Data-Driven Computational Algorithms for Predicting Electricity Consumption Missing Values: A Comparative Study

Bavly Hanna
Department of Computer Science
University of Technology Sydney
Sydney, Australia
Bavly.s.Hanna@student.uts.edu.au

Guandong Xu
Department of Computer Science
University of Technology Sydney
Sydney, Australia
Guandong.Xu@uts.edu.au

Xianzhi Wang
Department of Computer Science
University of Technology Sydney
Sydney, Australia
Xianzhi.Wang@uts.edu.au

Jahangir Hossain
School of Electrical and Data Engineering
University of Technology Sydney
Sydney, Australia
Jahangir.Hossain@uts.edu.au

Abstract: **Data availability is a significant issue and barrier for modeling and analyzing low voltage networks. This paper develops, implements, and compares several prediction algorithms for finding missing values in energy usage for commercial consumers. Four predictive machine learning models, such as random forest regression, linear regression, multi-layer perceptron, and decision trees, are utilized in this paper. Four commercial users from a regional city in Australia are selected as a dataset based on 30-minute intervals. Firstly, the obtained data is analyzed and pre-processed and then utilized for model training and testing. RMSE and MAE measures are used to compare the effectiveness of each machine learning model. This paper concludes that the multi-layer perceptron model provides better performance than that of random forest, decision tree, and linear regression. The RMSE and MAE of MLP model are 4.7472 and 4.2103, respectively, when using individual users as a training set.**

Keywords: energy consumption prediction, machine learning, random forest regression, linear regression, multi-layer perceptron, decision tree, missing values.

I. INTRODUCTION

In order to meet growing electrical demands in an effective and economical way while reducing glasshouse emissions, there has been an increasing focus on developing and deploying smart grids (SGs) and smart buildings [1]. The expanding intermittent renewable energy sources like wind and solar can support the need for smart grids [2]. A set of dispersed loads are served by a cluster of distributed energy resources known as a microgrid (MG) in both linked and isolated grid modes. Different distributed generating technologies are taken into consideration, with the loads presumed to be variable. System reliability and supply security-related considerations are considered during the design of MGs [3].

The accuracy of the network and demand data determines the quality of the knowledge retrieved, learning, and decision-making issues. Particularly with MGs, the issue of missing values (MVs) can significantly affect the decisions that can be drawn from this data. The quantity and quality of load demand data are crucial for the planning, design, and operation of MGs. As for most rural communities, full data sets are not available; it is important to find effective measures to handle and utilize the partial data set. The determination of the full data from the given partial information can boost the operational flexibility, resilience, integrated energy management capabilities, self-sufficiency, and dependability of power systems [4].

MVs are a genuine issue in the planning and design of any system, but they seem to be more prevalent in power systems because of insufficient sensors and network visibility issues. For utilizing data from smart meters, MV imputation techniques have advanced in the field of power systems research, but further studies are required to determine the most effective way to manage missing data.

In order to calculate the MVs, the following requirements need to be satisfied: (i) The distribution of the data should not be changed by the MV prediction algorithm; (ii) The prediction algorithm must preserve the connections between the data set's attributes; and (iii) The prediction approach should not be very complicated or expensive in terms of time and cost. MVs should be correctly identified and updated in order to make it simple to apply all algorithms for different applications. Consequently, numerous precise and sophisticated machine learning (ML) techniques have been developed as a result of recent developments in computing technology. However, further comparative analyses need to be undertaken to find the best and most reliable method for determining missing values in the data set.

The issue of MVs is often addressed from the perspective of pre-processing. Using the mean value to replace an unknown attribute is a typical MV imputation procedure that could provide results that deviates from ideal results [5]. Lakshminarayan et al. [6] investigate the application of ML-based substitutes to conventional "statistical data" for computing MVs. A unique approach for MVs reconstruction using fuzzy similarity is presented by Barladi et al. [7].

Planning and running sustainable energy systems require a number of fundamental building blocks, one of which is the dynamics of power use, quality, and volume of load data. The problem is that there are missing values in electricity consumption data. This paper compares the accuracy in calculating missing electricity energy consumption values in regional Australia using four machine learning algorithms: random forest regression, linear regression, multi-layer perceptron, and decision trees. The remainder of the paper is organized as follows: ML models: random forest, linear regression, and multi-layer perceptron are discussed in Section 3. Section 4 discusses the proposed methodology and Section 5 represents the dataset. Section 6 includes the

experimental analysis and results, followed by conclusions in Section 7.

II. Related Work

It is unlikely to get complete data from any meters and recorded systems. The completeness of the collected data is mostly related to the dependability of transmission and storage. Between 3 and 4 percent of MVs are recorded in smart meter systems that have been put in place, for instance, because of scheduled outages [8]. MVs in recorded data are typical issues as a result of these difficulties. The majority of applications that involves the MVs, can be handled by pre-processing the data, even if certain applications can accommodate partial data [9].

The time series data from observed power generation or consumption in the context of smart meters often depends on a variety of variables, including the weather, daily routines, societal conventions (such as weekends or vacations), and more [10]. These components frequently result in well-known patterns with various periodicities (intra-day, daily, monthly, etc.) [11]. Studies that address missing data for the building energy system are few. One strategy is to remove any missing numbers and then analyze the behavior of the building using the data that is now available. The problem with this approach is that it may only have a tiny collection of observations to simulate the behavior of the building [12].

Mean imputation is another strategy in which any MV is substituted with the average amount of the remaining variables [13]. The variable's distribution and the connections among variables are distorted by this technique, which can lead to significant discrepancies between anticipated and actual values. The alternative approach to dealing with missing data is to replace those values with constants (e.g., average or zero). This has been utilized for situations where gaps in the data are intolerable [14]. Regression analysis has historically been the most widely used modeling method for estimating energy use [15]. The prediction of energy usage has also been made using artificial neural networks (ANN) [16]. An artificial neural network was trained in [16] on simulated data to create "a mapping between input and output," and then the predictive model was applied to forecast energy usage.

In reality, NNs have shown to be effective tools for data analysis across a variety of fields. The use of decision trees (DTs) as a "decision support" tool for a "production system" has also been demonstrated to be effective [17]. Although infrequently used in energy consumption prediction, a comparison of these diverse "data analysis and modeling" methodologies has been taken into consideration in a number of applications.

The hourly electricity usage of two educational facilities in Florida was predicted using a random forest (RF) [18]. They investigated how well an RF model performed predictions with various parameter values. The outcomes of the simulation showed that the RF was less susceptible to the set of variables and that the empirical methodology was superior. González and Zamarreno [19] employed straightforward "back-propagation NN" for the prediction of short-term construction loads. For the models to forecast hourly energy consumption, actual and predicted amounts of the "current load", "temperature", "hour", and "day" were utilized as parameters. It has been shown that the suggested model produces reliable outcomes.

A straightforward NN may be utilized to connect energy use to several inhabitants and the weather (such as outdoor air temperature and relative humidity), as shown by Nizami and Al-Garni [20]. They determined that ANN performed better after comparing the data with a regression model. Due to its speed and ability to be employed for real-time control applications, ANN models have been constructed in the majority of research in place of sophisticated dynamic simulation programs.

III. Machine Learning Models

The presented predictive models in this section can be utilized to derive imputed figures in which significant portions of the data are not available to deal with the MVs problem in the power consumption dataset.

A. Random Forest Regression

Breinman and Cutle [21] developed the initial RF model. With the help of a voting system, a group of trees is employed in the ensemble technique known as RF to achieve the desired result. Each tree is created employing a randomly chosen "training subset" and a randomly chosen feature subset. This suggests that the trees are reliant on the variables' amounts in the dataset that was individually sampled while applying a uniform distribution for all trees.

The average of each induced tree's predictions serves as the final forecast in the case of regression. Additionally, the caret implementation was applied to this procedure. Due to its relative insensitivity to hyperparameter settings, RF has a significant benefit [22]. Due to its feature of being a group of DTs trained on various portions of the same training set, RFs are also less prone to overfitting.

The attractive qualities that RF provides make it a desirable instrument for predicting energy usage. The first of RF's properties is that it takes into account predictor interaction [23]. The ensemble learning theory on which it is built, enables it to learn both basic and complicated tasks. Finally, compared to other ML approaches (such as ANN, SVM, etc.), RF's hyper-parameters do not require as much fine-tuning, and frequently default settings can produce great results.

B. Linear Regression

The regression approach is frequently employed in forecasting techniques, and when updating MVs using one or more auxiliary variables, the same concepts are used. While an LR technique is more appropriate for datasets with non-binary numerical variables than logistic regression is for datasets with binary variables.

Despite its ease of use and benefits over more complicated prediction methods, the use of LR could result in data that are poorly correlated and have a skewed distribution. The issue is frequently avoided by inserting a noise factor to LR, which also minimizes the "bias" while boosting each anticipated value with a "residual term". Benefits of this "stochastic regression" include the replacement of each unavailable data with a fresh "imputed value" rather than a previously used one [24].

The simplest and most basic regression analysis procedure, known as multiple LR (MLR), builds the connection model between a response variable and a number of explanatory variables. According to the following equation, Eq. (1), its response variable Y is taken to be a linear function:

$$Y = \beta + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

where Y represents the response variable, the explanatory variable is expressed as X_i, and β is the constant coefficient. MLR models were employed in forecasting home energy demands for a very long time because of how simple they are to use.

MLR was used by Trigo-González et al. [25] to calculate Chile's hourly solar power production. In order to determine the energy needs in connection to any weather conditions, Ciulla [26] employed MLR and created a straightforward but accurate energy forecast model. In order to create a model for predicting the performance of a ground source heat pump system time-by-time based on MLR, Park et al. [27] examined the factors that have an impact on the performance of "large-scale ground source heat pump" systems.

C. Multi-Layer Perceptron

ANN is the third approach used in this paper to forecast energy usage. A lot of ANNs have been utilized to forecast building energy demand. Their aptitude for dealing with nonlinear issues has been shown in the literature. ANNs are robust, "fault-tolerant," and "noise-immune" by nature, making it simple for them to model erratic home energy system data. Due to its benefits, such as its capacity to learn complicated behavior, ANNs are regularly used for pattern recognition and predictions [28]. The "input layer," "hidden layer," and "output layer" are the three basic layers that make up the structure of an ANN model. The synaptic weight of each connection connecting the neurons was adjusted until the difference was small (minimizing "Sum Squared Error"), consequently providing regularization for the model. The original output was compared to the intended output [29].

The weight is a graphical depiction of how important a neuron's input is. The network solution structure for this paper was an ANN structure of the Multilayer Perceptron Model (MLP) type with an "error backpropagation" learning technique. The data received by the input layer was computed using an appropriate nonlinear transfer function in the hidden layer. Equation (2) shows the ANN model in detail:

$$y_t = \alpha_0 + \sum_{j=1}^{n} \alpha_j f\left(\sum_{i=1}^{m} \beta_{ij} y_{t-i} + \beta_{0j}\right) + \varepsilon_t \quad (2)$$

where m refers to how many input nodes, n refers to how many hidden nodes, f is the "Sigmoid Transfer function", $\{\alpha_j, j = 0, 1, \ldots, n\}$ is the weights vector from the output layer to the hidden layer and $\{\beta_{ij}, i = 0, 1, \ldots, m;\ j = 0, 1, \ldots, n\}$ is the input to the hidden nodes' weight.

D. Decision Tree

An empirical tree in DT modeling is a segmentation of the data produced by the application of a set of straightforward rules. Through the repeated process of splitting, these models provide a set of rules that may be utilized for prediction. The DT produces a model that may represent interpretable rules or logic statements, which gives it a significant advantage over other modeling techniques.

An essential aspect of trees that produce axis parallel decision surfaces is their explanatory capabilities [30]. Additionally, the classification may be carried out without the need for intricate calculations, and both continuous and categorical data can be employed with the approach. The outcomes of DT models also clearly show how important particular aspects are for categorization or prediction. Though, DT induction is prone to noisy data and typically does not outperform neural networks for nonlinear data [31]. Generally speaking, the method is better suited for categorical result prediction, and DTs are less ideal for use with time series data unless clear trends and sequential patterns are present.

IV. Proposed Methodology

The prediction of the energy consumption is carried out using random forest regression, linear regression, and multi-layer perceptron models. We transfer the knowledge obtained from existing users to more users through ML models. The procedure of the proposed method is illustrated in Figure 1.

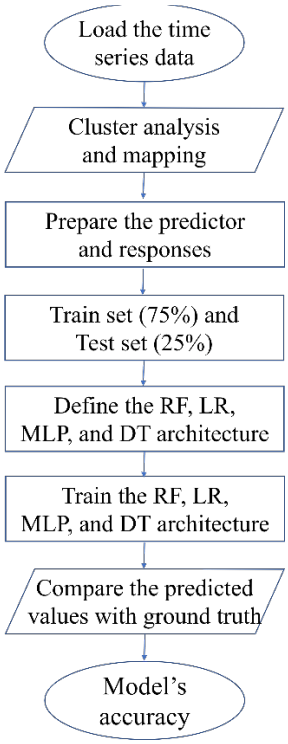


Fig. 1. The procedure of energy consumption prediction using ML models

V. Simulation Data

We use real energy data (kWH) from a regional town in Australia. The training data corresponds to a time period between July 1, 2017, 12:30 AM, and June 1, 2022, 12:00 AM, and was at 30-min time resolution. The model includes four commercial users, and they are divided into 75% "training set" and 25% "test set." The four users are classified in phase 3.

The summary statistics of energy consumption for the four users are illustrated in Table 1. User 4 has the highest average 30-min consumption of 55.35 kWh, while user 3 has the lowest average 30-min consumption of 7.90 kWh during the entire period. The minimum 30-min is 13.27 kWh for user 4, while it is zero for all other users, which means there was no energy consumption during a certain time of the day. User 4 has the highest maximum 30-min of 131.96 kWh, followed by 58.84 kWh, 52.74 kWh, and 39.82 kWh for users 1, 2, and 3, respectively. User 4 has the highest standard deviation of 15.20, while user 3 has the lowest standard deviation of 5.06. User 3 has the highest skewness of 1.06, while user 2 has the lowest skewness of 0.33. User 1 has the highest kurtosis of 1.22, while user 2 has the lowest kurtosis of -0.62. All users have the same 30-min data points of 86, 207, which covers around five years.

TABLE I. SUMMARY STATISTICS OF 30-MIN ENERGY CONSUMPTION DATA FOR FOUR USERS (KWH)

	User 1	User 2	User 3	User 4
Average	13.81	27.44	7.90	55.35
Min.	0.00	0.00	0.00	13.27
Max.	58.84	52.74	39.82	131.96
Standard deviation	7.44	7.97	5.06	15.20
Skewness	1.01	0.34	1.06	0.61
Kurtosis	1.22	-0.62	1.07	0.23
Count	86,207	86,207	86,207	86,207

We aggregate the energy consumption values at each time interval for three users and then get the average. Figure 2 shows the 30-min energy consumption for the average aggregated values of three users over the period of analysis.

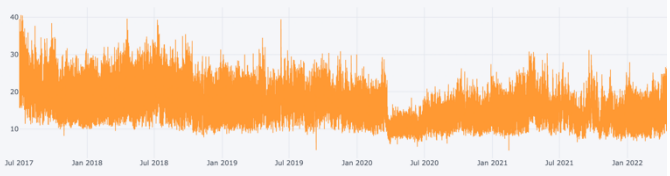


Fig. 2. Timeseries of 30-min energy consumption

Figure 3 represents the Box plot of yearly versus quarterly energy consumption for the average aggregated values of three users. It is obvious that the interquartile range is the highest for the year 2017. The year 2020 has the lowest interquartile range between 12-15 kWH, mainly due to COVID-19. The interquartile range is the highest in quarter three but the lowest in quarter four.

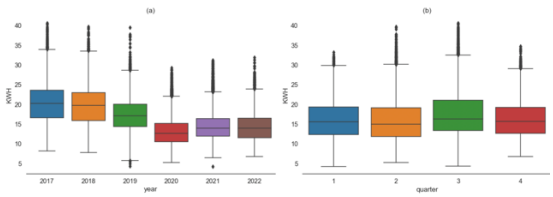
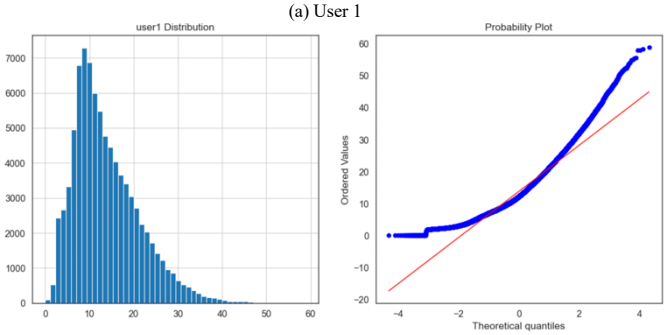


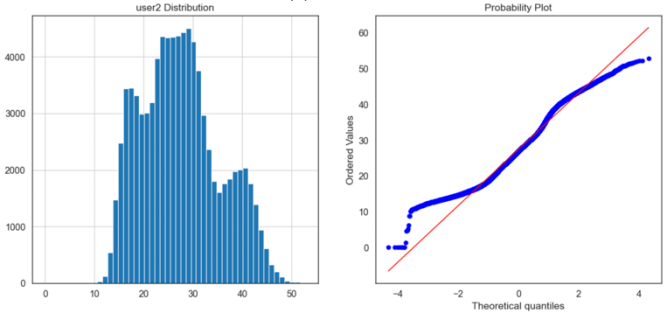
Fig. 3. Box plot of energy consumption: (a) yearly vs. (b) quarterly

Figure 4 shows the energy consumption distribution for each user and the average of aggregated three users. All data are positively skewed with the exception of user 4.

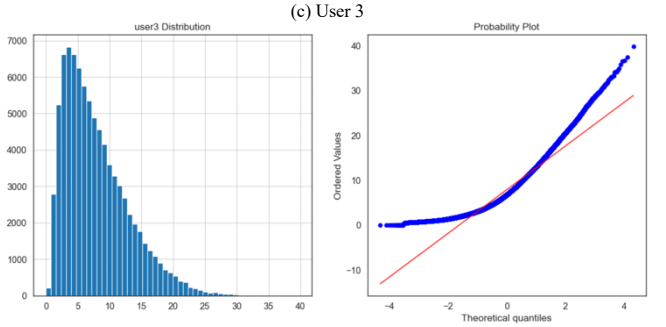
(a) User 1



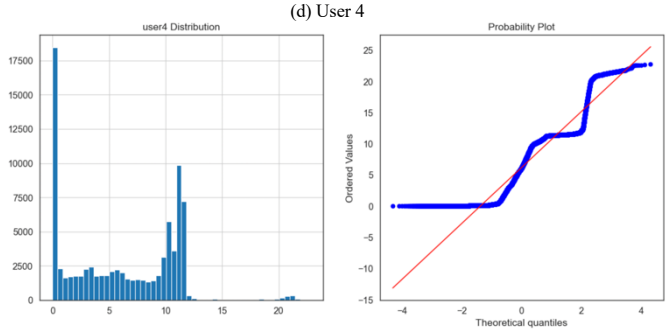
(b) User 2



(c) User 3



(d) User 4



(e) Average aggregated values of three users

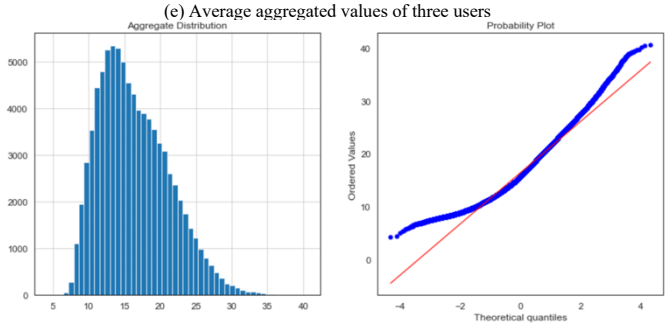
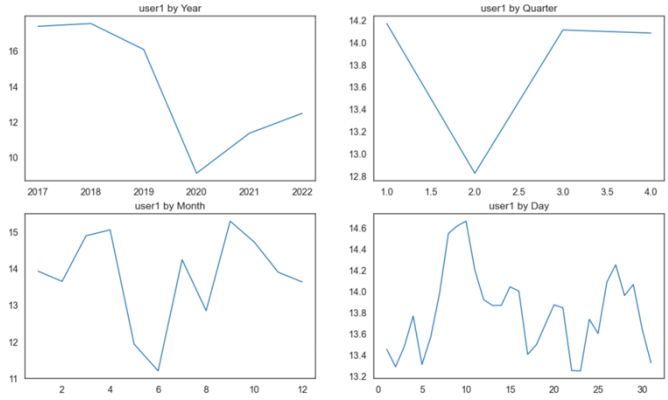


Fig. 4. Energy consumption distribution

Figure 5 shows the mean energy consumption of for each user and an average of aggregated three users grouped by year, quarter, month, and day. For yearly analysis, energy consumption dropped in 2020 due to COVID-19. For quarterly analysis, quarter two has the lowest mean energy consumption while quarter three has the highest mean. For monthly analysis, July has the highest mean energy consumption, while March has the lowest. For daily analysis, day 8 has the highest mean energy consumption, while the lowest is day 2.

(a) User 1



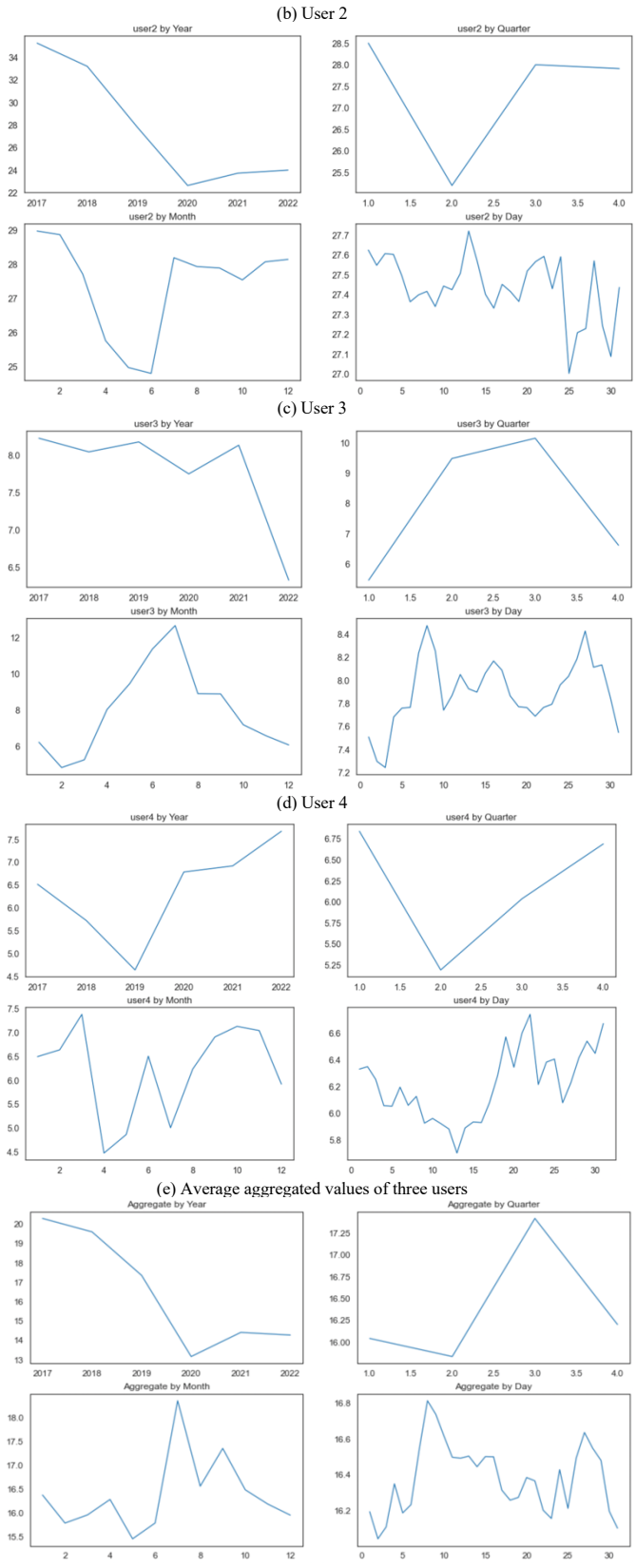


Fig. 5. Mean KWH grouped by year, quarter, month, and day

VI. Experimental Results and Discussion

The data was split into two groups before being fed to the ML algorithm, with 75% of the dataset being used as training data groups and the remaining 25% being utilized for testing. Each ML algorithm was trained using the training sets of data to produce a prediction model. These models produce results that correspond to the data on observed energy use. The remaining data was saved to test the trained prediction model.

Based on "Root Mean Square Error (RMSE)" and "Mean Average Error (MAE)", each forecasting model is assessed. Given that A_t are the actual values of energy consumption and P_t are the forecast values for n data points, the formula is as illustrated in equations (3)-(4) correspondingly. These methods of measurement are useful for comparing the three imputation algorithms.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - P_t)^2}{n}} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - P_t| \quad (4)$$

These two metrics are frequently used to evaluate a technique's effectiveness for time series forecasting [32-33]. These two measurements have the benefit that the average prediction error of a model is given in the same units of the predicted variable. Lower values are preferred for the two measures, which can be assumed to have values larger than or equal to 0. The similarity between the simulated and observed values is shown by both RMSE and MAE.

The sample standard deviation of the discrepancy between the real and the estimated is represented by RMSE. Due to the fact that prediction mistakes are squared, RMSE penalizes big errors severely. As a result, the RMSE might be helpful when we wish to avoid making significant forecasting mistakes.

The MAE calculates the average size of the predictions' mistakes. Since MAE expresses the absolute mistake, it is simple to comprehend.

We provide two case studies, an individual user's training set and an average aggregated training set. We analyze individual users because it is important in the operation of the microgrid, while we analyze the average of aggregated users because it is important in the planning phase.

A. Case Study 1

The first case study refers to training the dataset of individual users (1, 2, and 3) and using user 4 as a test set. Table 2 reports the RMSE and MAE of the predicted values against the actual values of the four ML models. MLP model shows superior predictive power in comparison with RF, LR and DT models. The RMSE value of MLP, LR, RF, and DT are 4.7472, 4.7480, 9.6486, and 6.7745, respectively. The MAE value of MLP, LR, RF, and DT are 4.2103, 4.2111, 7.8740, and 5.2971, respectively.

TABLE II. Scenario Analysis (Individual Users Training Set)

	RMSE	**MAE**
Random Forest Regression	9.6486	7.8740
Linear Regression	4.7480	4.2111
Multi-Layer Perceptron	4.7472	4.2103
Decision Trees	6.7745	5.2971

B. Case Study 2

As a robust test, we tested one user against the average of aggregated values of three users (1, 2, and 3) as a training set. The results of RMSE and MAE are shown in Table 3. The values of RMSE and MAE are close in case study 1 and 2 but case study 1 is slightly better than case study 2 for RF, LR, and MLP models. Having the training set of individual users yields better results for all ML models compared to having the training set of average aggregated values of the remaining individual users. This is because the higher the number of

users, the higher the predictive power of the ML models, as we have more data in the training set to interpolate trends.

TABLE III. SCENARIO ANALYSIS (AVERAGE AGGREGATED USERS)

	RMSE	MAE
Random Forest Regression	9.6500	7.8918
Linear Regression	4.7486	4.2115
Multi-Layer Perceptron	4.7481	4.2114
Decision Trees	6.1337	4.9200

VII. CONCLUSION

The ability to estimate energy usage is crucial for facility managers, building owners, and energy suppliers to make well-informed decisions. In this paper, four ML models were used to impute the missing values (RF, LR, MLP, and DTs). The RMSE and MAE techniques were used to calculate the difference between the anticipated values and actual values. The fact that the RMSE and MAE variations between the four types of models are often relatively minimal shows that the four modeling approaches are largely equivalent in forecasting energy consumption. We get to the conclusion that, generally, the MLP model outperforms both RF, DTs, and conventional LR in terms of predictive power. The RMSE and MAE of MLP model are 4.7472 and 4.2103, respectively, when using individual users as a training set. In terms of future work, we intend to expand the training set by using more data and adding other ML models (naïve mean and MLP tuned), since this method has been successful in solving other time series forecasting issues. We will also include a performance comparison of these algorithms with real-time prediction.

REFERENCES

[1] M.E. El-hawary, "The smart gridstate-of-the-art and future trends," Electr Power Compon Syst, 42(3-4): 239-50, 2014. http://dx.doi.org/10.1080/15325008.2013.868558.

[2] E. Mocanu, P.H. Nguyen, W.L. Kling and M. Gibescu, "Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning," Energy Build 116: 646-55, 2016.

[3] S.A. Arefifar, Y.A.R.I. Mohamed, and T.H.M. El-Fouly, "Optimum microgrid design for enhancing reliability and supply-security," IEEE Transactions on Smart Grid, vol. 4, no. 3, pp. 1567-1575, 2013. Doi: 10.1109/TSG.2013.2259854.

[4] S.A. Arefifar and Y.A.R.I. Mohamed, "DG mix, reactive sources and energy storage units for optimizing microgrid reliability and supply security," IEEE Transactions on Smart Grid, vol. 5, no. 4, pp. 1835-1844, 2014. Doi: 10.1109/TSG.2014.2307919.

[5] V. Tresp, A. Ahmad and R. Neuneier, "Training neural networks with deficient data," Advances in Neural Information Processing Systems 6, pp. 128-135, 1994.

[6] K. Lakshminarayan, S.A. Harp, R. Goldman, and T. Samad, "Imputation of missing data using machine learning techniques," in Proceedings: Second International Conference on Knowledge Discovery and Data Mining, pp. 140-145, 1996.

[7] P. Baraldi, F. Di Maio, D. Genini, and E. Zio, "Reconstruction of missing data in multidimensional time series by fuzzy similarity," Applied Soft Computing, vol. 26, pp. 1-9, 2015.

[8] J. Peppanen, M.J. Reno, M. Thakkar, S. Grijalva, and R.G. Harley, "Leveraging AMI data for distribution system model calibration and situational awareness," IEEE Transactions on Smart Grid, vol. 6, no. 4, pp. 2050-2059, Jul. 2015.

[9] S.J. Taylor, and B. Letham, "Forecasting at scale," The American Statistician, vol. 72, no. 1, pp. 37-45, Jan. 2018.

[10] J. Peppanen, X. Zhang, S. Grijalva, and M.J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2016, pp. 1-5, 2016.

[11] J.A.G. Ordiano, S. Waczowicz, V. Hagenmeyer, and R. Mikut, "Energy forecasting tools and services," WIREs Data Mining and Knowledge Discovery, vol. 8, no. 2, p. e1235, 2018.

[12] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. Brown, and R. Pendyala, "Machine learning approaches for estimating commercial building energy consumption," Applied energy, vol. 208, pp. 889-904, 2017.

[13] D. Cabrera, and H. Zareipour, "Data association mining for identifying lighting energy waste patterns in educational institutes," Energy and Buildings, vol. 62, pp. 210-216, 2013.

[14] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520-525, 2001.

[15] G.K.F. Tso, and K.K.W. Yau, "A study of domestic energy usage pattern in Hong Kong," Energy, 28:1671-82, 2003.

[16] S.A. Kalogirou, and M. Bojic, "Artificial neural networks for the prediction of the energy consumption of a passive solar building," Energy, 25:479-91, 2000.

[17] W. Muller, and E. Wiederhold, "Applying decision tree methodology for rules extraction under cognitive constraints," Eur J Oper Res, 136: 282-9, 2002.

[18] Z. Wang, Y. Wang, R. Zeng, R. Srinivasan, and S. Ahrentzen, "Random Forest based hourly building energy prediction," Energy Build, 171, 11-25, 2018.

[19] P.A. González, and J.M. Zamarreno, "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network," Energy Build. 37(6), 595-601, 2005. http://dx.doi.org/10.1016/j.enbuild.2004.09.006, ISSN 0378-7788.

[20] S.J. Nizami, and A.Z. Al-Garni, "Forecasting electric energy consumption using neural networks," Energy Policy, 23(12), 1097-1104, 1995. http://dx.doi.org/10.1016/0301-4215(95)00116-6, ISSN 0301-4215.

[21] L. Breiman, "Random Forests," Mach. Learn., 45, 5-32, 2001.

[22] G. Dudek, "Short-Term Load Forecasting Using Random Forests". In: Intelligent Systems '2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS' 2014, Warsaw, Poland 2 (2015), pp. 821-828.

[23] A. Statnikov, L. Wang, and C.F. Aliferis, "A comprehensive comparison of random forests and support vector machines for micro array based cancer classification", BMC Bioinform, 9 (1), 2008.

[24] D. Schunk, "A Markov chain Monte Carlo algorithm for multiple imputation in large surveys," AStA Adv. Stat. Anal. 92(1), 101-14, 2008.

[25] M. Trigo-González, F.J. Batlles, J. Alonso-Montesinos, P. Ferrada, J. del Sagrado, M. Martínez-Durbán, M. Cortés, C. Portillo, and A. Marzo, "Hourly PV production estimation by means of an exportable multiple linear regression model," Renew. Energy, 135, 303-312, 2019. https://doi.org/10.1016/j.renene.2018.12.014.

[26] G. Ciulla, and A. D'Amico, "Building energy performance forecasting: a multiple linear regression approach," Appl. Energy, 253, 113500, 2019. https://doi.org/10.1016/j.apenergy.2019.113500.

[27] S.K. Park, H.J. Moon, K.C. Min, C. Hwang, and S. Kim, "Application of a multiple linear regression and an artificial neural network model for the heating performance analysis and hourly prediction of a large-scale ground source heat pump system," Energy Build. 165, 206-215, 2018. https://doi.org/10.1016/j.enbuild.2018.01.029.

[28] S.L. Karunathilake, and H.R.K. Nagahamulla, "Artificial neural networks for daily electricity demand predicitons of Sri Lanka," In: International Conference on Advances in ICT for Emerging Regions (ICTer), 2017.

[29] Z. Liu, D. Wu, Y. Liu, Z. Han, L. Lun, J. Gao, G. Jin, and G. Cao, "Accuracy analyses and model comparison of machine learning adopted in building energy consumption prediction," Energy Explor. Exploit., 1-26, 2019.

[30] P. Perner, U. Zscherpel, C. Jacobsen, "A comparison between neural networks and decision trees based on data from industrial radiographic testing,". Pattern Recognition Lett 2001; 22:47-54.

[31] S.P. Curram, J. Mingers, "Neural networks, decision tree induction and discriminant analysis: an empirical comparison," J Oper Res Soc 1994; 45:440-50.

[32] R. Talavera-Llames, R. Pérez-Chacón, A. Troncoso, and F. Martínez-Álvarez, "MV-kWNN: A novel multivariate and multi-output weighted nearest neighbours algorithm for big data time series forecasting," Neurocomputing, 353, 56-73, 2019.

[33] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, and F. Martínez-Álvarez, "Multi-step forecasting for big data time series based on ensemble learning," Knowl.-Based Syst., 163, 830-841, 2019.