subset of a residential energy forecasting model.

The presented study extends this line of research by evaluating a more comprehensive dataset. The dataset includes demographic information, housing unit characteristics (e.g., age, size, number of rooms), region, and energy consumption (that is broken down into different pre-defined energy consumption categories) to determine the most discriminative features for building energy performance prediction to be used in decision-making in energy retrofits.

METHODOLOGY

The proposed methodology for this study is shown in Figure 1. The major steps include data selection and pre-processing, feature ranking and selection, and the prediction model development using a machine learning approach. The methods used in each step are described in details in this section.

Figure 1. The proposed methodology

Data Collection and Pre-Processing

The U.S. Energy Information Administration (EIA) dataset of 2015 Residential Energy Consumption Survey (RECS) is used for this study. RECS is a nationally representative sample of housing units that include their energy consumption information. The raw data included more than 5,600 data points with more than 700 variables. A pre-processing technique is applied to the raw noisy data to identify and remove outliers and resolve inconsistencies. The purpose of data pre-processing is to transform the dataset so that the information content is best exposed to the data-mining tool. The following methods are used for pre-processing of the data:

Data Selection

A selection process is used to remove data points having missing information about electricity and gas consumption, along with demographic information, housing unit characteristics (e.g., age, size, number of rooms), and region, among others. Besides, since about 60% of the data represented the single-family detached houses, only this type of housing units are considered for data analysis in this study. Such information richness at the micro-level is appropriate for an empirical analysis of the energy consumption prediction at the household level.

Outlier Detection and Removal

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Outliers detection is a task that finds

objects that are dissimilar or inconsistent regarding the remaining data. Different methods such as box plot, scatter plot, interquartile range (IQR), or percentage-based methods can be used. Due to the nature of the data, a 5-95% method is used to detect and remove outliers, in which 5% of data from the minimum side and 5% data from the maximum side are detected and removed from the dataset.

The resulting dataset includes 2,084 data points each representing a detached single-family housing unit that only uses electricity or natural gas as the main energy sources. There are 225 attributes as independent variables for each house, including housing unit characteristics, location characteristics, and energy consumption information. The ultimate dependent variable to be predicted in this study is the total energy consumption of units (i.e. the sum of electricity and natural gas consumption) that can be calculated as the total Btu of consumed energy per year (TOTALBTU).

Feature Ranking and Selection

Feature selection is a key step in developing shallow machine learning models as it helps the classifier to be fast, computationally effective, and more accurate (Karabulut et al. 2012). For this study, a two-step approach is used to select the most important and relevant independent variables. In the first step, a correlation matrix between the dependent and independent variables is created. The variables with the p-value less than 0.01 are kept for the next step and the rest are removed as they do not contribute to statistical significance. In the second step, the list of correlated variables is reviewed and the most relevant ones are selected. For example, for lighting energy consumption category, if the variable "*Number of inside light bulbs turned on at least 4 hours a day"* exists in the initial list of correlated variables, it is kept in that category. However, another correlated variable such as "*If the electricity is used for space heating"* is removed from the final list of related variables since it was not relevant to the category. This will ensure there is no multicollinearity in the data.

Prediction Model Development

Different machine learning models are used in this study to find the best energy consumption predictor including. Regression, Decision Trees, and Neural Networks are used as commonly used algorithms and Bootstrap Forest and Boosted Trees are also employed as ensemble models. For the analysis a statistical software called JMP by SAS Institute Inc. is used to train and validate the models (JMP 2019). JMP software is partly focused on exploratory data analysis and visualization. It is designed for users to investigate data to learn the unexpected as opposed to confirming a hypothesis.

To develop the model, the data is divided to 60% training, 20% validation, and 20% testing. While training and building the models, the required coefficients (based on the selected model) are learned and fitted to training data. The training aim is to find the best fit model such that cost function is minimized. The cost function helps in measuring the error. During the training process, the error between actual and predicted values as well as the cost-function are minimized. A simple and common cost function is Mean Squared Error (MSE) which is equal to the average squared difference between an observation's actual and predicted values. While validating the models, the dataset is used to minimize overfitting to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. Finally, while testing the model, the final model fit on the training dataset is evaluated and compared in different models. Finally, the R-Square is calculated based on the predicted and actual total energy consumption for each

model, and the machine learning model with the highest R-Square is selected as the best-fit approach. It is worth mentioning that other methods such as k-fold training and validation to avoid under- or overfitting can be also used. However, since a variety of models have been compared here in terms of training and validation R-Squares, it was guaranteed that there is no under- or overfitting involved.

(a) Analyzed Data **(b) EPA U.S. Average Home (2009)**

Figure 2. Energy consumption percentages

MODEL DEVELOPMENT

Data Summary

For the purpose of this study, the ultimate dependent variable is broken down into six main energy consumption categories, including *Space Heating*, *Air Conditioning*, *Water Heating*, *Lighting*, *Appliances and Electronics*, and *Others*. The *Others* category consists of any other variable that cannot be categorized in the named categories such as *Hot Tube Heater* and *Pumping Equipment*. Since the category of *Appliances and Electronics* includes a wide range of items, it is also broken down into four sub-categories, including *fridge*, *cooking*, *laundry*, and *others*. Figure 2 presents the average value of these energy consumption categories in the analyzed dataset. It also compares the results with the percentages of energy usage in a U.S. average home by the end-users (EPA, 2009). The average energy consumption of a housing unit in the dataset is 93.6 MBtu per year; while on average about 50% of that is for space heating based on both the analyzed data and EPA (2009).

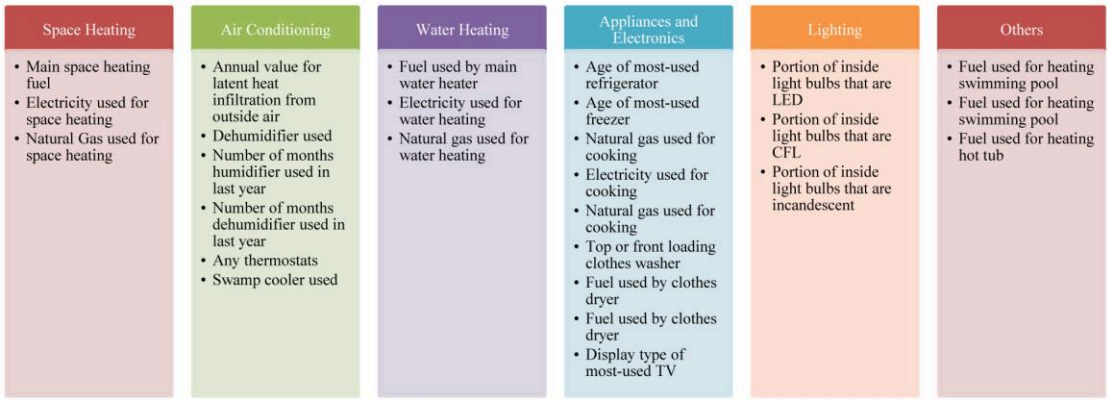

Figure 3. List of variables that can be changed through building energy retrofits Prediction Model Development

Table 1. Selected Variables for Each Energy Consumption Category

Category / Sub-Category	Variable	Description
Space Heating	FUELHEAT	Main space heating fuel
	ELWARM	Electricity used for space heating
	EGWARM	Natural Gas used for space heating
	DIVISION	Census Division
	UATYP10	Census 2010 Urban Type
	IECC_CLIMATE_PUB	IECC Climate Code
	CDD65	Cooling degree days in 2015, base temperature 65F
	HDD50	Heating degree days in 2015, base temperature 50F
	HDD65	Heating degree days in 2015, base temperature 65F
	DBT1	Dry bulb temperature expected to be exceeded 1% of times
	DBT99	Dry bulb temperature expected to be exceeded 99% of times
	BASEHEAT 2	Heating used in basement
	YEARMADERANGE	Range when housing unit was built
	TEMPHOME	Winter temperature when someone is home during the day
	WINDOWS	Number of windows
	TOTCSQFT	Total cooled square footage
Air Conditioning	CDD65	Cooling degree days in 2015, base temperature 65F
	OA_LAT	Annual value for latent heat infiltration from the outside air
	HDD50	Heating degree days in 2015, base temperature 50F
	TOTCSQFT	Total cooled square footage
	NOTMOIST	Dehumidifier used
	USEMOISTURE	Number of months humidifier used in last year
	USENOTMOIST	Number of months dehumidifier used in last year
	WINDOWS	Number of windows
	YEARMADERANGE	Range when housing unit was built
	NUMCFAN	Number of ceiling fans used
	IECC_CLIMATE_PUB	IECC Climate Code
	NUMBERAC	Number of individual air conditioning units used
	CENACHP 2	Central air conditioner is a heat pump

Table 1. Selected Variables for Each Energy Consumption Category (Continued)

Category / Sub-Category	Variable	Description
	THERMAIN	Any thermostats
	CLIMATE_REGION_PUB	Building America Climate Category
	SWAMPCOL	Swamp cooler used
	NHSLDMEM	Number of household members
	IECC_CLIMATE_PUB	IECC Climate Code
Water Heating	FUELH2O	Fuel used by main water heater
	ELWATER	Electricity used for water heating
	UGWATER	Natural gas used for water heating
	YEARMADERANGE	Range when housing unit was built
	ICE	Through-the-door ice on most-used refrigerator
	NUMFRIG	Number of refrigerators used
	AGERFRI1	Age of most-used refrigerator
Fridge	TYPERFR2	Door arrangement of the second most-used refrigerator
	NUMFREEZ	Number of separate freezers used
	UPRTFRZR	Door arrangement of the most-used freezer
	AGEFRZR	Age of most-used freezer
	SIZFREEZ	Size of most-used freezer
	COOKTUSE	Frequency of use of cooktop part of the stove
	OVENUSE	Frequency of use of oven part of the stove
	SEPOVENUSE	Frequency of separate oven use
	UGCOOK	Natural gas used for cooking
Cooking	SEPCOOKTUSE	Frequency of separate cooktop use
	ELFOOD	Electricity used for cooking
	OVEN	Number of separate ovens
	UGCOOK	Natural gas used for cooking
	COOKTUSE	Frequency of use of cooktop part of the stove
	AMTMICRO	Frequency of microwave use
	WASHLOAD	Frequency of clothes washer use
	TOPFRONT	Top or front-loading clothes washer
	CWASHER	Have clothes washer in home
	DRYRUSE	Frequency of clothes dryer use

Appliances and Electronics

Appliances and Electronics — Laundry

Table 1. Selected Variables for Each Energy Consumption Category (Continued)

Category / Sub-Category	Variable	Description
Others	DRYRFUEL	Fuel used by clothes dryer
	DRYER	Have clothes dryer in home
	DRYRFUEL	Fuel used by clothes dryer
	DRYRUSE	Frequency of clothes dryer use
	DRYER	Have clothes dryer in home
	DWASHUSE	Frequency of dishwasher use
	COMBODVR	Number of cable or satellite boxes with DVR
	SEPDVR	Number of separate DVRs
	CABLESAT	Number of cable or satellite boxes without DVR
	TVONWD1	Most-used TV usage on weekdays
	TVSIZE1	Size of most-used TV
	TVTYPE1	Display type of most-used TV
Lighting	LGTIN4	Number of inside light bulbs turned on at least 4 hours a day
	LGTINLED	Portion of inside light bulbs that are LED
	LGTINCFL	Portion of inside light bulbs that are CFL
	LGTINCAN	Portion of inside light bulbs that are incandescent
	LGTOUTNUM	Number of light bulbs installed outside the home
	TOTUSQFT	Total unheated square footage
Others	MONPOOL 2	Months swimming pool used in the last year
	POOL 2	Heated swimming pool
	RECBATH	Hot tub
	FUELPOOL 2	Fuel used for heating swimming pool
	FUELPOOL 2	Fuel used for heating swimming pool
	SWIMPOOL	Swimming pool
	MONTUB 2	Months hot tub used in the last year
	FUELTUB 2	Fuel used for heating hot tub

Variable Selection

In the next step, the most relevant independent variables are selected, considering each energy consumption category as a dependent variable. These variables are shown in Table 1. The selected variables are not only statistically correlated to the energy consumption of each energy category but also they are filtered by an expert's insight to make sure they are logically correlated as well.

As Table 1 illustrates, there are some variables that their value can be changed through building energy retrofits. Therefore, the model could help understand and predict the impact of implementing such EEM on the building. The list of variables and impacted energy categories to be investigated for building energy retrofit are shown in Figure 3.

As mentioned before, four different machine learning models are used in this study to find the best energy consumption prediction model. Only the selected variables from the previous section (among all 225 attributes) are used to develop the energy prediction models. The accuracy and R-square of each developed model are shown in Table 2. The ANN model had the best performance to be used for building energy consumption prediction (R-Square of 67%). It is worth mentioning that the low number of data points usually makes ANN not the best predictor. But since having no under- and over-fitting was ensured through validation R-Square, it was decided to proceed with ANN. A one hidden layer ANN model is used with ten nodes chosen for Tanh, Linear, and Gaussian. A Learning rate of 0.1 is used as well as transform covariates and robust fit for fitting options. Besides, Weight Decay is used as the penalty method.

Table 2. Comparison of Different ML Models

Dependent Variable	**Models – R-Sqaure Values (Validation)**				
	Regression	Decision Trees	BootStrap Forest	Boosted Trees	**Neural Network**
Space Heating	0.581	0.606	0.639	0.630	**0.739**
Air Conditioning	0.616	0.509	0.595	0.611	**0.669**
Water Heating	0.662	0.750	0.460	0.750	**0.844**
Appliances	0.874	0.559	0.754	0.760	**0.876**
Lighting	0.683	0.621	0.456	0.603	**0.825**
Others	0.183	0.231	0.197	0.219	**0.369**
Total	***0.599***	***0.501***	***0.402***	***0.650***	***0.670***

MODEL IMPLEMENTATION

A simple example case is presented in this section to illustrate how the developed model can help predict building energy performance through energy retrofits. A sample average housing unit is determined based on the data, using mean values for numerical/continuous variables, the median value for numerical/discrete and categorical/ordinal variables, and mode value for categorical/nominal variables. The sample house is a unit with 2,513 square feet, three bedrooms, two bathrooms, located on a cold/very cold climate, occupied with three households. According to the developed model, the total energy consumption of the sample house is predicted to be 125.9 MBtu per year, which is equal to 36,899 kWh if only electricity is used in that housing unit (12,300 kWh per person per year or 14.7 kWh per square feet per year). This

predicted value is in line with the EIA average annual electricity consumption of 10,400 kWh per person per year for a U.S. residential utility customer. This example case is used to investigate the impact of some EEM examples on building energy consumption. The EEM selected to be investigated in this section are (1) installing thermostat; (2) replacing a higher portion of inside light bulbs with LED; and (3) replacing the appliances with a newer (and more energy efficient) ones. These EEM impacts the energy consumption on air conditioning, lighting, and appliances and electronics, respectively.

Table 3. The Impact of EEM on Average Building Energy Consumption

EEM	Variable	Value	Description	Impacted Energy Category	Predicted Energy Category Consumption (MBtu per year)	Predicted Energy saving (kWh per year)
1	THERMAIN	1	There is a thermostat	Air Conditioning	9.0	101.6
		0	There is no thermostat		9.3	-
2	LGTINLED	4	All of the inside light bulbs that are LED	Lighting	2.3	444.4
		3	Most of the inside light bulbs that are LED		2.6	372.0
		2	About half of the inside light bulbs that are LED		2.7	326.4
		1	Some of the inside light bulbs that are LED		3.3	164.1
		0	None of the inside light bulbs that are LED		3.9	-
3	AGERFRI1	1	Age of most-used refrigerator is less than 2 years old	Appliances / Fridge	11.6	30.1
		2	Age of most-used refrigerator is 2 to 4 years old		11.7	69.1
		3	Age of most-used refrigerator is 5 to 9 years old		11.7	72.5
		4.1	Age of most-used refrigerator is 10 to 14 years old		11.8	48.2
		4.2	Age of most-used refrigerator is 15 to 19 years old		11.8	44.6
		5	Age of most-used refrigerator is 20 years or older		12.0	-

The results of model implementation are shown in Table 3. As the results show, the annual energy saving for implementing each EEM can be calculated using the developed model. For example, installing a thermostat can reduce the required energy for air conditioning for around 101.6 kWh per year. Assuming the average electricity rate of 13.20 cents per kWh in the U.S., installing a thermostat could result in saving $13.4 per year for air conditioning for the sample house. On the other hand, installing more LED lights could increase the saving from $21.6 to $58.7 per year, based on the portion of light bulbs that are LED. Finally installing a newer refrigerator could increase the energy savings up to $4.0 per year. Such information could help decision-makers to predict the impact of EEM on building energy consumption and make more effective decisions for energy retrofits.

CONCLUSION AND FUTURE WORK

The proposed framework aims at determining the best set of features for a machine learning-based building energy consumption prediction as well as developing such model. This study addresses areas that need further research attention as identified by Amasyali and El-Gohary (2018) in their comprehensive review of data-driven building energy consumption prediction studies by targeting big energy data analytics and the residential sector. The performance of the developed framework is tested on 2015 Residential Energy Consumption Survey (RECS) data through the use of a two-step feature selection method as well as different machine learning algorithms. Results indicate that the total energy consumption can be predicted with more than 81% accuracy and that the output can be used to implement EEM for energy retrofit. One limitation of the presented paper is that the model is not validated through real case studies which is being addressed in an ongoing study by the research team.

REFERENCES

Ahmad, M.W., Mourshed, M., and Rezgui, Y. (2017) “Trees vs. Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption.” *Energy and Building*, 147, 77–89.

Amasyali, K., and El-Gohary, N. (2018) “A review of data-driven building energy consumption prediction studies.” *Renewable and Sustainable Energy Reviews,* 81, 192–205.

Ekici, B.B., and Aksoy, U.T. (2011) “Prediction of building energy needs in early stage of design by using ANFIS.” *Expert Systems with Applications*, 38(5), 352–5358.

EPA (Environmental Protection Agency), (2009) “Building and Their Impact on the Environment: a Statistical Summary”

Hsu, D. (2015) “Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data.” *Applied Energy*, 160, 153–163.

Kolter, J., and Ferreira, J., (2011) “A large-scale study on predicting and contextualizing building energy usage.” *Proceedings of the 25th Conference on Artificial Intelligence*, 1349–56.

JMP®, Version <14>. SAS Institute Inc., Cary, NC, <https://www.jmp.com> (Jul. 5, 2019).

Jafari, A. and Valentin, V. (2017) “An optimization framework for building energy retrofits decision-making.” *Building and Environment*, 115, 118-129

Jafari, A. and Valentin, V. (2018) “Selection of optimization objectives for decision-making in building energy retrofits.” *Building and Environment,* 130, 94–103.

Jain, R.K., Damoulas, T., and Kontokosta, C.E. (2014) “Towards data-driven energy consumption forecasting of multi-family residential buildings: feature selection via the

lasso." *In Computing in Civil and Building Engineering*, 1675-1682.
Karabulut, E.M., Özel, S.A., İbrikçi, T. (2012) "A comparative study on the effect of feature selection on classification accuracy." *Procedia Technology*, 1, 323-327.
Mena, R., Rodríguez, F., Castilla, M., and Arahal, M.R. (2014) "A prediction model based on neural networks for the energy consumption of a bioclimatic building." *Energy and Building*, 82, 142–155.
Naji, S., Shamshirband, S., Basser, H., Keivani, A., Alengaram, U.J., Jumaat, M.Z., and Petkovic, D. (2016a) "Application of adaptive neuro-fuzzy methodology for estimating building energy consumption." *Renewable and Sustainable Energy Reviews,* 53, 1520–1528.
Naji, S., Keivani, A., Shamshirband, S., Alengaram, U.J., Jumaat, M.Z., Mansor, Z., Lee, M. (2016b) "Estimating building energy consumption using extreme learning machine method." Energy, 97, 506–516.
Štreimikien˙e, S. (2014) "Residential energy consumption trends, main drivers and policies in Lithuania." *Renewable and Sustainable Energy Reviews*, 35, 285–293.
Xuemei, L., Lixing, D., Jinhu, L., Gang, X., and Jibin, L. (2010a) "A Novel Hybrid Approach of KPCA and SVM for Building Cooling Load Prediction." *Third International Conference on Knowledge Discovery and Data Mining (IEEE)*.
Xuemei, L., Yuyan, D., Lixing, D., Liangzhong, J. (2010b) "Building cooling load forecasting using fuzzy support vector machine and fuzzy C-mean clustering." *International Conference on Computer and Communication Technologies in Agriculture Engineering (IEEE).*
Zhang, Y., and Chen, Q. (2014) "Prediction of building energy consumption based on PSO-RBF neural network." *In Proceedings of the IEEE International Conference on System Science and Engineering*, Shanghai, China, 11–13 July, 60–63.
Zhao, H., and Magoulès, F., (2012) "Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method." *Journal of Algorithms & Computational Technology*, 6, 59–78.
Zuo, J. and Zhao, Z. (2014) "Green building research–current status and future agenda: A review." *Renewable and Sustainable Energy Reviews*, 30, 271–281.