., 2022) and Bünning et al. (Bünning et al., 2022) also show results where the prediction models imitate a base controller. For the ANNs modelling ΔT_{zone}, the application of the SHAP methodology reveals that the influences of some features on ΔT_{zone} are not correctly captured by the base ANN. Differences are also visible in the feature time of day in cosine format. The base ANN learns the influence of user behaviour on ΔT_{zone} via this feature, whereas the explorative ANN learns no correlation.

6. Discussion and Conclusion

In this work, the popular explainable artificial intelligence method SHAP is combined with data-driven model predictive control. The aim is to increase the interpretability of artificial neural networks used as process models for MPC during the model creation process. Using the standardised BESTEST Hydronic Heat Pump test case of the *Building Optimization Testing Framework*, an in-depth analysis of how the models make predictions is performed. In addition, models using a base data set and models using an explorative data set with more operating points are compared in terms of control performance. The comparison is interesting to analyse how the training data affect the generalisation ability of ANNs.

The control using the explorative data set outperforms the control using the base data set. The difference in control performance can be explained by analysing the SHAP values. It is shown that the influence of some features on ΔT_{zone} is not correctly captured by the base model. In addition, unimportant features are identified using SHAP. Removing these unimportant features from the model results in an improved control performance.

The results show that it is possible to increase the interpretability of ANNs using the SHAP methodology. This paper shows how the characteristics of a feature affect the prediction. Furthermore, the most important features are identified, and those that only have a minor influence on the prediction. Using the SHAP approach, it is also possible to identify weaknesses in the model. In addition to the analysis of a single ANN, it is shown how the SHAP methodology also allows a comparison between different ANNs.

A critical point to discuss is that the choice of background data can have a significant impact on the SHAP values. In addition, the SHAP approach reveals which correlations the model has learned, but not why the model has learned these correlations. Furthermore, it should not be neglected that the SHAP approach is subject to a crucial simplification, namely the assumption of independence of the features. As a result, unrealistic inputs are generated by the SHAP approach and evaluated by the model.

Future work should discuss the sensitivity of the SHAP methodology to the background data. In addition, the results of SHAP should be compared with other explainable artificial intelligence methods and should be applied to more models for comparison. Finally, the performance of data-driven model predictive control should be further increased. This can be done by improving the feature selection for example by analysing the feature importance as shown in this paper. In addition, physical prior knowledge should be incorporated into the model creation process by using modified neural networks.

Acknowledgments

We gratefully acknowledge the financial support from the Federal Ministry for Economic Affairs and Climate Action (BMWK), promotional reference 03SBE0006A.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion, 58:82–115, 2020.

Jakub Białek, Wojciech Bujalski, Konrad Wojdan, Michał Guzek, and Teresa Kurek. Dataset level explanation of heat demand forecasting ann with shap. Energy, 261:125075, 2022.

David Blum, Javier Arroyo, Sen Huang, Ján Drgoňa, Filip Jorissen, Harald Taxt Walnum, Yan Chen, Kyle Benne, Draguna Vrabie, Michael Wetter, and Lieve Helsen. Building optimization testing framework (boptest) for simulation-based benchmarking of control strategies in buildings. Journal of Building Performance Simulation, 14(5):586–610, 2021.

Felix Bünning, Benjamin Huber, Adrian Schalbetter, Ahmed Aboudonia, Mathias Hudoba de Badyn, Philipp Heer, Roy S. Smith, and John Lygeros. Physics-informed linear regression is competitive with two machine learning methods in residential building mpc. Applied Energy, 310: 118491, 2022.

Zhe Chen, Fu Xiao, Fangzhou Guo, and Jinyue Yan. Interpretable machine learning for building energy management: A state-of-the-art review. Advances in Applied Energy, 9:100123, 2023.

Chika E. Ugwuanyi. Using Interpretable Machine Learning for Indoor CO2 Level Prediction and Occupancy Estimation: PhD Thesis. 2021.

L. Di Natale, B. Svetozarevic, P. Heer, and C. N. Jones. Physically consistent neural networks for building thermal modeling: Theory and analysis. Applied Energy, 325:119806, 2022.

Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L. Vrabie, and Lieve Helsen. All you need to know about model predictive control for buildings. Annual Reviews in Control, 50: 190–232, 2020.

Ján Drgoňa, Aaron R. Tuor, Vikas Chandan, and Draguna L. Vrabie. Physics-constrained deep learning of multi-zone building thermal dynamics. Energy and Buildings, 243:110992, 2021.

Tryambak Gangopadhyay, Sin Yong Tan, Zhanhong Jiang, Rui Meng, and Soumik Sarkar. Spatiotemporal attention for multivariate time series prediction and interpretation, 2020.

Tadeusz A. Grzeszczyk and Michal K. Grzeszczyk. Justifying short-term load forecasts obtained with the use of neural models. Energies, 15(5):1852, 2022.

Achin Jain, Truong Nghiem, Manfred Morari, and Rahul Mangharam. Learning and control using gaussian processes. In 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS), pages 140–149. IEEE, 2018. ISBN 978-1-5386-5301-2.

Anjukan Kathirgamanathan, Mattia de Rosa, Eleni Mangina, and Donal P. Finn. Data-driven predictive control for unlocking building energy flexibility: A review. Renewable and Sustainable Energy Reviews, 135:110120, 2021.

Olivera Kotevska, Jeffrey Munk, Kuldeep Kurte, Yan Du, Kadir Amasyali, Robert W. Smith, and Helia Zandi. Methodology for interpretable reinforcement learning model for hvac energy control. In 2020 IEEE International Conference on Big Data (Big Data), pages 1555–1564. 2020.

Krzysztof Arendt, Muhyiddine Jradi, Hamid Reza Shaker, and Christian Veje. Comparative analysis of white-, gray- and black-box models for thermal simulation of indoor environment: Teaching building case study. In Proceedings of the 2018 Building Performance Modeling Conference and SimBuild co-organized by ASHRAE and IBPSA-USA, pages 173–180. ASHRAE, 2018.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. volume 30. 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

R. Machlev, L. Heistrene, M. Perl, K. Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities. Energy and AI, 9:100169, 2022.

Jianqiao Mao, Ryan Grammenos, and Konstantinos Karagiannis. Data analysis and interpretable machine learning for hvac predictive control: A case-study based implementation. Science and Technology for the Built Environment, 29(7):698–718, 2023.

Christoph Molnar. Interpretable machine learning: A guide for making Black Box Models interpretable. Lulu, Morisville, North Carolina, 2019. ISBN 978-0-244-76852-2.

M. Z. Naser. An engineer's guide to explainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference. Automation in Construction, 129:103821, 2021.

R. Judkoff and J. Neymark. International energy agency building energy simulation test (bestest) and diagnostic method, 1995. URL http://dx.doi.org/10.2172/90674.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? In Balaji Krishnapuram, Mohak Shah, Alex Smola, Charu Aggarwal, Dou Shen, and Rajeev Rastogi, editors, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016. ISBN 9781450342322.

Phillip Stoffel, Patrick Henkel, Martin Rätz, Alexander Kümpel, and Dirk Müller. Safe operation of online learning data driven model predictive control of building energy systems. Energy and AI, 14:100296, 2023a.

Phillip Stoffel, Laura Maier, Alexander Kümpel, Thomas Schreiber, and Dirk Müller. Evaluation of advanced control strategies for building energy systems. Energy and Buildings, 280:112709, 2023b.

David Sturzenegger, Dimitrios Gyalistras, Manfred Morari, and Roy S. Smith. Model predictive climate control of a swiss office building: Implementation, results, and cost–benefit analysis. IEEE Transactions on Control Systems Technology, 24(1):1–12, 2016.

United Nations Environment Programme. 2022 global status report for buildings and construction: Towards 2022 global status report for buildings and construction: Towards a zero-emission, efficient and resilient buildings and construction sector. Nairobi, 2022.

Min Gyung Yu and Gregory S. Pavlak. Extracting interpretable building control rules from multi-objective model predictive control data sets. Energy, 240:122691, 2022.