ICIC Express Letters
Part B: Applications
Volume **15**, Number **4**, April **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **399–407**

INTERPRETATION OF THE DOMINANT FEATURES IN THE PREDICTION RESULTS OF ELECTRICAL ENERGY CONSUMPTION USING SMART HOME DATASETS BASED ON KNN MACHINE LEARNING

Mochammad Haldi Widianto[1,*], Alexander Agung Santoso Gunawan[2]
Yaya Heryadi[1] and Widodo Budiharto[2]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan Palmerah, Jakarta 11480, Indonesia
{ aagung; yayaheryadi; wbudiharto }@binus.edu
*Corresponding author: mochamad.widianto@binus.ac.id

Received August 2023; accepted November 2023

Abstract. *The Internet of Things (IoT) is essential for generating data in Smart Homes, and Machine Learning (ML) can help save household electricity energy from prediction results. However, IoT generates many features in Smart Homes that must be considered for prediction and interpretation. Therefore, this study proposes a model based on an ML prediction algorithm and model interpretation with many datasets and features of Smart Homes. The author's proposed model uses the K-Nearest Neighbors (KNN) learning algorithm and the SHapley Additive exPlanations (SHAP) model interpretation to explain the prediction results. The results describe that the proposed model produces optimal predictions compared to other distances. The results of the proposed model represent the effects of Mean Absolute Error (MAE) = 0.0533, Mean Squared Error (MSE) = 0.0893, and Root Mean Squared Error (RMSE) = 0.268. Optimal prediction results occur because the proposed model utilizes the Manhattan distance to better deal with many Smart Home features and datasets. Furthermore, the proposed model can interpret the prediction results with the dominant feature being "Furnace" in Smart Home Datasets. In the future, the proposed model can provide prediction and interpretation results for saving electricity in Smart Homes.*

Keywords: Dominant features, K-Nearest Neighbors (KNN), Smart Home, SHapley Additive exPlanations (SHAP).

1. **Introduction.** The energy crisis has occurred almost worldwide, significantly impacting energy consumption for all sectors, such as business, health, manufacturing, agriculture, education, and other service sectors [1,2]. The energy consumption problem is also a consideration and a challenge in saving or switching to a different clean energy [3]. One of the sectors with the highest energy consumption is the household sector. This sector does many activities to increase the economy at home [4].

Therefore, saving electricity consumption in the household sector is essential. This can be applied by using the Internet of Things (IoT) management information system. This management is used for monitoring, automating, and switching electrical energy consumption [5]. This management is usually combined with Machine Learning (ML) in Smart Homes for flexible and intelligent responses to user requirements [6]. The ML model can produce high-accuracy energy consumption predictions depending on the data conditions [7]. Another trend is knowing how ML works and seeing the interpretation

DOI: 10.24507/icicelb.15.04.399

of dominant features in the prediction results. Therefore, users can understand why predictions describe high or low results based on the explanation from the ML used [8]. Interpretation of this prediction result can be assisted by model interpretation such as Local Interpretable Model-Agnostic Explanations (LIME) [8] and SHapley Additive exPlanations (SHAP) [9]. However, Smart Home has many multi-features and datasets. The problem of prediction and interpretation of a Smart Home requires solving by the author.

Previous research used ML as a prediction of electricity consumption. For example, Shao et al. [10] and Tabrizchi et al. [11] used a Support Vector Machine (SVM) learning algorithm to detect electricity consumption. Other research, such as Afuosi and Zoghi [12], Li and Jin [13], used the K-Nearest Neighbors (KNN) learning algorithm in predicting electricity savings. Several studies have made comparisons of all ML methods, including Priyadarshini et al. [14], Fard and Hosseini [15], and Abdul Malek et al. [16].

However, previous studies usually only compare the best prediction with several other ML methods, and selection based on k (neighbors) on the KNN learning algorithm is one of the things to look for optimal prediction. Furthermore, there are quite a lot of hyperparameters in the KNN learning algorithm [17] that have yet to be exported [18]. In addition, previous studies have yet to consider the large number of multi-features and data sets on Smart Home.

Therefore, the study focuses on taking advantage of these limitations, where previous research on ML prediction has yet to be widely studied to interpret the dominant features and find optimal predictions. Interpretation of prediction results is usually used by humans to better understand the effect of prediction results on existing features. This explanation is possible with SHAP model interpretation. Another advantage is that many previous studies have focused on the "Euclidean" distance, which is very susceptible to multi-feature. The proposed model will find an algorithm that is more robust than the Euclidean by utilizing tuning on the number of k, distance [19], and model interpretation using SHAP as a global model interpretation.

The goal is to minimize model error and quickly set up "interpretation dominant feature" factors to help save electrical energy. The following are the key contributions.

1) This study builds on the proposed model KNN techniques at the prediction and SHAP model interpretation to describe how KNN works and the interpretation of dominant features.

2) The proposed model KNN learning architecture is used for predicting electricity consumption and comparing its findings with other existing hyperparameter KNN models, especially regarding multi-feature conditions.

3) Several main scenarios are assessed against various results from the KNN hyperparameter to provide the best results for predictions, and the best results will be interpreted using SHAP and compared to another model interpretation.

4) The work obtained promising results in predicting the KNN model with reduced error, and interpretation results that Smart Home users can use to save electricity consumption.

The article is divided into the following four sections. Section 1 describes the problem description and introduction. Section 2 describes the proposed model. Section 3 describes the implementation and trial results. Finally, Section 4 is followed by conclusions and suggestions for further research.

2. **Proposed Method.**

This section will explain the stages of using the proposed model. Authors have widely used this model using predictions based on the KNN applied to Smart Homes to predict electricity consumption using IoT. Figure 1 describes the nine-step research flow process for predicting and interpreting electrical energy consumption in Smart Homes.

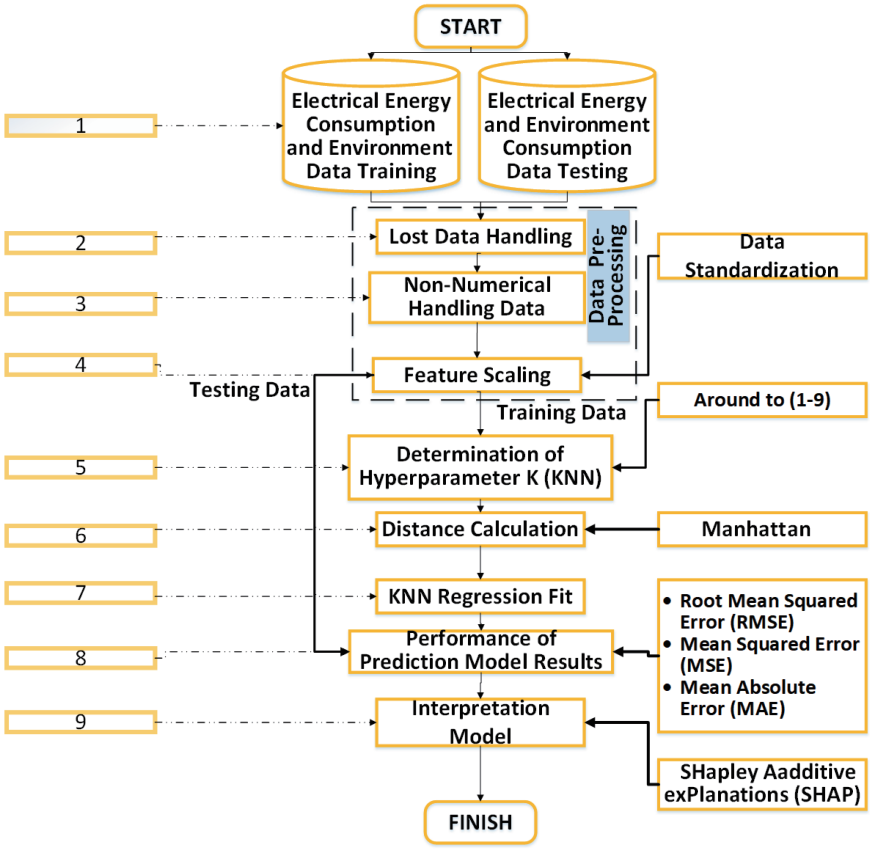


FIGURE 1. Research proposed model

Steps such as "Start" and "Finish" are not considered in order from 1 to 9, and finally, the end of the thread. Figure 1 is explained in more depth as follows.

1) The first stage is the data introduction stage. Information on electricity consumption is collected from dataset files taken from [20] Kaggle secondary data. The data set contains 500000+ data with 32 features. At this stage, the data is also divided by 2 with a comparison between training data and testing data, which is 90/10 for training data.

2) The flow begins with carrying out load handling, where the data will be processed and searched for some features that do not affect the prediction results following previous correlation studies in [21]. The development of 24 features will be processed to the next step.

3) In the next stage, the authors reconfirm whether the data is numeric. If there is data other than numeric, then an encoder will label it as non-numeric data.

4) In this stage, the data will be adjusted to a normal distribution because datasets are generally not distributed. Then, at this stage, the standardization of the Z-Score is used with a range of results [−3 to 3].

5) The first hyperparameter setting is carried out at this stage: the number of neighbors (k) from 1 to 9. This aims to determine how many k matches are in this data set.

6) The second hyperparameter is set using several distance algorithms using Manhattan because this hyperparameter is very influential in retrieving training data like data testing.

7) After obtaining several selected hyperparameters, this section learned all training data using the KNN learning algorithms. The following result is compared with the error.

8) The results in this section evaluate several measurement metrics, such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

9) Finally, the model's results will be interpreted to explain how the KNN algorithm learning can work by providing dominant features using SHAP model interpretation.

2.1. **Data information.** This section uses Smart Home data generated by IoT devices. The data have multi-feature as described in Table 1.

TABLE 1. Data information

No	Feature	Description data
1	Use	All data on total electricity consumption on Smart Home
2	Solar Cell	Electric energy harvest data from solar panels
3	Dishwasher	Energy consumption data on the Dishwasher
4	Office Room	Energy consumption data in Office Room
5	Fridge	Energy consumption data on Fridge
6	Wine Cellar	Energy consumption data at Wine Cellar
7	Garage Door	Energy consumption data on Garage Door
8	Barn	Energy consumption data on Barn
9	Well	Energy consumption data on Well
10	Microwave	Energy consumption data on Microwave
11	Living Room	Energy consumption data in the Living Room
12	Temperature	Environmental Temperature data
13	Humidity	Environmental Humidity data
14	Visibility	Environmental Visibility data
15	apparentTemperature	Environmental apparentTemperature data
16	Pressure	Environmental Pressure data
17	windSpeed	Environmental windSpeed data
18	cloudCover	Environment cloudCover data
19	windBearing	Environmental windBearing data
20	precipIntensity	Environment precipIntensity data
21	dewPoint	Environmental dewPoint data
22	precipProbability	Environmental precipProbability data
23	Furnace	Energy consumption data on Furnace
24	Kitchen	Energy consumption data in the Kitchen

Table 1 explains the data used in this study based on Smart Home data with 24 features. The target of the prediction is "**Use**". Twenty-three other elements are used to predict the target feature.

2.2. **K-Nearest Neighbors.** This algorithm has several k numbers needed to describe the optimal parameters and Manhattan distance for the proposed model. Manhattan distance, as described in Equation (1) and compared with several distance algorithms, utilizes several algorithms (Euclidean, Minkowski, Cosine, Chebyshev) in Equations (2)-(5) based on [22]:

Manhattan distance:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

Minkowski distance:

$$(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{2}$$

Euclidean distance:

$$(x, y) = \|x - y\| \tag{3}$$

Cosine distance:

$$d(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4}$$

Chebyshev distance:

$$d(x, y) = \max(|x_2 - x_1|, |y_2 - y_1|) \tag{5}$$

where $i = 1$ to n, p = positive integer.

The variables x and y from Equations (1) to (5) represent only two vectors in the feature space, and x_i and y_i are their coordinates each [22].

2.3. **Error evaluation.** This section will explain how error evaluation uses RMSE, MSE and MAE with Equations (6)-(8):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - \bar{y}| \tag{6}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 \tag{7}$$

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2} \tag{8}$$

where $\hat{y}$ = predicted value of y, $\bar{y}$ = true value of y.

In Equations (6)-(8), it is used to evaluate the proposed model's prediction model. These results will be compared for all existing hyperparameters, and the best model suggestions will be given based on the results of the slightest error in the KNN model.

2.4. **Model interpretation.** The proposed model will be interpreted using SHAP (Equation (9)) and LIME (Equation (10)) [8,9]:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \mho(g) \tag{9}$$

where $\xi(x)$ = interpretation of results based on data x, G = interpretable model family, f = ML complex models, g = simple model of interpretation, π_x = local neighbourhoods, $\mathcal{L}(f, g, \pi_x)$ = base estimates on local neighborhood's, $\mho(g)$ = manages the complexity of the simple replacement model.

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(F - |z'| - 1)!}{F!}\left[f_x(z') - f_x\left(\frac{z'}{i}\right)\right] \tag{10}$$

where ϕ_i = Shapley's score for feature i, f = model BlackBox, x = input data, $z' \subseteq x'$ = all input feature data, x' = sample input data, F = set of all features.

Equation (9) is the SHAP inspired by game theory. It looks for how much a character contributes to getting exp in each contribution made, especially how the contribution details are, and how appropriately a character can be given results based on their global contribution.

Equation (10) is a LIME widely used in the initial approach to interpreting ML models. LIME's primary goal is to identify models that can be solved locally on predictions. In model interpretation, it is essential to distinguish between data representation and features, which are data representations that humans can understand regardless of the actual features.

3. **Main Results.** In this section, two results from experiments will be discussed. The evaluation results of the proposed model are compared with other distance algorithms. Next, the prediction results will be interpreted, and the interpretation results will be compared.

3.1. **KNN result evaluation.** In this section, the authors focus on the prediction results of the proposed model by utilizing a range of k (1-9), and the results are presented in Figure 2.

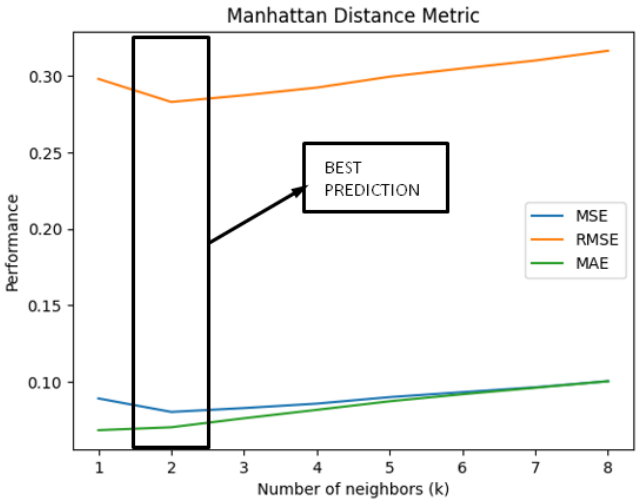


FIGURE 2. Proposed model performance for prediction electricity energy consumption

The proposed model utilizes the Manhattan distance. Figure 2 gets the best k value, 2, and becomes the best prediction. Table 2 presents the comparison results between other distances and the proposed model that utilizes Manhattan.

TABLE 2. Comparing the result performance of distance

No	k	Distance	MSE	MAE	RMSE
1	2	Minkowski	0.0983	0.0578	0.307
2	2	Euclidean	0.0982	0.0554	0.305
3	2	Cosine	0.0985	0.0545	0.318
4	2	Chebyshev	0.1378	0.0876	0.347
5	2	**Manhattan (proposed model)**	**0.0893**	**0.0533**	**0.268**

Table 2 shows that Manhattan distance and $k = 2$ are the best models with an evaluation value, MAE = 0.0533, MSE = 0.0893, and RMSE = 0.268. Manhattan is more robust than other distances, because each distance has an Lk norm, a distance used to measure the distance between two points in a multi-feature space, which is sensitive to high dimensions [19].

3.2. **Interpretation result.** This section will explain which features are most dominantly used by the proposed model for prediction. The fetched features are described in Figure 2, with the prediction target feature being "**Use**", resulting in a comparison with interpretation using LIME, as illustrated in Figures 3 and 4.

Figure 3 describes the most dominant feature as the "Furnace" in high prediction value. Figure 4 explains the dominant feature with the low prediction result is the "Solar Cell" and "Furnace" features. The difference is that when the "Solar Cell" feature generates high energy, it helps reduce electricity consumption, whereas the "Furnace" feature causes low electricity consumption. The use of LIME is limited to looking for markers for low and high predictions. However, the proposed model uses SHAP to look for global feature contributions, not only at the effect of forecasts on low and high predictions. The results of global feature contributions are presented in Figures 5(a) and 5(b).

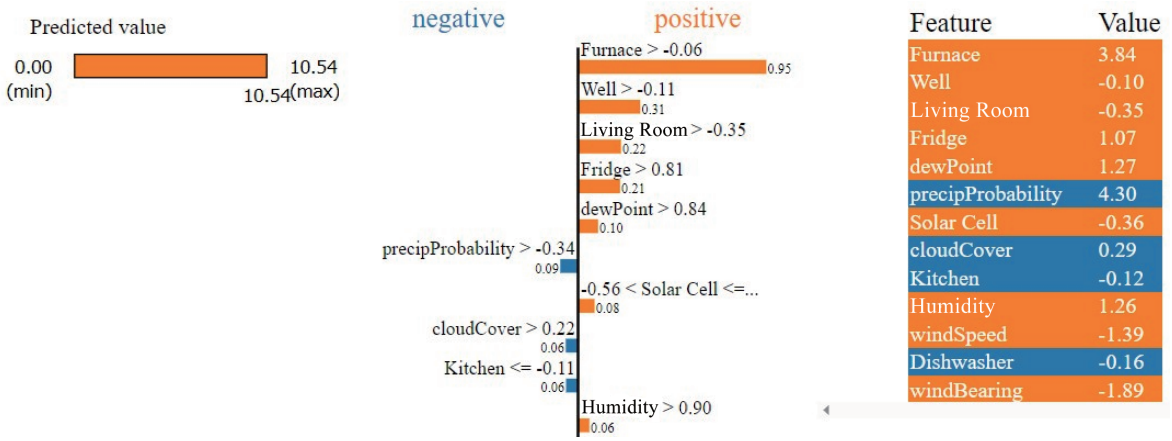


FIGURE 3. LIME result for high prediction value

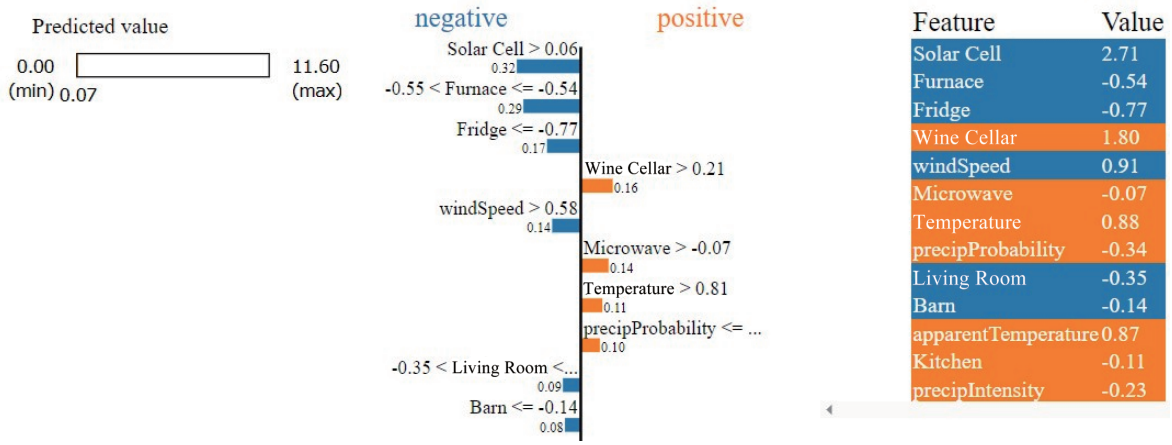
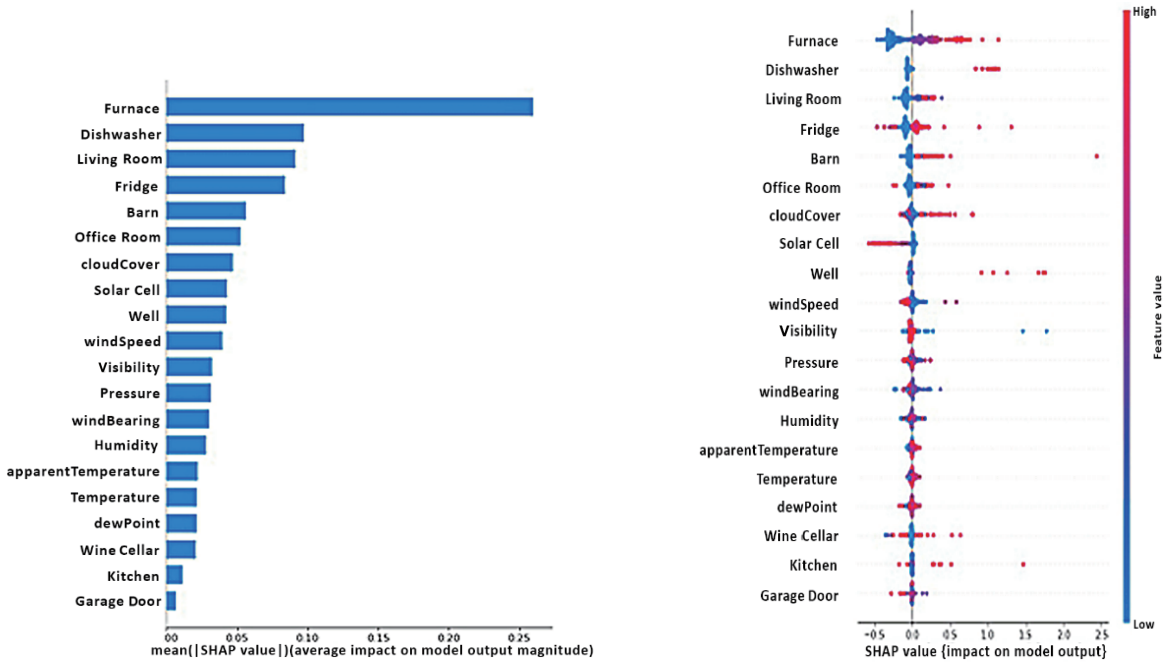


FIGURE 4. LIME result for low prediction value



(a) (b)

FIGURE 5. (a) (Bar Plot/Absolute Value) Dominant feature contribution to prediction globally; (b) (BeeSwarm Plot) feature contribution to prediction globally

The contribution and explanation of the dominant feature are evident, as described in Figure 5(a). The best feature with a SHAP value of 0.25 (a high SHAP value explains high predicted results and vice versa) is "Furnace"; therefore, it is considered a dominant feature in the electrical energy consumption prediction results.

Figure 5(b) shows the result of an experiment that highlights the importance of relationship features of the model, such as

1) Furnace: A high feature value impacts the prediction of electricity consumption results more significantly. Conversely, a low feature value affects the more minor prediction of electricity consumption.

2) Dishwasher: A high feature value impacts the greater electrical energy consumption. Conversely, A low feature value affects the more minor prediction of electricity consumption.

3) Solar Cell: Unlike other features, this feature has a high-value impact on low electricity consumption. Low feature values affect high electricity consumption.

4. **Conclusions.** The proposed model used Manhattan, k = 2, and SHAP has result evaluation as MAE = 0.0533, MSE = 0.0893, and RMSE = 0.268. The results of the proposed model are better than the others because the proposed model is more resistant to many multi-features and data sets on a Smart Home. Furthermore, to describe prediction, SHAP can explain the most dominant features. The SHAP interpreted that "Furnace" is a dominant part of predictions. Meanwhile, the "Solar Cell" feature is the opposite. When the feature value is high, the predictive value is low. It happens because Solar Cells generate electricity. Unlike LIME, which can only interpret features for high or low-value prediction, SHAP can interpret dominant features globally. For further research, the determination of hyperparameters owned by the proposed model is not only k and the distance algorithm. However, many other hyperparameters can usually be searched with search algorithms in AI.

REFERENCES

[1] Y. Ali, Z. Rasheed, N. Muhammad and S. Yousaf, Energy optimization in the wake of China Pakistan Economic Corridor (CPEC), *Journal of Control and Decision*, vol.5, no.2, pp.129-147, DOI: 10.1080/23307706.2017.1353929, 2018.

[2] W. Ostrowski, The twenty years' crisis of European energy security: Central and Eastern Europe and the US, *Geopolitics*, vol.27, no.3, pp.875-897, DOI: 10.1080/14650045.2020.1835863, 2022.

[3] M. Vallés, A. Bello, J. Reneses and P. Frías, Probabilistic characterization of electricity consumer responsiveness to economic incentives, *Applied Energy*, vol.216, pp.296-310, DOI: 10.1016/j.apenergy.2018.02.058, 2018.

[4] M. Yusuf, B. Surya, F. Menne, M. Ruslan, S. Suriani and I. Iskandar, Business agility and competitive advantage of SMEs in Makassar City, Indonesia, *Sustainability*, vol.15, no.1, DOI: 10.3390/su15010627, 2023.

[5] M. Shahjalal, M. K. Hasan, M. M. Islam, M. M. Alam, M. F. Ahmed and Y. M. Jang, An overview of AI-enabled remote smart-home monitoring system using LoRa, *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp.510-513, DOI: 10.1109/ICAIIC48513.2020.9065199, 2020.

[6] T. Mladenova and I. Valova, A review of the application of machine learning in home automation, *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp.1-4, DOI: 10.1109/HORA58378.2023.10156796, 2023.

[7] A. Mosavi and A. Bahmani, Energy consumption prediction using machine learning: A review, *Energies (Basel)*, no.3, pp.1-63, DOI: 10.20944/preprints201903.0131.v1, 2019.

[8] M. T. Ribeiro, S. Singh and C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, New York, NY, USA, pp.1135-1144, DOI: 10.1145/2939672.2939778, 2016.

[9] S. M. Lundberg, P. G. Allen and S.-I. Lee, A unified approach to interpreting model predictions, *Proc. of the 31st International Conference on Neural Information Processing Systems*, 2017.

[10] M. Shao, X. Wang, Z. Bu, X. Chen and Y. Wang, Prediction of energy consumption in hotel buildings via support vector machines, *Sustainable Cities and Society*, vol.57, 102128, DOI: 10.1016/j.scs.2020.102128, 2020.

[11] H. Tabrizchi, M. M. Javidi and V. Amirzadeh, Estimates of residential building energy consumption using a multi-verse optimizer-based support vector machine with k-fold cross-validation, *Evolving Systems*, vol.12, no.3, pp.755-767, DOI: 10.1007/s12530-019-09283-8, 2021.

[12] M. B. Afuosi and M. R. Zoghi, Indoor positioning based on improved weighted KNN for energy management in smart buildings, *Energy Build*, vol.212, 109754, DOI: 10.1016/j.enbuild.2019.109754, 2020.

[13] F. Li and G. Jin, Research on power energy load forecasting method based on KNN, *International Journal of Ambient Energy*, vol.43, no.1, pp.946-951, DOI: 10.1080/01430750.2019.1682041, 2022.

[14] I. Priyadarshini, S. Sahu, R. Kumar and D. Taniar, A machine-learning ensemble model for predicting energy consumption in smart homes, *Internet of Things*, vol.20, 100636, DOI: 10.1016/j.iot.2022.100636, 2022.

[15] R. H. Fard and S. Hosseini, Machine learning algorithms for prediction of energy consumption and IoT modeling in complex networks, *Microprocessors and Microsystems*, vol.89, 104423, DOI: 10.1016/j.micpro. 2021.104423, 2022.

[16] M. R. A. Malek, N. A. Ab. Aziz, S. Alelyani, M. Mohana, F. N. A. Baharudin and Z. Ibrahim, Comfort and energy consumption optimization in Smart Homes using bat algorithm with inertia weight, *Journal of Building Engineering*, vol.47, 103848, DOI: 10.1016/j.jobe.2021.103848, 2022.

[17] S. Zhang, Challenges in KNN classification, *IEEE Transactions on Knowledge and Data Engineering*, vol.34, no.10, pp.4663-4675, DOI: 10.1109/TKDE.2021.3049250, 2022.

[18] C. Ma and Y. Chi, KNN normalized optimization and platform tuning based on Hadoop, *IEEE Access*, vol.10, pp.81406-81433, DOI: 10.1109/ACCESS.2022.3195872, 2022.

[19] S. S. Khan, Q. Ran, M. Khan and M. Zhang, Hyperspectral image classification using nearest regularized subspace with Manhattan distance, *Journal of Applied Remote Sensing*, vol.14, no.3, 32604, DOI: 10.1117/1.JRS.14.032604, 2019.

[20] T. S. Anttal, *Smart Home Dataset with Weather Information*, https://www.kaggle.com/datasets/taranvee/smart-home-dataset-with-weather-information, Accessed on Oct. 29, 2022.

[21] M. H. Widianto, A. A. S. Gunawan, Y. Heryadi and W. Budiharto, Evaluation of machine learning on Smart Home data for prediction of electrical energy consumption, *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pp.434-439, DOI: 10.1109/ICCoSITE57641.2023.10127700, 2023.

[22] X. Gao and G. Li, A KNN model based on Manhattan distance to identify the SNARE proteins, *IEEE Access*, vol.8, pp.112922-112931, DOI: 10.1109/ACCESS.2020.3003086, 2020.