MULTICOLLINEARITY IN MODELS OF APPLIANCE ENERGY USE

ABSTRACT

This paper is concerned with the impact of multicollinearity on regression models of appliance energy use. It uses the 'Appliances Energy Dataset' and the same models as its authors (Candanedo, Feldheim and Deramaix, 2017) - multiple linear regression, support vector machine with radial kernel, random forest and gradient boosting machine. Multicollinearity was removed using manual selection (correlation / mutual information) and principal component analysis. Manual selection proved more effective than PCA, for all models except SVM. Neither method improved accuracy, but manual selection achieved approximately equal results with fewer variables (seven versus 34).

Keywords: energy use, appliances, regression, machine learning, LM, SVM, RF, GBM, PCA.

INTRODUCTION

Appliance use represents a significant proportion of residential energy consumption. Eurostat (2024) found 20.2% was due to appliances and lighting in 2022. DESNZ (2024) quote 22% (18% excluding lighting) in 2023, up from 17% (14%) in 2008 when appliance energy use was first considered separately. Given the increasing availability of smart appliances and improving housing energy efficiency, this trend seems unlikely to reverse. Appliance use prediction aims to identify demand peaks, detect malfunctions/human errors from unusual use patterns, and contribute to building energy models.

This paper uses the 'Appliances Energy Dataset' from UCI (Candanedo, 2017), which is from a Passive House in Belgium (Candanedo, Feldheim and Deramaix (2017), 'the original authors/paper'), with additional climate variables. The house has a network of smart sensors measuring temperature and humidity. Predicting a numerical value requires a regression model, and they chose linear regression (LM), support vector machine (SVM) with radial kernel, random forest (RF) and gradient boosting machine (GBM). This broad approach covers both linear and non-linear variable relationships, and bagging/boosting methods.

Unfortunately, their results were disappointing; even their best model (GBM) scored poorly (R^2 0.57, RMSE 66.65) on the test dataset. Scores on the training set were much better (GBM R^2 0.97, RMSE 17.56), suggesting overfitting. Similar results were obtained by Vakharia *et al.* (2024), Chammas, Makhoul and Demerjian (2019), Adams *et al.* (2019), Xiang, Xie and Xie (2020) and Shorfuzzaman and Hossain (2022).

There is lots of multicollinearity in the dataset, as Passive Houses are designed to maintain consistent conditions inside. There is also high correlation in temperature and humidity between outside and inside, and between temperature and humidity themselves. Multicollinearity does not necessarily affect prediction accuracy (Frost, no date), so may not cause the low scores. LM is most affected (which might explain why this model's scores are so low - R^2 0.16, RMSE 93.18 on the test set), while ensemble methods like RF and GBM are more resistant (Geeks for Geeks, 2024). Opinions on whether SVM is affected seem divided. However, it does obscure variable importance (Daoud, 2017), may cause overfitting (Geeks for Geeks, 2025), and increases computational demand.

Previous studies have not removed the multicollinearity. In this paper, two methods - manual variable selection using correlation and mutual information, and principal component analysis (PCA) - are compared on all four models from the original paper.

DATA EXPLORATION AND FEATURES SELECTION

Dataset

The dataset has 19735 instances of 28 variables (temperature/humidity readings within the home, climate data from a nearby weather station and two random variables), all numeric except for datetime. Measurements were taken at 10-minute intervals over 4.5 months (Jan to May). Appliances energy consumption and light energy consumption are rounded to the nearest 10Wh. The target variable is the total energy use due to appliances only.

Table 1: Original variables

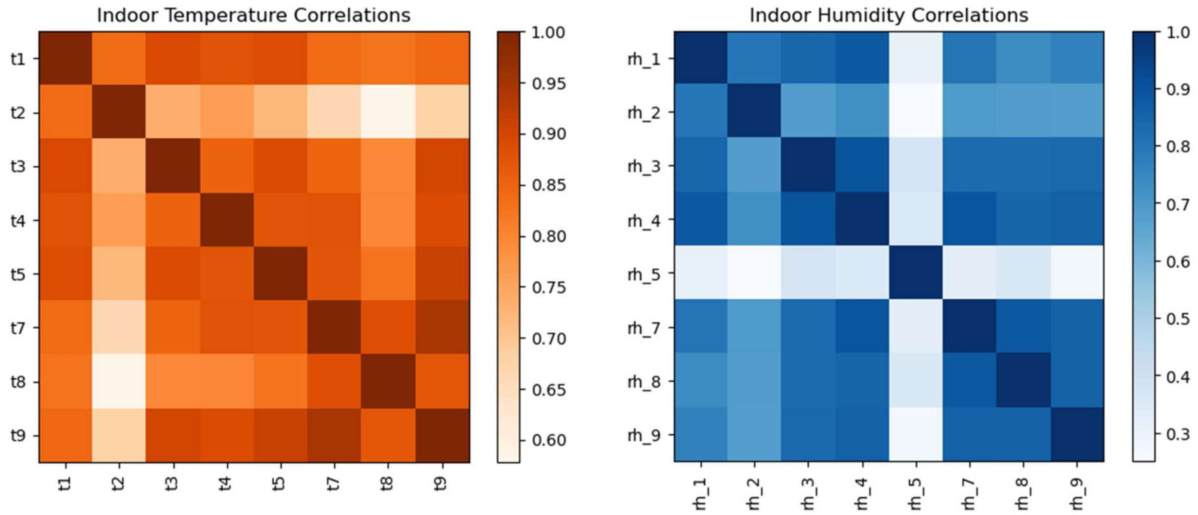
Variable	Units
Date and time	n/a
Appliances energy consumption (target)	Wh
Light energy consumption	Wh
T1, Temperature in kitchen area	◦ C
RH1, Humidity in kitchen area	%
T2, Temperature in living room area	◦ C
RH2, Humidity in living room area	%
T3, Temperature in laundry room area	◦ C
RH3, Humidity in laundry room area	%
T4, Temperature in office room	◦ C
RH4, Humidity in office room	%
T5, Temperature in bathroom	◦ C
RH5, Humidity in bathroom	%
T6, Temperature outside the building (north side)	◦ C
RH6, Humidity outside the building (north side)	%
T7, Temperature in ironing room	◦ C
RH7, Humidity in ironing room	%
T8, Temperature in teenager room	◦ C
RH8, Humidity in teenager room	%
T9, Temperature in parents' room	◦ C
RH9, Humidity in parents' room	%
T_out, Temperature outside (from Chièvres weather station)	◦ C
Pressure (from Chièvres weather station)	mm Hg
RH_out, Humidity outside (from Chièvres weather station)	%
Windspeed (from Chièvres weather station)	m/s
Visibility (from Chièvres weather station)	km
Tdewpoint (from Chièvres weather station)	◦ C
Random Variable 1 (RV 1)	n/a
Random Variable 2 (RV 2)	n/a

There are no missing values in the dataset. Some of the measurements might warrant double-checking with the occupants (if not already done). As this was not possible for this paper, and the values are not outside the possible range, no outliers were removed.

Feature Engineering and Selection

Due to the high correlations, only one measure of temperature or humidity was included. Figure 1 shows the multicollinearity among indoor temperatures/humidities.

Figure 1: Correlations between indoor temperatures and humidities.



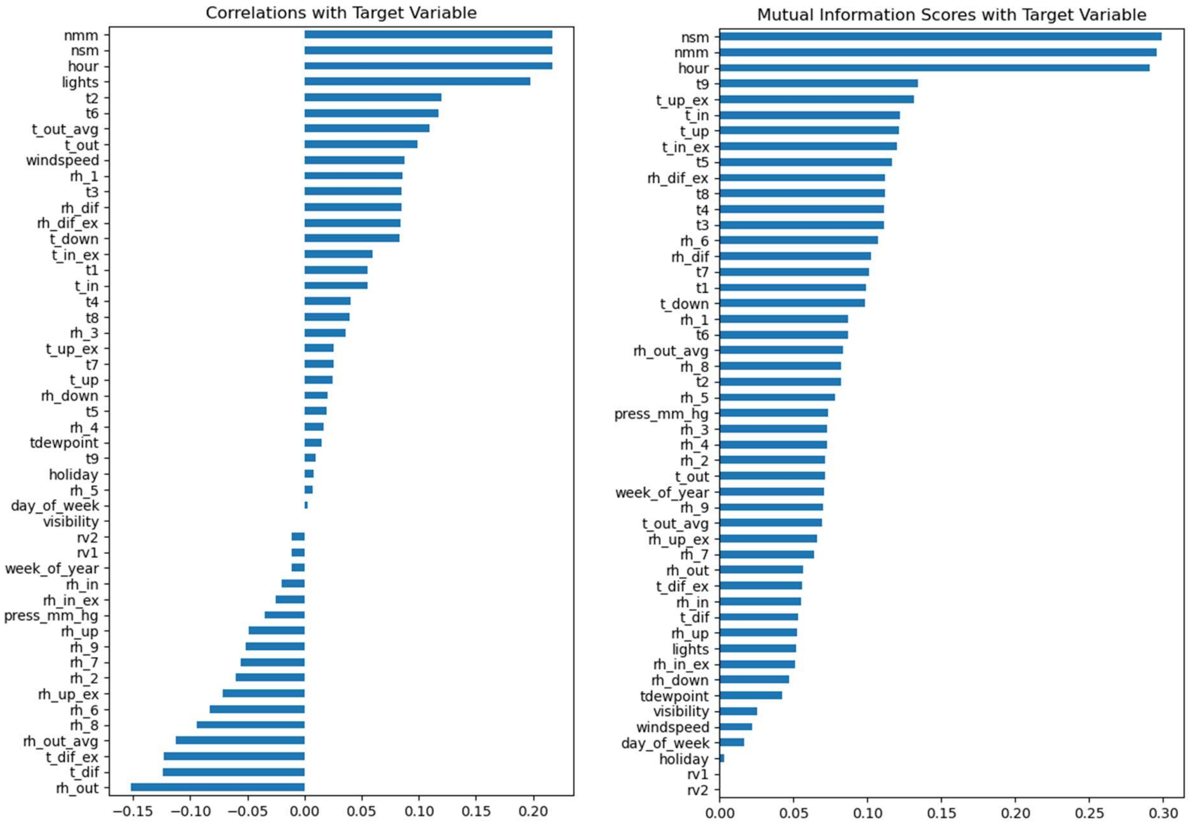
The exceptions here are RH5 (bathroom), where there are large spikes presumably due to the shower, and T2 (living room), probably because people tend to congregate there.

Outside temperature and humidity were measured both at the house and at a nearby weather station, so T6 and T_out measure the same thing, as do RH6, R_out and T_dewpoint (an alternative measure of humidity). This redundancy also needs removing.

In their paper, the original authors computed the number of seconds after midnight (NSM) as a measure of time. This paper also considered the hour and the number of minutes after midnight (NMM), but neither gave a better result. Week of the year was calculated as a measure of seasonal changes, but it correlated highly with other variables. As in the original paper, day of the week was added. Whether the date was a Belgian national holiday was also considered but showed little effect.

Figure 2 shows the correlations and mutual information scores of all the independent variables with the target variable. The final set of features used for the models was: NSM, T9, atmospheric pressure, lights, visibility, day of the week and windspeed.

Figure 2: Correlations and mutual information scores of independent variables with target variable.



Principal Component Analysis (PCA)

As a result of the scree plot (Figure 3), an alternative set of four variables was calculated using PCA.

Figure 3: Scree plot of variance explained by each PCA component.

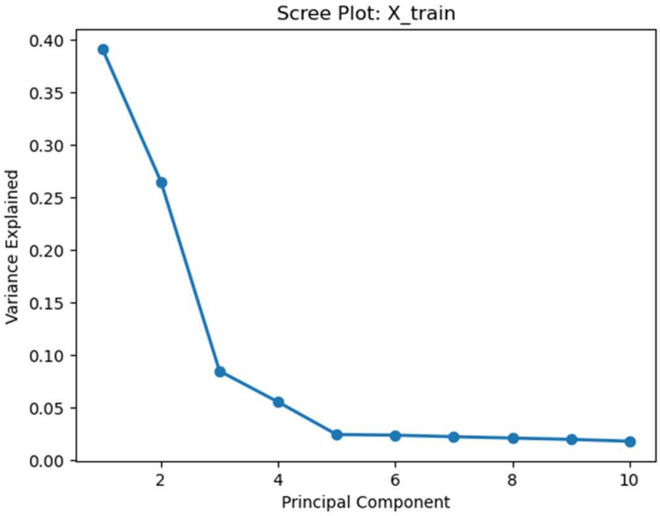
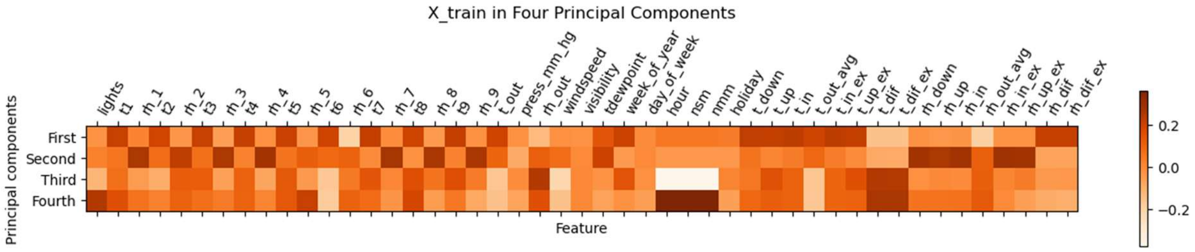


Figure 4: Composition of principal components.



Data Integrity

All the work was done in Python, using Scikit-learn. The train and test sets use an 80:20 split. Where applicable, scaling was performed within each fold of 10-fold cross validation and the test set was scaled separately, to prevent data leakage. All models use the same cross-validation split. The Robust Scaler was used (except for PCA, where Standard Scaler proved more effective), as it preserves variable relationships like the Min-Max Scaler but is less sensitive to outliers (Geeks for Geeks, 2025). Random state 123 was used throughout for reproducibility.

Models Used

Parameter values (Table 2) were chosen using RandomSearchCV and GridSearchCV, done separately for manually-selected and PCA-selected variables. If not otherwise stated, the default value performed best. Parameter tuning for RF and GBM followed Jain (2025).

Table 2: Model Parameters

Model	**Parameters (Manual Selection)**	**Parameters (PCA)**
Multiple linear regression	-	-
Support vector regression	epsilon=5, gamma=15, C=375	epsilon=5, gamma=5, C=650
Random forest regressor	n_estimators = 750, max_depth=33, max_features=3	n_estimators=800, max_depth=39, max_features=2
Gradient boosting regressor	n_estimators = 700, learning_rate = 0.05, max_depth = 30, min_samples_split = 60, max_features = 3,	learning_rate = 0.2, n_estimators = 2700, max_depth = 22, min_samples_split = 17, min_samples_leaf = 3, max_features = 1

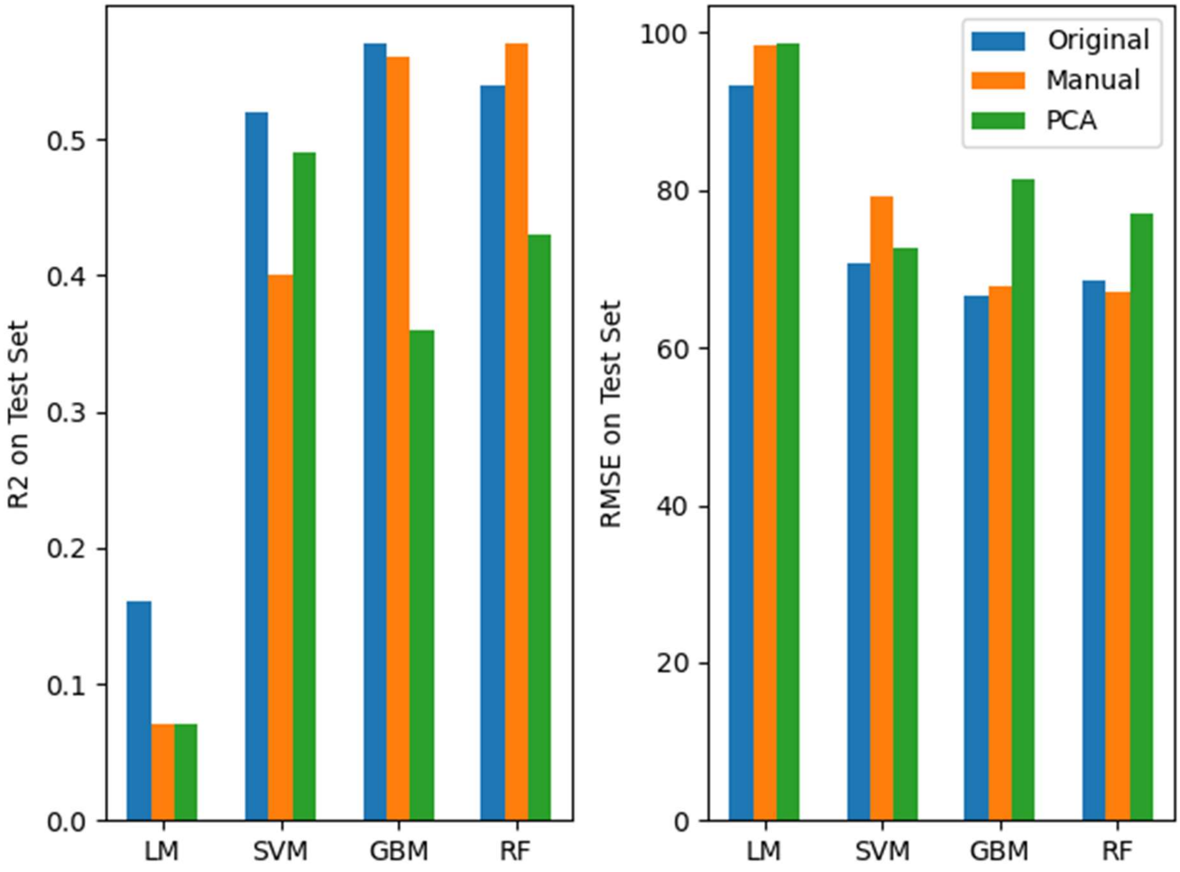
RESULTS

Model predictions were rounded to the nearest 10 Wh, as in the dataset.

Table 3: Model results in the original paper, with manual variable selection and PCA.

	Original Paper				**Manual Selection**				**PCA**			
	Train		**Test**		**Train**		**Test**		**Train**		**Test**	
Model	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
LM	0.18	93.21	0.16	93.18	0.08	98.76	0.07	98.35	0.07	99.16	0.07	98.57
SVM	0.85	39.35	0.52	70.74	0.77	49.22	0.40	79.23	0.88	36.07	0.49	72.68
GBM	0.97	17.56	0.57	66.65	1.00	4.84	0.56	67.91	1.00	0.00	0.36	81.39
RF	0.92	29.61	0.54	68.48	0.94	24.55	0.57	67.01	0.93	28.01	0.43	77.14

Figure 5: Clustered bar chart of model results.



DISCUSSION AND CONCLUSIONS

Results are consistent with previous work. Removing the multicollinearity from the dataset has not significantly increased model accuracy nor reduced overfitting. However, the two best-performing models (GBM and RF) are almost as accurate with seven variables (manual selection) as with 34 (original paper). So, variable selection has decreased model complexity/dimensionality and thus running time with no significant drop in performance. The models are not good enough to know if it has improved estimates of variable importance, as permutation feature importance is model-specific, and features important in a bad model can be unimportant in a good one.

LM performed worse with variable selection, even though it should be most affected by multicollinearity. This could indicate that important variables were excluded (although the other models still appear overfitted). Since the LM scores are so low, the relationships between the independent and dependent variables are probably not linear. This might also explain why PCA did not perform better than manual selection in 3 of 4 models, as it also assumes linear relationships (Keboola, 2022). Non-linear feature selection methods such as t-SNE might do better (Przyborowski, 2024).

The energy consumption data is rounded to the nearest 10Wh rather than continuous, which is not ideal for regression models. Whether this was a sensor limitation or the original authors' choice is not clear. In this paper, the predictions were rounded before evaluation, but it is not known whether the original authors did the same, so error comparisons may not be like-for-like.

Appliance use is closely linked with occupancy. The original authors included lights as a measure of this, but in modern houses lights are often not required during the day. Also, lamp use would fall under appliances (although only two are listed for this house). The original authors saw no drop in performance when excluding lights from their best (GBM) model. However, here permutation feature importance on RF and GBM manual found visibility to be the least important factor.

Alternative ways to measure occupancy could be considered, e.g. video doorbells or door closing sensors, but this would be an invasion of privacy. In conjunction with address details, this information could also be used for criminal activity. Smart meter data is considered personal data and so is subject to the GDPR (European Commission, 2018). Energy use patterns for individual households can also be used to infer a surprising amount of demographic and socio-economic information (Teng *et al.* 2022), especially in conjunction with other data sources like social media use. This could be exploited for targeted advertising or discriminatory pricing.

The available variables may not be sufficient for accurate predictions. Other possible relevant factors are precipitation, rolling consumption averages (Basu *et al.*, 2013, cited in Candanedo, Feldheim and Deramaix, 2017), school/family holidays and power cuts (Adams *et al.* 2019). However, long short-term memory network (LSTM) models have achieved much better results: Xiang, Xie and Xie (2020), Syed *et al.* (2021) and Shorfuzzaman and Hossain (2022) all achieved $R^2 > 95\%$ and RMSE <40 on the test set. Further research on predicting appliance energy is likely to focus on these rather than traditional machine learning models, although LSTMs require more computing power and training time and lack interpretability (Srivatsavaya, 2023).

REFERENCES

Adams, S., Greenspan, S., Velez-Rojas, M., Mankovski, S. and Beling, P.A. (2019) 'Data-driven simulation for energy consumption estimation in a smart home.' *Environment Systems and Decisions*, 39(3), pp. 281-294. doi: https://doi.org/10.1007/s10669-019-09727-1

Candanedo (2017) *Appliances Energy Prediction*. Available at: https://archive.ics.uci.edu/dataset/374/ (Accessed: 14 Jan 2025).

Candanedo, L.M., Feldheim, V. and Deramaix, D. (2017) 'Data driven prediction models of energy use of appliances in a low-energy house', *Energy and Buildings*, 140, pp. 81-97. doi: https://doi.org/10.1016/j.enbuild.2017.01.083

Chammas, M., Makhoul, A., and Demerjian, J. (2019) 'An efficient data model for energy prediction using wireless sensors', *Computers and Electrical Engineering*, 76, pp.249-257. doi: https://doi.org/10.1016/j.compeleceng.2019.04.002

Daoud, J.I., (2017) 'Multicollinearity and Regression Analysis', *Journal of Physics: Conference Series*, 949, Article ID: 012009, doi: 10.1088/1742-6596/949/1/012009

Department for Energy Security and Net Zero (DESNZ) (2024) *Energy consumption in the UK 2024: End uses data tables*. Available at: https://www.gov.uk/government/statistics/energy-consumption-in-the-uk-2024 (Accessed: 21 Mar 2025).

European Commission: Smart Grid Task Force 2012-2014 (2018) *Data Protection Impact Assessment Template for Smart Grid and Smart Metering Systems*. Available at: https://energy.ec.europa.eu/topics/markets-and-consumers/smart-grids-and-meters/data-protection-impact-assessment-smart-grid-and-smart-metering-environment_en (Accessed: 29 April 2025).

Eurostat (2024) *Energy consumption in households*. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households (Accessed: 21 Mar 2025).

Frost, J. (no date) *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Available at: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/ (Accessed: 28 Mar 2025).

Geeks for Geeks (2024) 'Solving the Multicollinearity Problem with Decision Tree.' Available at: https://www.geeksforgeeks.org/solving-the-multicollinearity-problem-with-decision-tree/ (Accessed: 28 Mar 2025).

Geeks for Geeks (2025) 'StandardScaler, MinMaxScaler and RobustScaler techniques – ML.' Available at: https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/ (Accessed: 10 Mar 2025).

Jain, A. (2025) 'Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python'. Available at: https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/ (Accessed: 4 Apr 2025)

Keboola (2022) 'A Guide to Principal Component Analysis (PCA) for Machine Learning' Available at: https://www.keboola.com/blog/pca-machine-learning (Accessed: 25 Apr 2025).

Przyborowski, M. (2024) 'Dimensionality Reduction - Popular Techniques and How to Use Them.' Available at: https://nexocode.com/blog/posts/dimensionality-reduction-techniques-guide/ (Accessed: 24 Apr 2025).

Robbia, G, R. (2022) 'Energy prediction of appliances using supervised ML algorithms', *Innovative Computing Review*, 2(1), pp. 73-89. doi: https://doi.org/10.32350/icr.0201.05

Shorfuzzaman, M. and Hossain, M.S. (2022) ‘Predictive Analytics of Energy Usage by IoT-Based Smart Home Appliances for Green Urban Development’, *ACM transactions on Internet technology*, 22(2), pp.1-26. doi: https://doi.org/10.1145/3426970

Srivatsavaya, P. (2023) ‘LSTM - Implementation, Advantages and Disadvantages’. Available at: https://medium.com/@prudhviraju.srivatsavaya/lstm-implementation-advantages-and-diadvantages-914a96fa0acb (Accessed: 1 May 2025)

Syed, D., Abu-Rub, H., Ghrayeb, A and Refaat, S.S. (2021) ‘Household-Level Energy Forecasting in Smart Buildings Using a Novel Hybrid Deep Learning Model’, *IEEE Access*, 9, pp. 33498-33511. doi: https://doi.org/10.1109/ACCESS.2021.3061370

Teng, F., Chhachhi, S., Ge, P., Graham, J., and Gunduz, D. (2022) *Balancing Privacy and Access to Smart Meter Data: An Energy Futures Lab Briefing Paper.* London: Imperial College. Available at: https://doi.org/10.25561/96974 (Accessed: 1 May 2025).

Vakharia, V., Dave, V., Borade, H., Agrawal, H. and Padia, N. (2024) ‘An efficient approach for appliances energy prediction using differential evolution optimization and Light GBM machine learning model’, *2024 International Conference on Sustainable Energy: Energy Transition and Net-Zero Climate Future (ICUE)*, Pattaya City, Thailand, 21-23 Oct 2024. IEEE. doi: https://doi.org/10.1109/ICUE63019.2024.10795607

Xiang, L., Xie, T. and Xie, W. (2020) ‘Prediction model of household appliance energy consumption based on machine learning’. *Journal of Physics: Conference Series*, 1453, Article ID: 012064. doi: https://doi.org/10.1088/1742-6596/1453/1/012064