

Modeling Baseline Energy Using Artificial Neural Network: A Small Dataset Approach

Wan Nazirah Wan Md Adnan¹, Nofri Yenita Dahlan², Ismail Musirin³

^{1,2,3}Faculty of Electrical Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

¹Faculty of Engineering and Life Sciences, Universiti Selangor, Bestari Jaya, Selangor, Malaysia

Article Info

Article history:

Received Jun 1, 2018

Revised Jul 30, 2018

Accepted Aug 7, 2018

Keywords:

Baseline Energy Model
Artificial Neural Network
Cross Validation
Bootstrap
Small dataset

ABSTRACT

In this work, baseline energy model development using Artificial Neural Network (ANN) with resampling techniques; Cross Validation (CV) and Bootstrap (BS) are presented. Resampling techniques are used to examine the ability of the ANN model to deal with a small dataset. Working days, class days and Cooling Degree Days (CDD) are used as ANN input meanwhile the ANN output is monthly electricity consumption. The coefficient of correlation (R) is used as performance function to evaluate the model accuracy. For this analysis, R is calculated for the entire data set (R_all) and separately for training set (R_train), validation set (R_valid) dan testing set (R_test). The closer R to 1, the higher similarities between targeted and predicted output. The total of two different models with several number of neurons are developed and compared. It can be concluded that all models are capable to train the network. Artificial Neural Network with Bootstrap Cross Validation technique (ANN-BSCV) outperforms Artificial Neural Network with Cross Validation technique (ANN-CV). The 3-6-1 ANN-BSCV, with R_train = 0.95668, R_valid = 0.97553, R_test = 0.85726 and R_all = 0.94079 is selected as the baseline energy model to predict energy consumption for Option C IPMVP.

Copyright © 2018 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Wan Nazirah Wan Md Adnan,
Faculty of Electrical Engineering,
Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia.
Email: wan_nazirah@unisel.edu.my

1. INTRODUCTION

In recent decades, electricity consumption has dramatically increased due to the population and economic growth. In Peninsular Malaysia, electricity consumption currently increased by 7.5% in 2 years, from 2012 to 2014 [1]. This situation has encouraged Malaysian government to implement energy efficiency of energy conservation measures to reduce electricity consumption. To evaluate the impact of this implementation, the reduction in energy consumption and energy saving must be determined. The evaluations are very dependent on Measurement and Verification (M&V) activities. The application of International Performance Measurement and Verification Protocol (IPMVP) is required as M&V reference protocol to evaluate energy savings. The IPMVP provides a framework for determining saving after implementation of an energy saving project [2].

There are four M&V options defined by the IPMVP i.e. 1) Option A–Retrofit isolation with key parameter measurement, 2) Option B – Retrofit isolation with all parameter measurement, 3) Option C–whole facility and 4) Option D–calibrated simulation. For this study, only IPMVP Option C- Whole facility energy use is chosen and data are normally derived from monthly utility bills. One of the M&V activities is developing the baseline energy model. Currently, baseline energy model is analyzed through a linear

regression [3-4]. However, this method may not accurately represent the data characteristic, especially for the non-linear characteristic.

Recently, Artificial Neural Network (ANN) is a popular technique for forecasting and classifying the output. It has been used in different fields by researchers to solve various problems. The ability of ANN to find the relationship between inputs-outputs is highly accurate. For this study, ANN has been selected to replace the linear regression. As previously mentioned, Option C data is derived from utility bills and usually, the available data is insufficient to train the network. The data are divided into three; training, validation and testing set. Training and validation set are participating in the ANN learning process, and the testing set is used to evaluate the performance of the trained model. ANN learns from examples; therefore, small data sets normally creates inaccurate results and produce a large training error.

A few researchers reported and explored several resampling techniques to deal with small data size in ANN application [5-8]. This study focuses on the development of baseline energy model and the integration of ANN with Cross Validation (CV) and Bootstrap (BS) to get a better accuracy of ANN prediction. This method may avoid any overfitting of the data. Overfitting creates the network to memorize training patterns, but they cannot generalize well to new data (testing set) and generates poor accuracy.

The structure of this paper is organized as follows: Section 2 briefly explains the methodology including data collection, baseline energy model development, and performance evaluation. This is followed by a presentation of the result and discussion of the proposed methods. Finally, the conclusion is summarized in section 4.

2. RESEARCH METHOD

This section describes the proposed methodology for the development of baseline energy model for IPMVP Option C small data set as illustrated in Figure 1.

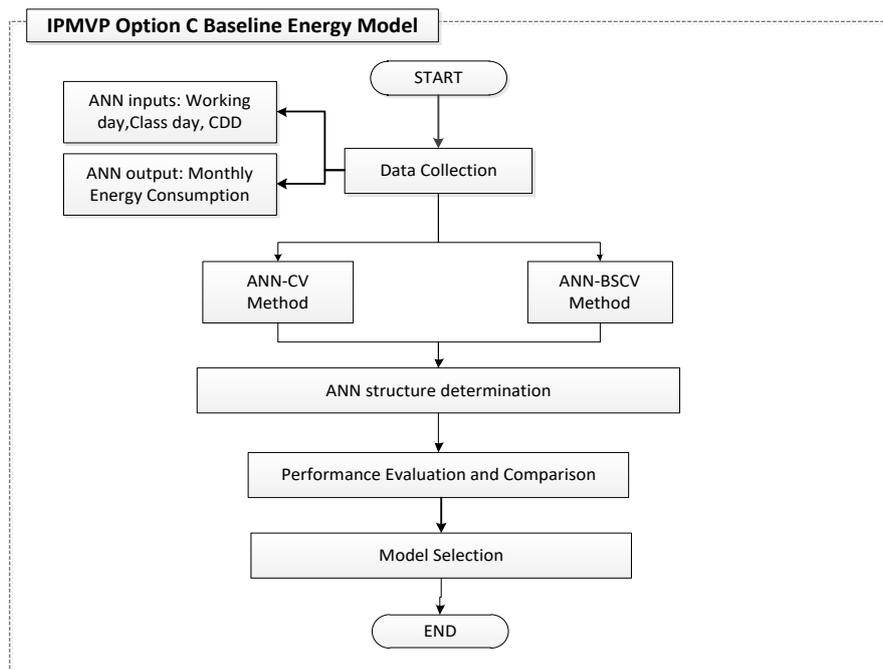


Figure 1. Proposed IPMVP Option C Baseline Energy Model

2.1. Data Collection

The baseline data is provided by the Facility Management Office, Universiti Teknologi Mara (UiTM), Shah Alam, Selangor, Malaysia for a 23 months period. The input variables are UiTM working days (WD), UiTM class days (CD) and Cooling Degree Days (CDD). These three variables are assigned as ANN input and targeted output is monthly electricity consumption, kWh.

Baseline statistical data; minimum, maximum and mean of baseline and post-retrofit are shown in Table 1.

Table 1. Statistical data for Baseline

Parameter	Range	Mean
Baseline input: working days	17 – 23 days	21 days
Baseline input: class days	0 – 23 days	14 days
Baseline input: cooling degree days	511 - 649	579
Baseline output: electrical energy consumption	63,002 kWh- 113,613 kWh	86,377kWh

2.2. IPMVP Option C Baseline Energy Model Development

Since limited data is available for this study, CV and BS techniques are embedded with ANN to develop an accurate baseline energy model. In this section, two baseline energy models, Artificial Neural Network–Cross Validation (ANN-CV) model and Artificial Neural Network–Bootstrap Cross Validation (ANN-BSCV) model are developed.

ANN structures and parameters need to be determined to develop the model. Feed-forward multilayer perceptron with single hidden layer is chosen due to this is the most popular ANN architecture [9]. The training algorithm used and recommended by the MATLAB and mostly used to train the network is `trainlm` (Levenberg-Marquardt) [10-11]. The selected transfer functions are `logsig` for hidden layer and `purelin` for the output layer. The 23 sets of data are divided into a training set, validation set and test set. The training set is used to train the network and to adjust ANN weights and biases. The validation set is used to validate the training process and to control overfitting through early stopping. The testing set is used to evaluate the performance of the trained model since the test set is not directly involved in the training process. In ANN-CV and ANN-BSCV, the networks are run for several times and average values of all four Rs are recorded.

Two types of CV are implemented in ANN-CV methods as illustrated in Figure 2, 1) K-Fold Cross Validation (KFCV): the data are partitions into k-samples and 2) Random Sampling Cross Validation (RSCV): each data splits randomly into k-subsamples. Training and validation set are divided into k-subsample and k=5 is selected in this study. In first iteration, k-1 subsamples are dedicated to training set to train the data meanwhile the remaining subsample is used to validate the date and the process is repeated for k iterations. The number of neurons in the hidden layer is set between 6 and 20 neurons only [9]. The graphical illustration of ANN_CV is presented in Figure 2.

The best structure of ANN-CV is used to developed ANN-BSCV model. The purpose of bootstrap resampling is to create a set of bootstrap samples that can provide a better understanding of the average and variability of the original unknown distribution or process [12]. Bootstrap resampling is performed by repeated sampling with replacement from the original sample. It involves generating a large number of new samples from the original sample.

Figure 4 illustrated the process of bootstrap resampling and Figure 5 graphically described ANN_BSCV model. The original sample of size N (N=8 in this illustration) randomly samples with replacement and repeated B times to produce B datasets, each contains different data with a size of N. B neural network models are developed to train and validate B bootstrap data, meanwhile the unused data is used to test the neural networks. Instead of selecting a single neural work that is the “best”, B neural network models are combined to approximate the relationship between model inputs and model outputs and also to improve model accuracy and robustness. In this study, B = 100 is selected since most common value is between 50 and 200 [7], [13-14]. The ANN_BSCV model fitted to each of the generated bootstrap dataset and ANN_BSCV estimate value is calculated as the average regression function of each model as in Equation 1.

$$F_{ANN_{BSCV}}(x) = \frac{1}{B} \sum_{b=1}^B f_{ANN_{BSCV}}(x) \quad (1)$$

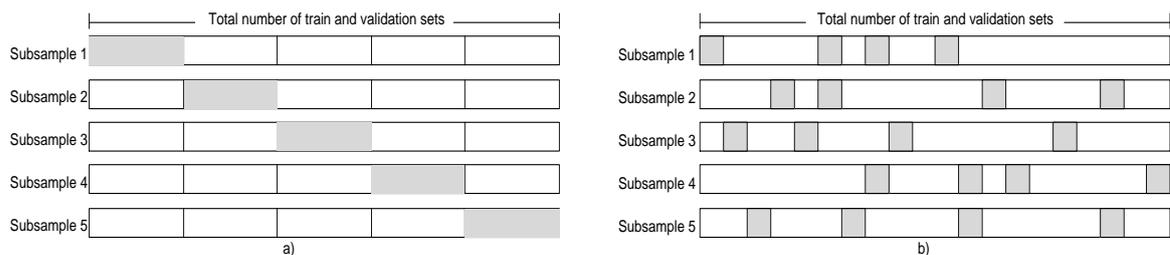


Figure 2. Cross-Validation resampling method: a) KFCV; b) RSCV

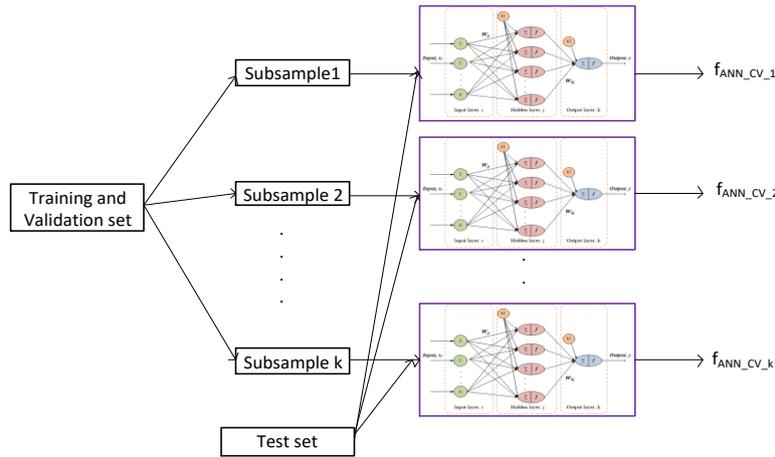


Figure 3. ANN_CV Model

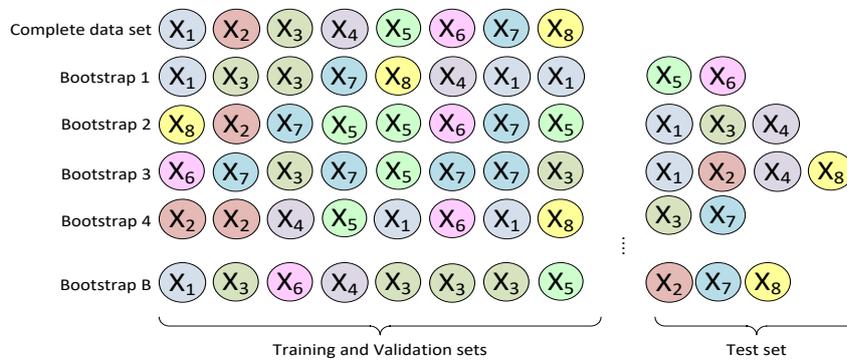


Figure 4. Bootstrap resampling method

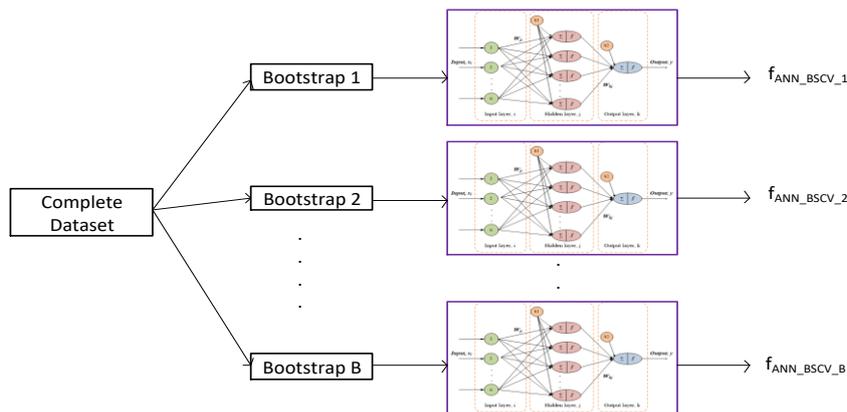


Figure 5. ANN_BSCV Model

2.3. Performance Evaluation and Model Selection

Several methods are integrated with ANN in order to overcome the small dataset problems and to increase accuracy in ANN. The performance of ANN is not only evaluated on single ANN but repeatedly measured with different configuration such as split of training, validation and testing set in CV; and randomly choosing data in BS. In each run, several ANN with various configurations are trained, producing a range of successful and unsuccessful ANN results.

To evaluate the ANN model performance, the Coefficient of Correlation, R between the targeted output and predicted output is measured. R measures the strength of correlation between two variables [18]. The closer R to 1, the higher similarities between targeted and predicted output. For this small dataset analysis, Coefficient of Correlation is calculated for the entire data set (R_{all}) and separately for training set (R_{train}), validation set (R_{valid}) dan testing set (R_{test}). For IPMVP, $R > 0.75$ defines statistically significant performance. The model with the highest R is selected as the baseline model.

3. RESULTS AND ANALYSIS

This section explains the result of baseline energy model using two methods, ANN-CV and ANN-BSCV. At the end of this section, performance comparisons of both models are evaluated. The model with the highest values of R is selected as the best baseline energy model.

3.1. Baseline Energy Model Result

As previously mentioned, ANN-CV model structure is developed using RSCV and KFCV. Since there is no principle in choosing the proper ANN architecture, the number of hidden neurons is determined using cross-validation technique to obtain the best structure. The hidden neurons in single hidden layer are varied and the network is trained for each addition of neurons. The total thirty structures are trained and evaluated for both ANN-CV model. In this ANN-CV method, data is divided into 5 subsamples and randomly runs for several times. The ability of ANN to learn is also supported by the accuracy of training, validation and testing data. The average values of training, validation, testing as well as entire data for each neuron is recorded and the effect of hidden neurons on overall performance for all R s are summarized in Figure 6 for ANN-KFCV and Figure 7 for ANN-RSCV.

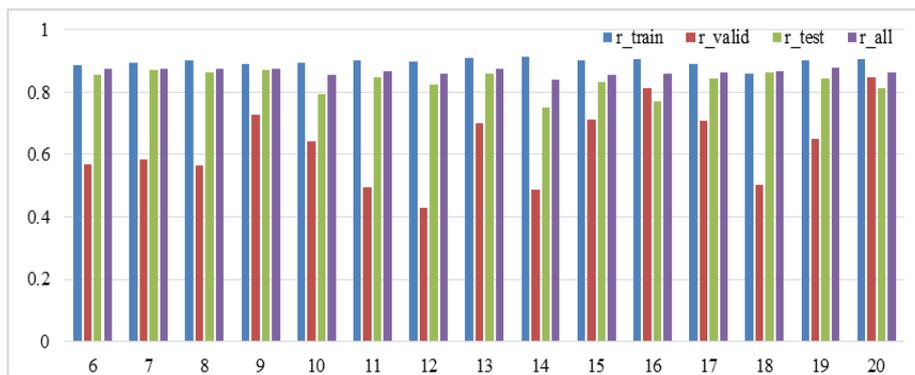


Figure 6. Performance comparison of Coefficient of Correlation (R_{train} , R_{valid} , R_{test} and R_{all}) for different neurons using ANN-KFCV model

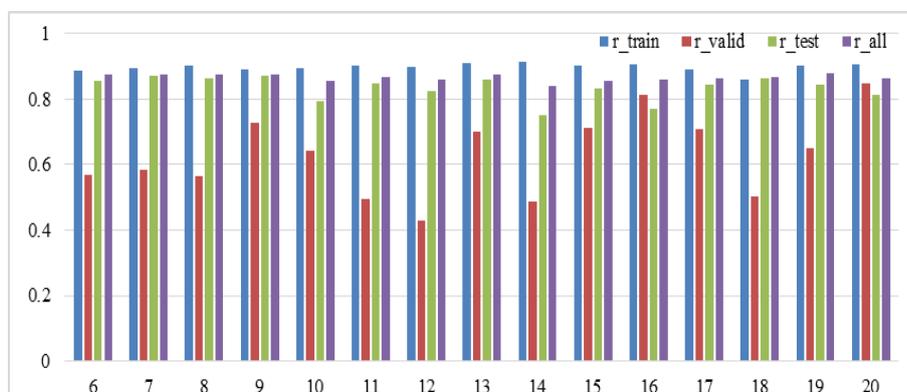


Figure 7. Performance comparison of Coefficient of Correlation (R_{train} , R_{valid} , R_{test} and R_{all}) for different neurons using ANN-RSCV model

In Figure 6, the lowest R_{train} is 0.8893, R_{valid} is 0.6888, R_{test} is 0.8200 and R_{all} is 0.8449, meanwhile the highest R_{train} is 0.9074, R_{valid} is 0.7591, R_{test} is 0.8580 and R_{all} is 0.8725. In Figure 7, the lowest R_{train} is 0.8570, R_{valid} is 0.4301, R_{test} is 0.7506 and R_{all} is 0.8391, meanwhile the highest R_{train} is 0.9139, R_{valid} is 0.8480, R_{test} is 0.8719 and R_{all} is 0.8758. In both figures, the training set gives the highest performance meanwhile the validation set gives the lowest performance. Therefore, it is important to select the best structure to avoid overfitting, especially in validation set.

In general, the increasing of hidden neurons not necessarily increase the performance of ANN. More hidden neurons are added to the network, the network becomes more complex and takes longer time to run. In fact, network with 6 and 9 of neurons give higher performance in ANN-KFCV but in ANN-RSCV, the network with 20 neurons gives the best results. This ANN-CV structure with 6,9 and 20 number of neurons in hidden layer which give the best performance has been selected and used in ANN-BSCV method.

In ANN-BSCV methods, bootstrap resampling method is used to generate different combinations of the data sets to create a set of 100 bootstrap samples using resampling with replacement. This means that a total of 3 structures are developed with 100 different set of bootstrap samples in each model are evaluated. The performance of average all R_s for 6, 9 and 20 neurons in hidden layer are summarized in several boxplots as in Figure 8 which shows the median, minimum and maximum values and also upper and lower quartiles of the results. The R_s values that away from the top or the bottom of the box are considered as outliers and denoted by the dot signs.

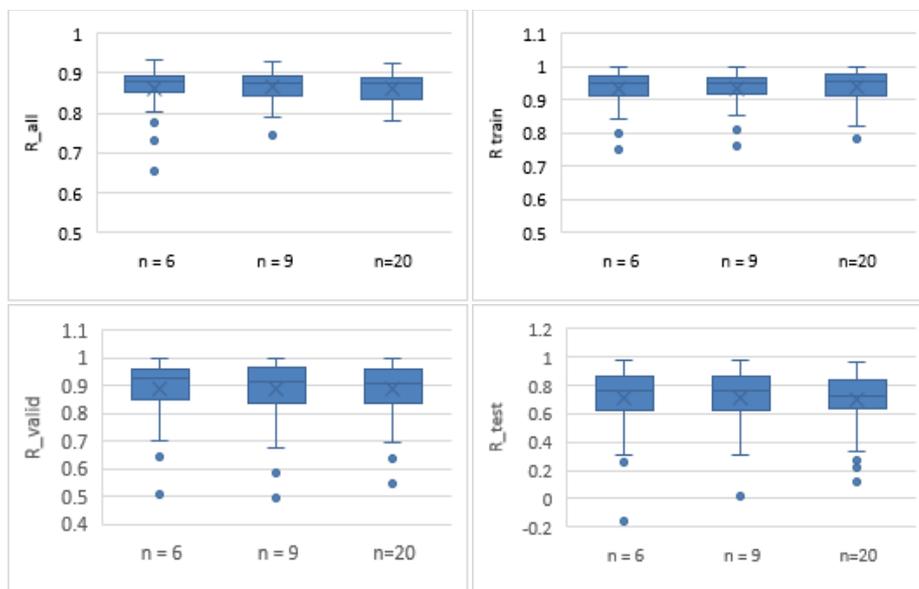


Figure 8. Boxplots of Coefficient of Correlation (R_{train} , R_{valid} , R_{test} and R_{all}) for different neurons using ANN-BSCV model

R_{all} varies between 0.6555 and 0.9270, R_{train} varies between 0.7518 and 1, R_{valid} varies from 0.5061 and 1, and R_{test} varies from -0.1531 and 0.9796 with 6 neurons in hidden layer. Meanwhile, R_{all} varies between 0.7465 and 0.9301, R_{train} varies between 0.7631 and 1, R_{valid} varies from 0.4969 and 1, and R_{test} varies from 0.0158 and 0.9736 with 9 neurons in hidden layer. R_{all} varies between 0.7817 and 0.9232, R_{train} varies between 0.7854 and 1, R_{valid} varies from 0.5484 and 1, and R_{test} varies from 0.1176 and 0.9691 with 20 neurons in hidden layer. As can be seen, all the three neurons tend to underpredict R_{valid} and R_{test} , this is due to the ANN poor ability to predict the values. The median for all R_s for 6 neurons in hidden layer is higher than the other two neurons. By comparison, all three structures have acceptable median values and also have smaller range for all R_{all} . In this case, R_{all} and R_{train} are less sensitive to the available data sets with less outliers.

3.2. Performance Comparison and Model Selection

The results of both proposed methods with three best structures for each method are selected and summarized as in Table 2. The results of ANN method without any resampling techniques are also presented in this table to highlight the strength of the proposed model. In view of the results obtained, the coefficient of

correlation for both proposed methods are above 0.9 which can be considered acceptable and adhere to the specification of IPVMP [2]. It can be seen that all R_{valid} for ANN without any resampling techniques are less than 0.4 which this situation creates overfitting during ANN training process. In selecting the best baseline energy model, all four Rs need to be considered to avoid overfitting.

These results indicate that all the proposed models have a high quality of prediction and capable to train the network. The ANN-BSCV method evidently outperforms ANN-CV method. Further comparison of all structures is made and clearly shows that the model of ANN-BSCV with 6 neurons in hidden layer has the highest value of R_{all} and is nominated as the best performance. It is based on the higher values of all four Rs which are close to unity and therefore the most accurate.

Table 2. Result summarization

Method	Number of hidden neurons	R_{train}	R_{valid}	R_{test}	R_{all}
ANN	6	0.91915	0.34414	0.88500	0.89849
	9	0.92242	0.25227	0.91236	0.88182
	20	0.91148	0.17804	0.88179	0.86935
ANN-CV	6	0.93322	0.87192	0.96103	0.90761
	9	0.84015	0.90649	0.94010	0.90508
	20	0.93421	0.89308	0.98286	0.90099
ANN-BSCV	6	0.95668	0.97552	0.85726	0.94079
	9	0.94506	0.97478	0.88684	0.93858
	20	0.98573	0.93317	0.83532	0.93928

4. CONCLUSION

The main objective of this paper is to develop baseline energy model for Option C IPMVP for small dataset available. In this paper, a combination of ANN with resampling techniques, BS and CV are presented. Two methods, ANN-CV and ANN-BSCV with several structures are proposed in this paper. It is found that the proposed techniques to integrate ANN with resampling techniques improved the prediction performance. The ANN input variables consisted of working days, class days and Cooling Degree Days and targeted output is monthly electricity consumption, kWh. Based on R_{all} , R_{train} , R_{valid} and R_{test} , a significant improvement is observed in the comparison of the results of both models and it is found that the ANN-BSCV model with 3-6-1 structure gave better accuracy based on highest values of all four Rs than the other structures. This structure is selected as the baseline energy model to predict energy consumption for Option C IPMVP. The results obtained from this paper show that the CV and BS method capable to train neural networks, avoid overfitting and create model diversity with limited dataset. For future works, optimization technique needs to be embedded with resampling techniques to obtain better ANN performance accuracy in developing baseline energy model.

REFERENCES

- [1]. Ministry Coordinator of Strategic Sectors, "National Energy Balance 2014," 2014.
- [2]. Efficiency Valuation Organization, "International Performance Measurement and Verification Protocol (IPMVP)," 2012.
- [3]. O. Akinsooto, D. De Canha, and J. H. C. Pretorius, "Energy savings reporting and uncertainty in Measurement & Verification," in *Australasian Universities Power Engineering Conference, AUPEC 2014, Curtin University, Perth, Australia*, 2014, no. October, pp. 1–5.
- [4]. S. M. Aris, N. Y. Dahlan, M. N. M. Nawi, T. A. Nizam, and M. Z. Tahir, "Quantifying energy savings for retrofit centralized hvac systems at Selangor state secretary complex," *J. Teknol.*, vol. 77, no. 5, pp. 93–100, 2015.
- [5]. A. Pasini, "Artificial neural networks for small dataset analysis," *J. Thorac. Dis.*, vol. 7, no. 5, pp. 953–960, 2015.
- [6]. K. Grolinger, A. L'Heureux, M. A. M. Capretz, and L. Seewald, "Energy forecasting for event venues: Big data and prediction accuracy," *Energy Build.*, vol. 112, pp. 222–233, 2016.
- [7]. G. Singh, R. K. Panda, and M. Lamers, "Modeling of daily runoff from a small agricultural watershed using artificial neural network with resampling techniques," *J. Hydroinformatics*, vol. 17, no. 1, p. 56, 2015.
- [8]. W. N. W. M. Adnan, N. Y. Dahlan, and I. Musirin, "Modeling baseline electrical energy use of chiller system by artificial neural network," in *PECON 2016 - 2016 IEEE 6th International Conference on Power and Energy, Conference Proceeding*, 2017, pp. 500–505.
- [9]. M. Costa *et al.*, "New multi-objective algorithms for neural network training applied to genomic classification data," *Stud. Comput. Intell.*, vol. 201, pp. 63–82, 2009.
- [10]. T. S. Gunawan, I. Z. Yaacob, and M. Kartiwi, "Artificial Neural Network Based Fast Edge Detection Algorithm for MRI Medical Images," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 123–130, 2017.
- [11]. N. Tehlah, P. Kaewpradit, and I. M. Mujtaba, "Artificial neural network based modeling and optimization of refined palm oil process," *Neurocomputing*, vol. 216, pp. 489–501, Dec. 2016.

- [12]. V. C. Ivănescu, J. W. M. Bertrand, J. C. Fransoo, and J. P. C. Kleijnen, "Bootstrapping to solve the limited data problem in production control: an application in batch process industries," *J. Oper. Res. Soc.*, vol. 57, no. 1, pp. 2–9, 2006.
- [13]. G. Singh, R. K. Panda, and M. Lamers, "Bootstrap-based artificial neural network analysis for estimation of daily sediment yield from a small agricultural watershed," *Int. J. Hydrol. Sci. Technol.*, vol. 5, no. 4, p. 56, 2015.
- [14]. E. Ferrario, N. Pedroni, E. Zio, and F. Lopez-Caballero, "Bootstrapped Artificial Neural Networks for the seismic analysis of structural systems," *Struct. Saf.*, vol. 67, pp. 70–84, 2017.
- [15]. D. Karaboga and B. Akay, "A comparative study of Artificial Bee Colony algorithm," *Appl. Math. Comput.*, vol. 214, no. 1, pp. 108–132, 2009.
- [16]. J. C. Bansal, H. Sharma, and S. S. Jadon, "Artificial bee colony algorithm: a survey," *Int. J. Adv. Intell. Paradig.*, vol. 5, no. 1/2, p. 123, 2013.
- [17]. C. Ozturk and D. Karaboga, "Hybrid Artificial Bee Colony algorithm for neural network training," in *2011 IEEE Congress of Evolutionary Computation (CEC)*, 2011, pp. 84–88.
- [18]. J. Marais, M. Kleingeld, and J. Van Rensburg, "Challenges in the scaling of energy savings baselines on mine compressed air systems," in *2011 Proceedings of the 8th Conference on the Industrial and Commercial Use of Energy*, 2011, pp. 88–92.