energies

MDPI

Article

Forecasting Energy Consumption of a Public Building Using Transformer and Support Vector Regression

Junhui Huang and Sakdirat Kaewunruen *

Department of Civil Engineering, School of Engineering, University of Birmingham, Birmingham B15 2TT, UK
* Correspondence: s.kaewunruen@bham.ac.uk

Abstract: Most of the Artificial Intelligence (AI) models currently used in energy forecasting are traditional and deterministic. Recently, a novel deep learning paradigm, called 'transformer', has been developed, which adopts the mechanism of self-attention. Transformers are designed to better process and predict sequential data sets (i.e., historical time records) as well as to track any relationship in the sequential data. So far, a few transformer-based applications have been established, but no industry-scale application exists to build energy forecasts. Accordingly, this study is the world's first to establish a transformer-based model to estimate the energy consumption of a real-scale university library and benchmark with a baseline model (Support Vector Regression) SVR. With a large dataset from 1 September 2017 to 13 November 2021 with 30 min granularity, the results using four historical electricity readings to estimate one future reading demonstrate that the SVR (an R^2 of 0.92) presents superior performance than the transformer-based model (an R^2 of 0.82). Across the sensitivity analysis, the SVR model is more sensitive to the input close to the output. These findings provide new insights into the research area of energy forecasting in either a specific building or a building cluster in a city. The influences of the number of inputs and outputs related to the transformer-based model will be investigated in the future.

Keywords: CO_2 emissions; energy consumption; transformer; machine learning; building energy performance; building physics; net zero energy building; artificial intelligence

Citation: Huang, J.; Kaewunruen, S. Forecasting Energy Consumption of a Public Building Using Transformer and Support Vector Regression. *Energies* **2023**, *16*, 966. https://doi.org/10.3390/en16020966

Academic Editor: Jesús Manuel Riquelme-Santos

Received: 6 December 2022
Revised: 4 January 2023
Accepted: 11 January 2023
Published: 15 January 2023

1. Introduction

Overpopulation has been considered a global problem leading to a surge in energy consumption [1]. World Energy Outlook 2022 by the International Energy Agency (IEA) reports that electricity forms around 20% of the world's total energy consumption. Due to many countries' pervasive electric vehicles and heat pumps, this trend could still go up by 25–30% by 2030. At the same time, the growth of economies relies on energy consumption, proliferating the demand for energy [2]. However, the rising demand for energy sparks considerable carbon emissions, which are harmful to the environment [3,4]. It is a difficult trade-off between economic growth and a green environment. Hence, there is a requirement to enhance energy efficiency by forecasting energy consumption. Energy prediction is an approach to building net-zero concepts. It analyses building consumption in the future with considerations of the environmental impacts. In addition to carbon neutrality, public buildings' electricity decarbonisation is imperative [5]. Forecasting energy consumption is beneficial for decision-makers and policy-makers such as governments. Determining how much energy will be consumed may assist in minimising energy consumption. Predicting future energy consumption in the short and long term will enable developers to understand which form of energy is most widely utilised and attempt to reverse the trend. Different variables, such as water, wind, and temperature, influence the quantity of energy used in various places. Forecasting energy can be challenging with various variables [6].

From air conditioners to light bulbs to power switches to intelligent interfaces, electricity is supplied to operate different forms of equipment. While all are energy consumers,

some equipment has excessive energy consumption. Mitigating such a high consumption becomes even more critical, with a need to set effective solutions and measures. Nevertheless, setting such feasible solutions manually may be useless as they may be set based on biased perspectives. It may be challenging for a human to consider interchangeable and complicated variables regarding the energy-consuming factors at one time, leading to setting inefficacious energy reduction measures. In the past, physical modelling techniques (also known as white box models or engineering approaches) were used to predict energy demand. Physical models established statistical models mainly based on thermodynamic principles. Some physical-based software such as eQuest and EnergyPlus relied on the buildings' thermodynamic knowledge and environmental factors such as operation schedules, building construction details, heating, ventilation, air conditioning, and weather conditions [6]. However, this is not always a solution for predicting energy demand owing to a lack of apparatuses in buildings recording the required information for the physical models. On the other hand, physical models are laborious, which can be more dominant for a modern big building.

To alleviate such laborious methods, data-driven methods referring to AI do not need to include buildings' energy consumption details instead of extracting patterns from historical data. The blooming of Artificial Intelligence enhances unprecedented advances in many domains. AI models have achieved great leaps in energy forecasting using a wide range of features, from simple date and time information to more complex features such as weather conditions and building designs. Much of the attention has focused on the more traditional algorithms referring to Artificial Neural Networks (ANNs) [7], Recurrent Neural Networks (RNNs) [8], Convolutional Neural Networks (CNNs) [9], Support Vector Regression (SVR) [10], etc. According to a review for energy consumption forecasting investigating conventional models and artificial-intelligence-based models [11], most of the studies (48% within 129 studies) were favourable to AI models, followed by 43% of the researchers using conventional models and the least number of users for other models. Collapsing across the AI models, 77% were ANNs, and SVR accounted for 23%. Very little about the transformer using self-attention for forecasting energy consumption is known. To address this gap, here, we shed light on the role of the transformer—a novel, simple network based on an attention mechanism showing superior performance and making a great stride in Natural language processing (NLP). For example, Ref. [12] achieved a leap in machine translation with features of more parallelisable and less time required to train. In [13], the authors proposed a pre-trained transformer-based model, BERT—Bidirectional Encoder Representations from Transformers, which delivered an unprecedented performance in 11 NLP tasks.

Motivated by the great success of transformers in the NLP domain, the transformer technique has been dominant in the computer vision area. Convolutional Neural Networks (CNNs) draw most of the attention in computer vision, such as [14,15], but transformers became a player in computer vision, such as [16] leveraged transformers for object detection and [17] for pixels prediction. Either of them in [16,17] presented comparable or better performances compared to CNNs. More recent studies shared similar ways with us. Saoud et al. utilised a hybridisation of stationary wavelet transform (SWT) and transformers to predict five households' power consumption in various resolutions (5 min and 10 min) [18]. Their results found that their proposed method outperformed the existing approaches such as a deep transformer, Long Short-Term Memory (LSTM), LSTM-CNN, support vector machine (SVM), and LSTM-SWT. Considering the limitation exhibited in [18], it is sensible to enlarge the size and variability of input data to ameliorate the model's generality. More importantly, a large and variable dataset is used in our study to evaluate a relatively new model in the energy consumption prediction domain—transformer—and a more traditional model—SVR.

This study provides insight into if the self-attention mechanism is helpful in the energy forecasting domain with a comprehensive dataset (1 December 2017—25 March 2021 with 30 min granularity) of the University of Birmingham library. The possible influence of

the number of inputs and outputs was discussed in our previous study [19], signifying that multiple outputs hindered the model's capability, especially since the number of outputs was large. Hence, we choose four historical timestamps to predict one following timestamp as a subject. SVR shows superior performance in [19] compared to Long Short-term Memory (LSTM) [20] and Extreme Gradient Boosting (XGBoosting) [21]. Therefore, SVR is used as a benchmarking model here.

The contribution of this study can be summarised in the following points:

1. The performance of the transformer-based model is estimated in the energy forecasting domain, which can provide some insights into the comparison between a conventional model SVR and the transformer-based model.
2. The sensitivity analysis's outcome shows there is no need to proliferate the input length as the prediction is insensitive to the data point far away.

The remainder of the paper is organised as follows: Section 2 unveils details of the dataset, how the raw data are pre-proceeded, knowledge of the transformer model and SVR, hyperparameter tuning, and evaluation criteria. Section 3 provides results for estimating the next 30 min, discussions with existing counterparts, and sensitivity analysis of the SVR. Section 4 provides conclusions, limitations, and future directions.

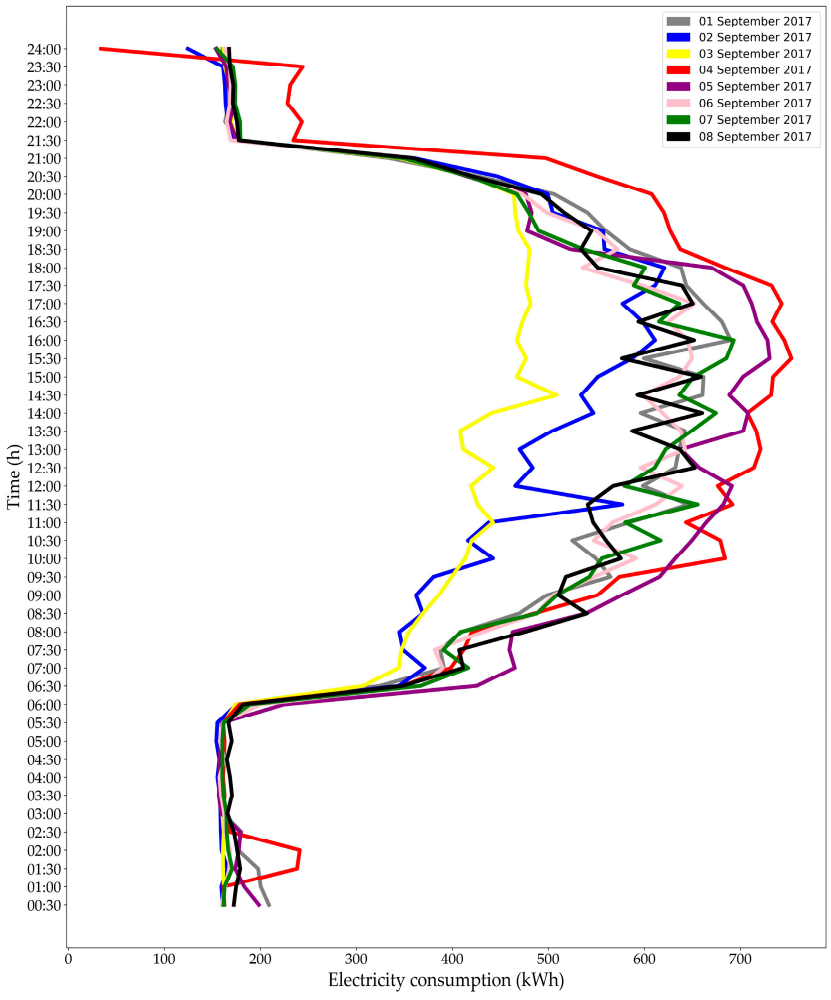
2. Materials and Methods

The Energy Performance of Building Directive and the further recasts envisaged a standard [22]—the Energy Performance Certificates (EPCs) to rate buildings' energy performance. However, the major obstacle to using EPCs has been the absence of the reliability of data collection approaches. Therefore, Salvalai G and Sesana MM evaluated the impact of the different methods to collect energy data [23], yielding that three levels of monitoring, namely the basic level, medium level, and advanced level, were required for different purposes. It is worth mentioning that the main library shown in Figure 1 uses the advanced level specified in [23], as there is sub-metering to monitor each room's energy consumption.

Figure 1. Main library at the University of Birmingham.

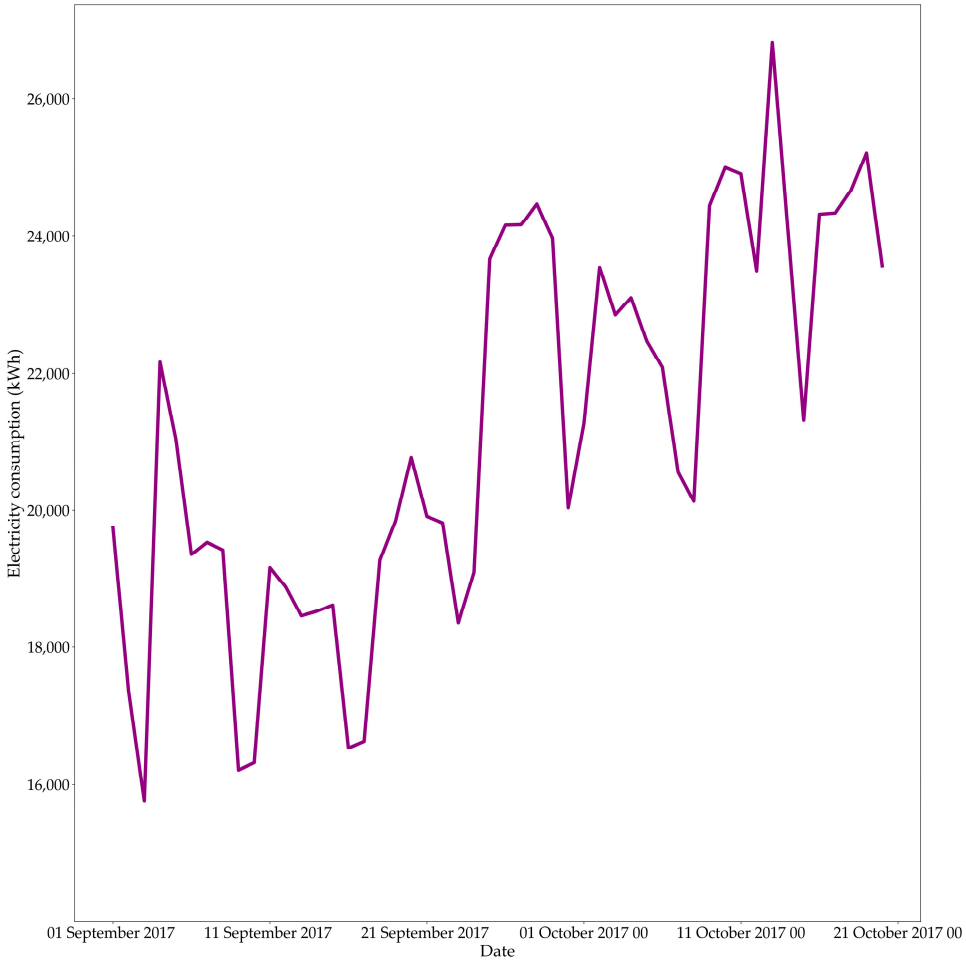
Electrical consumption in the library in Figure 1 constantly fluctuates throughout the day, which can be affected by a couple of factors, such as materials used in buildings, ventilation systems, heating systems, etc. These factors can be challenging to analyse due to most buildings' lack of energy-measuring technologies. However, the main library at the University of Birmingham is one of the largest academic libraries in the UK, which is

equipped with countless intelligent sensors to adjust ventilation, lighting, and air conditioning based on circumambient factors such as temperature, visibility, and other weather conditions. More importantly, the intelligent and modern facility allows the record of electricity usage at room level, providing an excellent dataset to study the energy forecasting model's efficacy. Figure 2a exemplifies the first eight days of electricity usage, showing that electricity consumption significantly fluctuates during the day. It is straightforward in the quiet period from 21:30 to 06:00: most energy consumptions fall into the range between 150 and 200 kWh. However, the rest of the day seems very hard to follow. The reason could be that the electricity consumption is recorded from 94 different rooms in the library, which implies enormous variability can be possible. Figure 2b glances at the variability of the dataset between days, also signifying a significant variability of the dataset. Within the 50 days, the most significant difference is 12,000 kWh resulting from the peak of 27,000 kWh and the bottom of 15,000 kWh. Therefore, this study can complement the study in [16] to further examine the potential of the transformer. At the same time, we do not consider the weather a necessary precondition. Hence, we seek to characterise the inherent electricity usage pattern feature, as seen in Figure 3, which results from all attributes affecting electricity consumption.



(**a**)

Figure 2. *Cont.*



(**b**)

Figure 2. Electricity consumption at two resolutions (**a**,**b**) overlook of the dataset (**a**) using 30 min resolution in the first eight days and (**b**) leveraging a day resolution in the first 50 days. The rationale for only eight days provided in (**a**) is the clear reading of the figure.

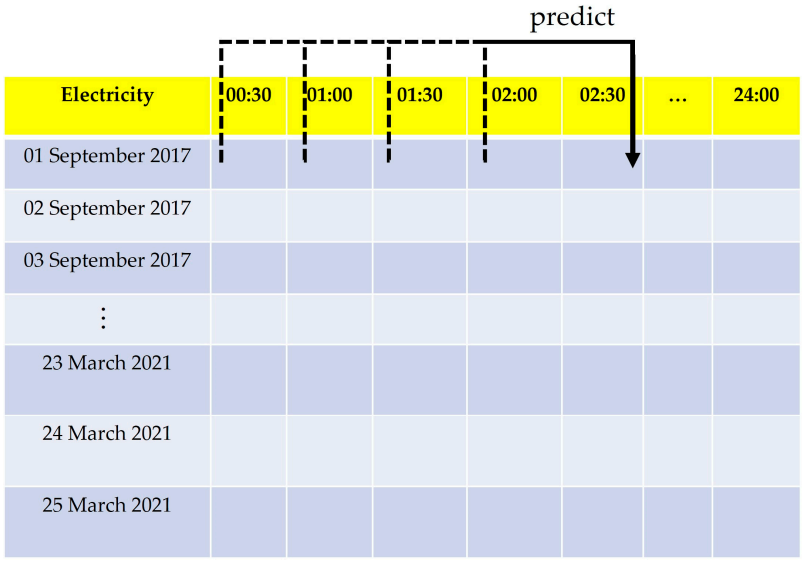


Figure 3. Four historical time stamps to predict one following timestamp.

2.1. Pre-Process Data

This section presents how we clear the dataset (addressing null values and checking the consecutive of the dataset) and data formatting (transforming and formatting the dataset to be used by Machine Learning (ML) models).

Figure 4 unveils the shape of the dataset after we remove the null values. The skip timeslot is not considered in this study, so we also check if the dataset is consecutive by checking the gap between every reading date and the previous date. The rationale for this action is to avoid biases introduced by the Nan value processing, such as the Nan value being replaced by the mean value or other methods. The whole dataset ranges from 1 September 2017 to 13 November 2021, but the section after 25 March 2021 is incomplete due to Nan values, so the incomplete part is removed. The summation of columns can manipulate the resolution to a larger granularity. However, the impact of granularity is not in our scope of interest.

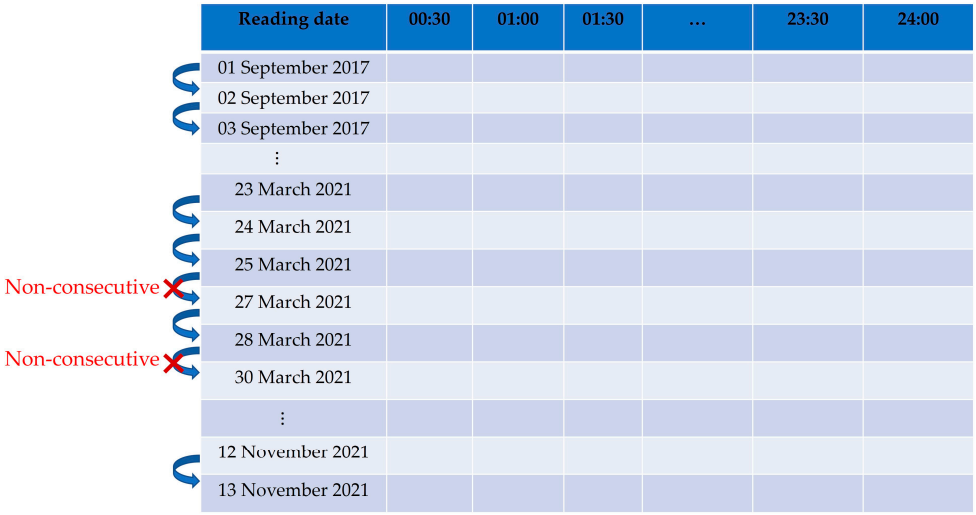


Figure 4. Dimension of the dataset.

2.2. Transformer

In this section, we outline the essential details to unfold how the transformer uses self-attention to extract information from input and some techniques, such as skip connection and layer normalisation, to help train the model.

2.2.1. Self-Attention

By addressing data in sequence problems, LSTM, RNN, and Gated Recurrent Unit (GRU) have been established and have shown compelling results. However, RNN suffered from the vanishing gradient problem [24], where LSTM and GRU mitigated but eliminated the issue. Considering the constraint of the RNN that the result at t is dependent on the result at t–1, the RNN is hard to parallel. A simple example in Figure 5a illustrates that y_3 is calculated by x_0 to x_3 in sequence but in parallel. For the more extended sequence, the RNN tends to lose some information due to the vanishing gradient problem, as the weight easily vanishes and explodes when the sequence is long.

To address the present lack of parallel calculation and memory loss due to long sequences, Ref. [12] proposed a self-attention mechanism to calculate the correlation. Figure 6 shows a block of self-attention where all samples are converted to queries, keys, and values (q, k, v in Figure 6) by the three matrices (W^q, W^k, and W^v). It is sensible to introduce more sets of the three matrices to increase the number of attentions that can extract more information from the input.

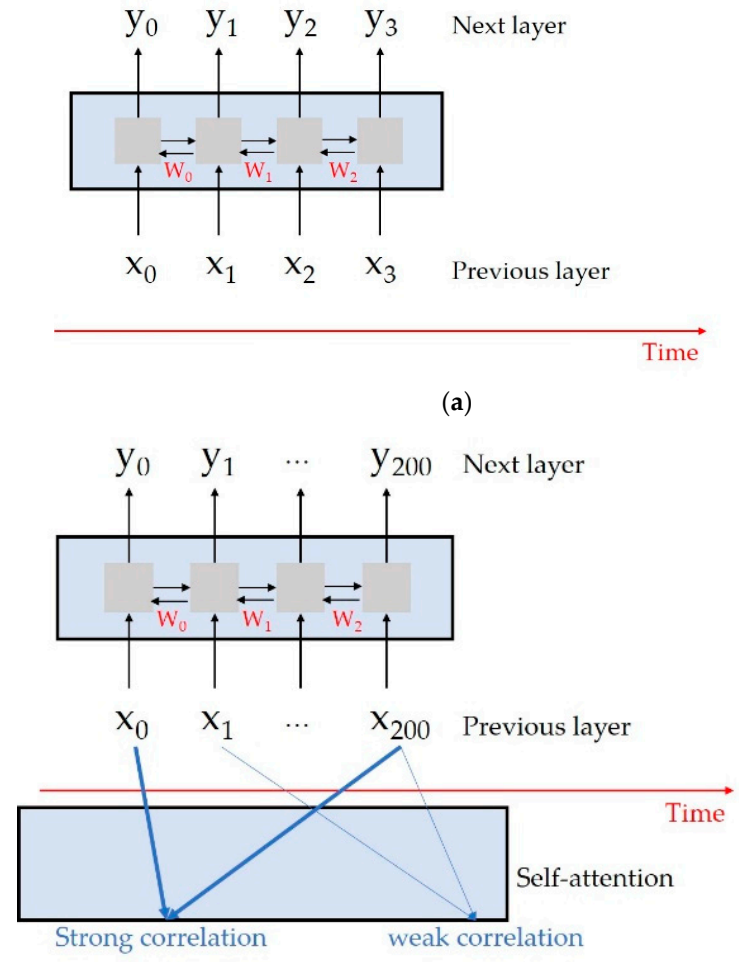


Figure 5. (**a**) shows the RNN mechanism that results at a later timestamp must be calculated in sequence. (**b**) illustrates a long sequence using a self-attention mechanism. To prevent information loss due to a very long sequence, self-attention computes the correlation between each input by assigning substantial weight to the inputs that have a strong correlation.

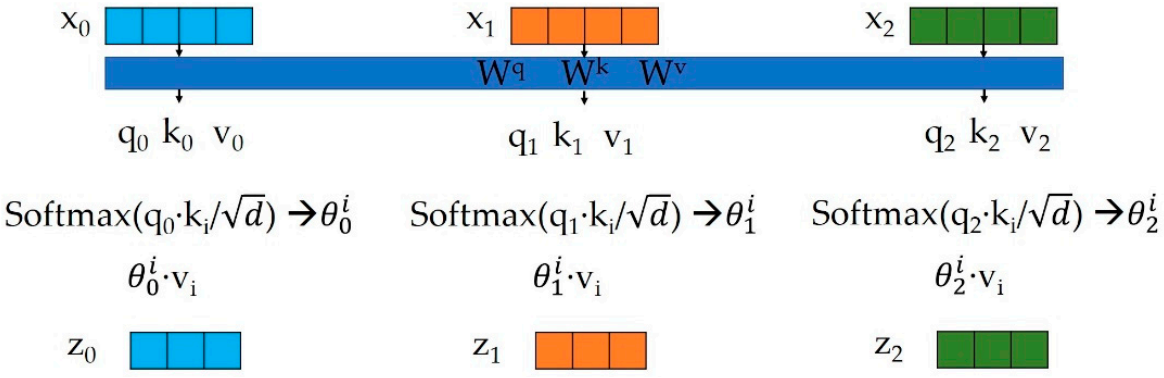


Figure 6. A block of self-attention.

Here, we show how self-attention works by using three samples. All samples share the same three matrices, W^q, W^k, and W^v, to gain each sample's query (q), key (k), and value (v) using Equations (1)–(3). To evaluate the interplay between each pair of samples, each sample's query dot products each sample's value to obtain the weights (θ) related to each sample before being converted to weight by the softmax function in Equation (4).

Noting that $\sqrt{d}$ is an empirical value to reduce the value inside the SoftMax function to prevent gradient vanishing, Equation (5) outputs the vector after the self-attention layer for each sample.

$$q_i = w^q \, x_i \tag{1}$$

$$k_i = w^k \, x_i \tag{2}$$

$$v_i = w^v \, x_i \tag{3}$$

$$\theta_0^i = \text{softmax}\left(\frac{q_0 \, k_i^T}{\sqrt{d}}\right) \tag{4}$$

$$z_0 = \theta_0^i \, v_i \tag{5}$$

2.2.2. Skip Connection and Layer Normalisation

Figure 7 exemplifies a skip function and normalization. The skip function was proposed by [14] specifically for a very deep network where the error can be minor in the deep layer leading to a difficulty that the model cannot learn from the minor error. This concern can be solved by a skip connection that can propagate the error from the shallow layer to the deep layer.

Normalisation is deployed to reduce the training time and the network's overfitting issue, especially for the deep neural network haunted by these two issues. Unlike batch normalisation [25], layer normalisation presents a more straightforward mechanism due to no new dependencies being introduced between training cases [26].

$$\mu^t = \frac{1}{N} \sum_{i=1}^{N} x_i^t \tag{6}$$

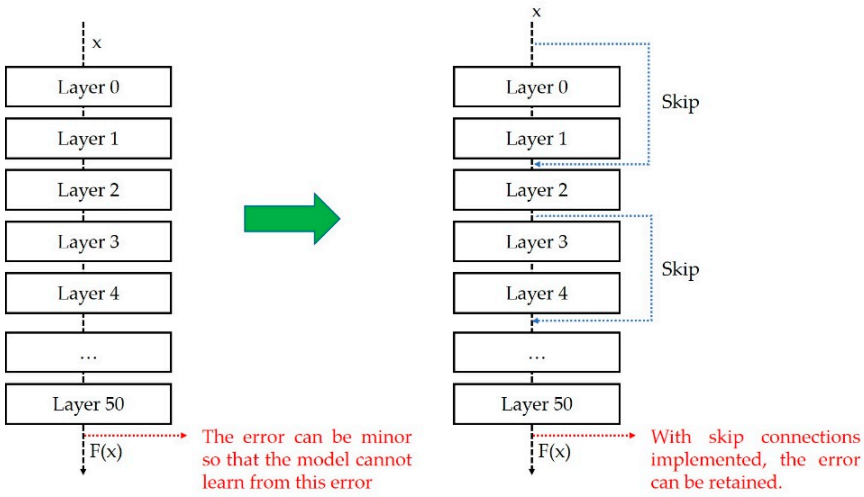
$$\sigma^t = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i^t - u^t)^2} \tag{7}$$

$$h^t = f[\, \frac{g}{\sigma^t} \odot \sum_{i=1}^{N} a^t - \mu^t) + b] \tag{8}$$

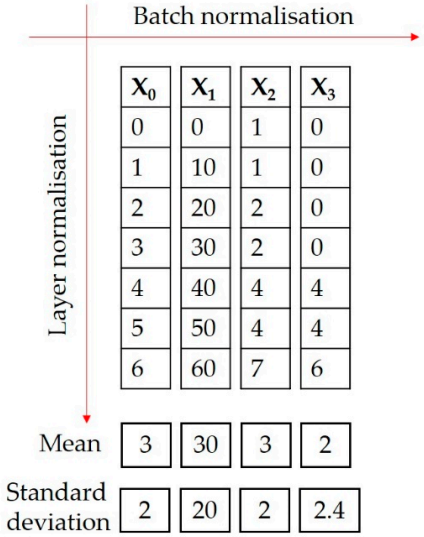
where μ is the mean value, σ is the standard deviation, t is the number of layers, h is the normalised value, and g and b are the learnable values to scale the normalised value.

2.3. SVR

Support vector machines were first developed by Vapnik [27] specifically for classification problems. SVMs aim to locate the maximum margin to separate the classes, namely hyperplane, which generally implies classifying as many training points as possible [28]. SVR, the variant of SVMs, was designed for regression problems by including an ε-tube, as shown in Figure 8. SVR's optimisation problem is now reformulated to find the ε-tube that best fits the training samples. The red samples outside the ε-tube are ditched, but the samples within the tube region that do not fall on the estimated line can still be retained and receive no penalty. One major strength of SVR is that its computational complexity is insensitive to the number of input features. In addition, SVR is selected for its accuracy and excellent generalizability.



(**a**)



X_0	X_1	X_2	X_3
0	0	1	0
1	10	1	0
2	20	2	0
3	30	2	0
4	40	4	4
5	50	4	4
6	60	7	6
3	30	3	2
2	20	2	2.4

(**b**)

Figure 7. (**a**) Skip function adds another path to pass the output of the shallow layer to the deep layer; (**b**) layer normalisation uses the mean and standard value procured along vertical direction by Equations (6) and (7), but batch normalisation uses horizontal direction.

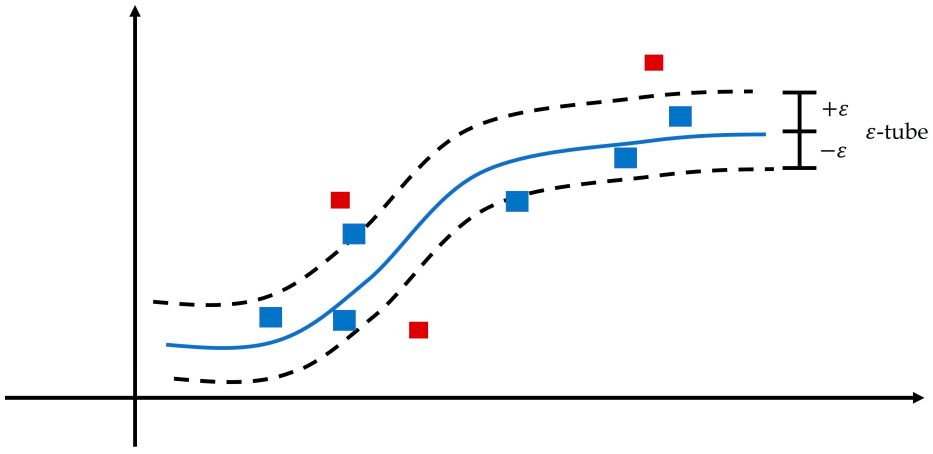


Figure 8. SVR.

2.4. Hyperparameter Tuning

Apart from the learnable parameters such as weights and biases, some parameters known as hyper-parameters are not learnable but pre-defined. To ensure the model can deliver optimal performance, it is necessary to conduct a hyper-parameter tuning process manually or by tuning techniques. Using the optimal subset of hyper-parameters for AI models directly influences models' performance. Manually selecting the best hyperparameters demands a comprehensive understanding of AI algorithms and good hyperparameter tuning capabilities. There are some automatic tuning techniques available. However, practical evaluation before choosing one is required as they have different benefits and drawbacks in various cases.

A hyperparameter tuning technique employed in the study is random search due to the time-saving and productivity compared to manual tuning and other methods such as grid search [29]. Grid search tends to take a long time. However, it shows no apparent benefit over the random search, especially when the search space is ample, according to research carried out by Bergstra et al. using seven different datasets trained on neural networks [30]. In Bergstra et al.'s findings, the random search showed equivalent or comparable results among four out of seven datasets and superior performance in one of seven datasets using the lower computational budget. It is possible that the random search can also outperform using the four datasets if the same computational budget as the grid search is provided for the random search. Table 1 illustrates the pre-defined space for the random search to identify the optimal subset. It is worth mentioning that SVR's kernel function RBF is used to alleviate the overfitting—a common issue that the model overfits the training set and lacks generalisation in other unseen or new datasets [31].

Table 1. The searching space of the hyperparameters.

Algorithms	Hyperparameters	Searching Space
Transformer	The No. of attention	1–256
	The no. of units for the feedforward layer	32–256
	The no. of attention blocks	1–20
	Dropout rate	1×10^{-1}–0.9×10^{-1}
	Learning rate	1×10^{-1}–1×10^{-6}
SVR	Epsilon	1×10^{-2}–2×10^{-1}
	C	1–2000
	Kernel	RBF

2.5. Metrics

Motivated by the ubiquitous method, ML models, used in the energy forecasting domain, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) unified a metric, the coefficient of variation (CV), to interpret and benchmark the model's result [32]. At the same time, the model's performance in the study is also evaluated by the Root Mean Squared Error (RMSE) [33] and coefficient of determination (R^2) [34] to circumvent potential biased interpretation of the result. All three metrics are given below.

$$CV = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_I)^2}}{\overline{y}} \times 100 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2} \tag{11}$$

where y_i indicates the actual value; $\hat{y}_I$ is the predicted value; $\overline{y}$ is the mean value of actual energy consumption; N refers to the number of samples.

3. Results and Discussions

Table 2 presents results from two optimal models based on the testing dataset (70% training set, 10% validation set, and 20% testing set) using the tuned values shown in Table 3. SVR procures the least CV (12%), followed by 18% gained by the transformer. During the training process, the transformer and SVR are fed with [X_{t-3}, X_{t-2}, X_{t-1}, X_t] to predict X_{t+1}, where X is the dataset and t refers to the timestamp right now. A widely used optimiser—Adam optimiser [35]—is adapted with the benefits of being computationally efficient, having little memory requirements, and being suitable for problems with large datasets. A key hyper-parameter learning rate from 0.1 to 1×10^{-6} with log sampling is set to find the optimal learning rate for the model. The transformer model is regularised by the dropout layer from overfitting in a range of 1×10^{-1} to 0.9×10^{-1} with an increment of 0.1. The transformer's architecture is delivered and implemented by Keras backend [36] and Tensorflow [37], while SVR is implemented with scikit-learn [38].

Table 2. Regression performances for the two models.

Models	CV	R^2	RMSE
Transformer	17.0652%	0.8238	76.9611
SVR	12.1949%	0.9196	54.9363
SVR in [18]	N/A	N/A	0.0296
The transformer in [18]	N/A	N/A	0.0182
The proposed method in [18]	N/A	N/A	0.009

Table 3. Optimal hyperparameters for the two models.

Algorithms	Hyperparameters	Tuned Value
Transformer	The No. of attention	256
	The no. of units for the feedforward layer	224
	The no. of attention blocks	4
	Dropout rate	0.2
	Learning rate	0.0013
SVR	Epsilon	0.089
	C	19
	Kernel	RBF

In view of past work [39], electricity consumption is weather-dependent. Therefore, they used 140 sensors to monitor weather attributes and achieved 20.05% CV. Only one feature is used here and delivers better results than [10], implying that weather attributes are not a precondition to predict future electricity load. At the same time, it is noted that SVR outperforms the transformer with about 10% more R^2, 20 kWh less RMSE, and 5% less CV.

In counterpart [18], they also used transformer and SVR to predict five households' energy consumption in multiple time granularities—5 min, 10 min, 20 min, and 30 min. The dataset acquired by the UK-DALE project [40] in 2015 was used to develop their models. In the following, we discuss and compare our results with the result of the resolution of 30 min in [18]. The common metric is RMSE between our study and [18], which delivered an optimal RMSE of 0.0296 for the SVR, 0.0182 for the transformer, and 0.009 for the transformer + SWT (proposed by them). It is noticeable that the RMSEs of our models are considerably larger than those in [18] because the RMSE is a scale-dependent measure, and it is not comparable to the different cases using RMSE. This is one of the reasons why CV, a measure independent of the unit or widely different means, has been introduced. However, in our case, the SVR shows superior performance than the transformer, which is the other way around in [18].

Sensitivity analysis is only conducted on the outperformed model—SVR—to showcase the main contributing factor to the prediction. Dimopoulos and Bakas [41] used a variant method from [42,43] to analyse and characterise each independent variable's impact on the dependent variable. A significant advantage of sensitivity analysis conducted here is that we can examine if the feature close to the prediction has a more significant effect. Figure 9 unveils how the sensitivity analysis factor is determined. The original dataset is modified to (x_0, x^0_{mean}, x^1_{mean}, x^2_{mean}), which subsequently is used to train the optimal SVR to gain y^0_{max} and y^0_{min}. Equations (12) and (13) are given to calculate the corresponding sensitivity analysis for the features.

$$I_n = y^n_{max} - y^n_{min} \tag{12}$$

$$S_n = \frac{I_n}{\sum_n I_n} \tag{13}$$

where I is the difference between the estimated electricity consumption, S is the sensitivity analysis factor, and n is the feature. Figure 10 shows the sensitivity analysis results from the timestamp far away (X_0) and closest (X_3) to the predicted values. Although all four features see similar sensitivity to electricity consumption, the impact of the features close to the predicted value still presents a more significant impact on the prediction.

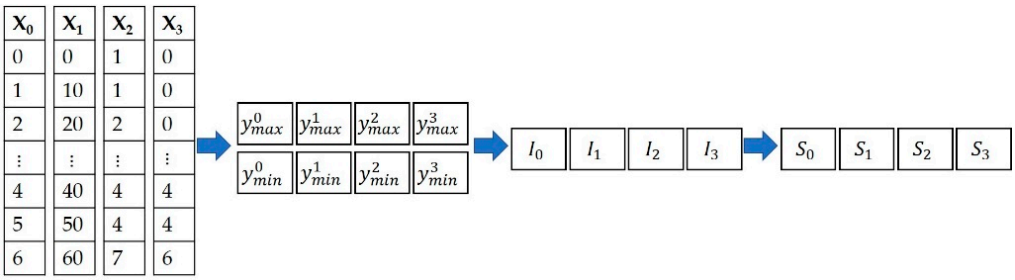


Figure 9. Sensitivity analysis.

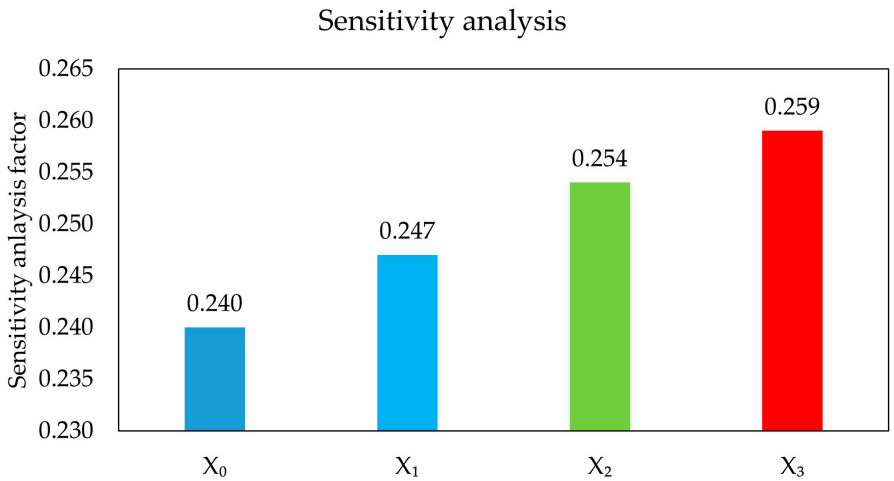


Figure 10. Sensitivity analysis parameters for electricity consumption.

4. Conclusions

This study is designed to determine the real-world efficacy of the transformer model in the energy forecasting domain, as the transformer shows a dominant performance in language translation compared to the RNN. The transformer also allows parallelisation that the RNN is not able to provide. Comparing the transformer and SVR yields a conclusion that the transformer fails to provide outstanding results as it does in other domains. The second significant finding is that the predicted value is proportional to its closeness found in the sensitivity analysis section. This study also underpins and strengthens the finding in our previous research [19] that the weather feature is not a prerequisite to forecast

energy demand for a public residential building or a public building such as a library. These findings contribute to existing knowledge of a transformer-based model used in the energy forecasting area by proving that the transformer does not perform better than a traditional SVR.

A sound model for predicting electricity consumption benefits the supplier, the end-user, and the environment. It helps save unnecessary costs and resources from undesigned activities of the buildings and the power plants. There is no consideration of the length of the input and output. However, this can be important to the transformer as the mechanism is designed for a long sequence which can be a future investigation. It is also unfortunate that the study does not include ensemble methods in which the combination of the transformer-based model and the SVR model can achieve an enhancement. Predicting other utility services, such as water and natural gas, will be a fruitful area for further work.

Author Contributions: Conceptualisation: S.K. and J.H.; Investigation: S.K. and J.H.; Methodology: S.K. and J.H.; Data Analysis: J.H., Validation: S.K. and J.H.; Draft: S.K. and J.H.; Review and Editing: S.K. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Commission, grant number 691135. The APC is sponsored by MDPI's Invited Paper Initiative.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are sincerely grateful to European Commission for the financial sponsorship of the H2020-MSCA-RISE Project No. 691135 "RISEN: Rail Infrastructure Systems Engineering Network," which enables a global research network that tackles the grand challenge of railway infrastructure resilience and advanced sensing in extreme environments [44].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Uniyal, S.; Paliwal, R.; Kaphaliya, B.; Sharma, R.K. Human Overpopulation: Impact on Environment. In *Megacities and Rapid Urbanization: Breakthroughs in Research and Practice*; IGI Global: Hershey, PA, USA, 2020; pp. 20–30.
2. Mahi, M.; Phoong, S.W.; Ismail, I.; Isa, C.R. Energy–finance–growth nexus in ASEAN-5 countries: An ARDL bounds test approach. *Sustainability* **2019**, *12*, 5. [CrossRef]
3. Alola, A.A.; Joshua, U. Carbon emission effect of energy transition and globalization: Inference from the low-, lower middle-, upper middle-, and high-income economies. *Environ. Sci. Pollut. Res.* **2020**, *27*, 38276–38286. [CrossRef] [PubMed]
4. Gu, W.; Zhao, X.; Yan, X.; Wang, C.; Li, Q. Energy technological progress, energy consumption, and CO_2 emissions: Empirical evidence from China. *J. Clean. Prod.* **2019**, *236*, 117666. [CrossRef]
5. Xiang, X.; Ma, M.; Ma, X.; Chen, L.; Cai, W.; Feng, W.; Ma, Z. Historical decarbonization of global commercial building operations in the 21st century. *Appl. Energy* **2022**, *322*, 119401. [CrossRef]
6. Zhao, H.-x.; Magoulès, F. A review on the prediction of building energy consumption. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3586–3592. [CrossRef]
7. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, e00938. [CrossRef]
8. Dash, S.K.; Roccotelli, M.; Khansama, R.R.; Fanti, M.P.; Mangini, A.M. Long Term Household Electricity Demand Forecasting Based on RNN-GBRT Model and a Novel Energy Theft Detection Method. *Appl. Sci.* **2021**, *11*, 8612. [CrossRef]
9. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74. [CrossRef]
10. Maldonado, S.; González, A.; Crone, S. Automatic time series analysis for electric load forecasting via support vector regression. *Appl. Soft Comput.* **2019**, *83*, 105616. [CrossRef]
11. Wei, N.; Li, C.; Peng, X.; Zeng, F.; Lu, X. Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. *J. Pet. Sci. Eng.* **2019**, *181*, 106187. [CrossRef]
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

13. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef] [PubMed]
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020.
17. Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; Sutskever, I. Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning (PMLR), Virtual, 13–18 July 2020.
18. Saoud, L.S.; Al-Marzouqi, H.; Hussein, R. Household Energy Consumption Prediction Using the Stationary Wavelet Transform and Transformers. *IEEE Access* **2022**, *10*, 5171–5183. [CrossRef]
19. Huang, J.; Algahtani, M.; Kaewunruen, S. Energy Forecasting in a Public Building: A Benchmarking Analysis on Long Short-Term Memory (LSTM), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) Networks. *Appl. Sci.* **2022**, *12*, 9788. [CrossRef]
20. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
21. Chen, T.; He, T. *Xgboost: Extreme Gradient Boosting, R package version 0.4-2*; The XGBoost Contributors: San Francisco, CA, USA, 2015.
22. European Union. *Directive 2010/31/Eu of the European Parliament and of the Council of 19 May 2010 on the Energy Performance of Buildings*; European Union: Brussels, Belgium, 2010; pp. 13–35.
23. Salvalai, G.; Sesana, M.M. Monitoring Approaches for New-Generation Energy Performance Certificates in Residential Buildings. *Buildings* **2022**, *12*, 469. [CrossRef]
24. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **1998**, *6*, 107–116. [CrossRef]
25. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015.
26. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
27. Vapnik, V. The Support Vector Method of Function Estimation. In *Nonlinear Modeling*; Springer: Boston, MA, USA, 1998; pp. 55–85.
28. Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines*; Apress Open: Berkeley, CA, USA, 2015; pp. 67–80.
29. LaValle, S.M.; Branicky, M.S.; Lindemann, S.R. On the relationship between classical grid search and probabilistic roadmaps. *Int. J. Robot. Res.* **2004**, *23*, 673–692. [CrossRef]
30. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
31. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **2019**, *1168*, 022022. [CrossRef]
32. Kreider, J.F.; Haberl, J.S. Predicting hourly building energy use: The great energy predictor shootout—Overview and discussion of results. In Proceedings of the 1994 American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE) Annual Meeting, Orlando, FL, USA, 25–29 June 1994.
33. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [CrossRef]
34. Renaud, O.; Victoria-Feser, M.-P. A robust coefficient of determination for regression. *J. Stat. Plan. Inference* **2010**, *140*, 1852–1862. [CrossRef]
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
37. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. {TensorFlow}: A system for {Large-Scale} machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016.
38. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. *arXiv* **2013**, arXiv:1309.0238.
39. Edwards, R.E.; New, J.; Parker, L.E. Predicting future hourly residential electrical consumption: A machine learning case study. *Energy Build.* **2012**, *49*, 591–603. [CrossRef]
40. Kelly, J.; Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2015**, *2*, 150007. [CrossRef]
41. Dimopoulos, T.; Bakas, N. Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus. *Remote Sens.* **2019**, *11*, 3047. [CrossRef]
42. Gevrey, M.; Dimopoulos, I.; Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* **2003**, *160*, 249–264. [CrossRef]

43. Olden, J.D.; Jackson, D.A. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* **2002**, *154*, 135–150. [CrossRef]
44. Kaewunruen, S.; Sussman, J.M.; Matsumoto, A. Grand Challenges in Transportation and Transit Systems. *Front. Built Environ.* **2016**, *2*, 4. [CrossRef]