Day-Ahead Load Forecasting using Support Vector Regression Machines

Lemuel Clark P. Velasco, Daisy Lou L. Polestico, Dominique Michelle M. Abella,
Genesis T. Alegata, Gabrielle C. Luna
Mindanao State University-Iligan Institute of Technology
Premier Research Institute of Science and Mathematics
Iligan City, The Philippines

***Abstract*—Accurate day-ahead load prediction plays a significant role to electric companies because decisions on power system generations depend on future behavior of loads. This paper presents a strategy for short-term load forecasting that utilizes support vector regression machines. Proper data preparation, model implementation and model validation methods were introduced in this study. The SVRM model being implemented is composed of specific features, parameters, data architecture and kernel to achieve accurate pattern discovery. The developed model was implemented into an electric load forecasting system using the java open source library called LibSVM. To confirm the effectiveness of the proposed model, the performance of the developed model is evaluated through the validation set of the study and compared to other published models. The created SVRM model produced the lowest Mean Average Percentage Error (MAPE) of 1.48% and was found to be a viable forecasting technique for a day-ahead electric load forecasting system.**

Keywords—Support vector regression machines; day-ahead load forecasting; energy analytics

I. INTRODUCTION

Accurate day-ahead load prediction demand plays a very significant role to electric companies because operation decisions in power systems such as unit commitment, contingency analysis, field scheduling, reducing spinning reserve, reliability analysis, load flow, and scheduling device maintenance depend on future behavior of loads [1]-[3]. It was pointed out that there is no substantial energy storage in the electric transmission and distribution system [1], [4]. Because of this existing limitation, utility companies invest greatly in load prediction to ensure that the basis of their operational decisions are reliable and in keeping the equilibrium between production and consumption since electricity cannot be stored. It should be stressed that inaccurate load predictions are costly to utility companies. When there is an underestimation in energy demand, it may result in limited supply of electricity at the consumer end, which leads to energy quality reduction in system reliability and there is an increase in operating costs when there is an overestimation in forecasting error [2], [3].

A computational intelligence method involving Support Vector Machines (SVM) has gained attention in the field of forecasting. SVM is a machine learning method based on statistical learning theory [1], [2], [5]. It was originally used for pattern recognition and classification but its importance increased when its algorithm was extended to solve regression problems, thus the nomenclature: Support Vector Regression Machine (SVRM). Only in 2001 on a European Network on Intelligent Technologies for Smart Adaptive Systems (EUNITE) competition in Taiwan was it realized that SVRM can also be used for load forecasting. Since then there have been several studies that explored different techniques to use or optimize SVRM for load forecasting of unique data sets [1], [2], [6], [7].

In the Philippines, a certain power utility is faced with a challenge to predict electric load using the historical data they accumulated. The power utility's current attempt involves using linear regression model and SVRM has not been utilized in their context. The historical load data of the power utility company can be used to develop a day ahead electric load prediction model. This study attempted to develop a day-ahead electric load prediction model using SVRM using electric load data provided by the power utility. Specifically, the study developed a data preparation model for load forecasting and implemented a SVRM model into an electric load forecasting system using LibSVM and Java. Measures were then conducted to quantify the efficiency and validate the accuracy of the developed SVRM prediction model. With this, SVRM was utilized to analyze and predict day-ahead load forecasting using the historical load data that a power utility company has accumulated with the aim to solve the long term problem of managing the supply and demand of the locality's power system.

II. METHODOLOGY

A. Load Data Preparation Methodology

There are many factors that can affect the success of machine learning, one of which is the quality of the data, so pre-processing of the data is essential [8]-[10]. In this study, a series of steps was conducted involving data description, data representation and feature selection. Data description discusses the time range of the data, and the description of its contents. Daily historical delivered load data was acquired from an electric utility which details were thoroughly recorded and examined. A description of the features present in the data was formulated as well as its structure along with other attributes including additional features. In data representation, the load data were represented, scaled and then partitioned. Binary was used represent non-numeric features, e.g. one (1) marks weekday which is a day type feature, while one-one-zero (110)

This work is supported by the Mindanao State University-Iligan Institute of Technology (MSU-IIT) as an internally funded research under the Premier Research Institute of Science and Mathematics (PRISM).

marks January which is a month type feature. The rest of the possible features that require representation used the same method of binary representation [7]. Scaling was also implemented to the load data since a study suggested that scaling is an important stage for training the SVRM in such a way that the higher values should not suppress the lower values in order to retain the activation function [11]. Min - Max scaling method as shown in (1) was then used as suggested by authors [3], [8], [11].

$$x' = \frac{\text{x-min(x)}}{\text{max(x)-min(x)}} \quad (1)$$

After representation and scaling, the load data was partitioned into two data sets: the training set and validation set where the training set was used for training the model while the validation set was used for testing the design of the model to confirm its predictive accuracy. Features which are data attributes in a specific data set that can affect the performance of machine learning were then filtered by feature selection. Feature selection, also known as variable selection or attribute selection, aims to identify the most relevant input variables within a data set and removing irrelevant, redundant, or noisy data [9], [10], [12]. It is defined as the process of selecting D most discriminatory features of d > D available ones [13]. Proper selection of features can improve the prediction performance and provide a better understanding of the underlying process that generated the data [9], [10], [12], [13].

The expected output of feature selection is a list of relevant features that is used in making the predictive model. Given a subset of the whole data set, various feature selection algorithms were tested for performance. According to a study there are two broad categories in feature selection: the wrappers and the filters [9]. The study proved that filters are more practical than wrappers because they are much faster in large databases while supporting correlation-based approach. A study conducted suggested using the same approach through the Baye's function method while a separate study proposed another feature selection method based on information gain approach [13]. Thus in this study, correlation-based filter feature selection was used implemented in R programming and Weka software open source environments. After identifying the correct features the data in an .xls format were converted into a libsvm format using R programming.

To eliminate human opinion on selecting the appropriate attributes for the electric load prediction, this research used correlation-based feature selection, namely, Pearson's correlation, Spearman correlation matrix, and Kendall's correlation to improve the accuracy of machine learning. The value of a correlation coefficient ranges between -1 and 1. The strongest linear relationship is indicated by a correlation coefficient of -1 or 1 while the weakest linear relationship is indicated by a correlation coefficient equal to 0. A positive correlation means that if one variable gets bigger, the other variable tends to get bigger. A negative correlation means that if one variable gets bigger, the other variable tends to get smaller. This study used R programming language to implement the three correlation based approach. To implement these three correlation based approaches, the function cor() of R was used. The cor() function calculates the weighted correlation of the given data set. The format used for this function is shown in (2) where x is the matrix or the data set, use specifies the handling of missing data with options: .obs, complete.obs, pairwise.complete.obs and method to specify the type of correlation: pearson, spearman or kendall.

$$cor(x, use=, method =) \quad (2)$$

B. SVRM Model Implementation Methodology

The researchers used LIBSVM's Java library, an open source integrated software for support vector classification, regression and distribution estimation. It was developed by the National Taiwan University and written in C++ through a C API [14]. It has many bindings for programming languages such as Java, MATLAB and Python. Creating the model, selecting the kernel and the parameters were done through the use of LIBSVM functions such as svm-train() and svm-predict().A software was designed and developed to implement a SVRM model as well as carry out the process needed for the load prediction starting from data loading to the generation of the predicted values. The process started with the development of the module for the data loading and scaling so that it can be loaded into the system. Then the researchers proceeded in implementing a SVRM model into the system wherein the already prepared data was loaded and will then output the predicted load values in table form. System reports such as visualization for comparing the actual and forecasted load, exporting the results to a spreadsheet file, and printing the predicted load values were also included in the system adding to its functionality. After all the development has been done, the researchers performed various tests to the software to check for bugs and errors to assess the software's quality and usability and to ensure that it will produce accurate results. Once the testing has been performed and all the quality checks were done, the system was packaged ready for deployment.

The SVRM model was identified to consist of Radial Basis Function (RBF) kernel with parameters of c=125, g=0.001, e =0.01 and p=0.0045 to train the SVRM model. The model's architecture has day load consumption attributes in 15-minute resolution from each of the following: 1 day before, 2 days before, 7 days before, and 14 days before. This architecture is denoted *as i-1, i-2, i-7, i-14* where *i* represents the day to be predicted and the number after the subtraction sign represents the number of days before the predicted data. After implementing the SVRM model, the predictive accuracy of the model was validated by comparing the resulting forecasted electric load values to the actual delivered load values performed in ten repetitions of day-ahead load forecast. To assess the predictive accuracy of the SVRM model, Mean Absolute Percentage Error (MAPE) as suggested by researchers [6], [7], [11]. Since smaller MAPE values would indicate consistency and accuracy, daily MAPE were computed to test the day-to-day accuracy of the model along with the average MAPE to evaluate the performance of the model in a monthly scale. The resulting values were compared to the tolerance error and standard set by the local power utility. A tabular and graphical representation of the computations was then generated with the purpose of illustrating the comparison between the actual and predicted electric load values.

III. Results and Discussion

A. Load Data Preparation Results

Electric load data from January 2013 to December 2014 were stored in spreadsheets with .xls file format. Each file represents the electric load data for a single month and contains three sheets corresponding to the three metering substations of the power utility. As shown in Table I, each sheet contains the metering point (SEIL), the billing date (BDATE), time in 15 minute resolution, kilowatt delivered (KW_DEL), kilowatt per hour delivered (KWH_DEL), and kilo volt amps reactive hours delivered (KVARH_DEL). KW_DEL is the load to maintain and the basis for load prediction while KWH_DEL is for billing consumption and kilo volt amps reactive hours delivered (KVARH_DEL) represents the reactive power. The data used as input for the SVRM models is the attribute kilowatt delivered (KW_DEL) for the reason that it is also the column considered by power utilities in determining forecasted load values for the next day. A total of 70,109 rows of data was then fed into the SVRM model.

TABLE I. Sample Raw Data Format

SEIL	BDATE	TIME	KW_DEL	KWH_DEL	KVARH_DEL
XX	12/26/11	00:15	XXXX	XXXX	XXXX
…	…	…	…	…	…
XX	12/26/11	24:00	XXXX	XXXX	XXXXX

This study added attributes such as calendar day, holiday, day type, and day to the original data set since a study strongly suggest that these attributes will affect the electric load [2]. These attributes will then be represented by binary variables as suggested by authors [7], [15]. Table II shows the binary equivalent for the said attributes. The time attribute in 15-minute resolution was converted into numerical values since LIBSVM does not accept variables represented with a semicolon. Starting at 00:15, which is 12:15 AM, time was given a representation value of 1. And for every increment of 15 minutes, the representation value is incremented by 1. This process iterates until 00:15-24:00 is converted from 1 to 96. Part of data representation is data scaling which is expected to aid in a more accurate SVRM model. In this research, the predictive variable which is the KW_DEL has been scaled using Min - Max with a scale from zero to one. After data was scaled, the data was partitioned into training data set and validation data set. January 1, 2013 to November 2014 was set as training dataset while December 2014 was set as validation and testing dataset. The available dataset of December 2014 starts from December 1, 2014 to December 25, 2014. Partitioning the electric load data is necessary to group them according to their use. The training set which should comprise the largest part of the datasets was used to train the models with different parameters and the validation set was used to check the accuracy of the trained models [3], [8], [16].

TABLE II. Attribute-Binary Table

Day	Binary Value	Day	Binary Value	Day Type	Binary Value	Date Type	Binary Value
Monday	1000000	Friday	1111100	Week - Day	1	Holiday	1
Tuesday	1100000	Saturday	1111110	Week - end	0	Non-holiday	0
Wednesday	1110000	Sunday	1111111				
Thursday	1111000						

To improve the accuracy and effectiveness of the model, this study performed feature selection. The effect of not undergoing feature selection is that researchers will have to do more unnecessary data training with and without certain features to determine whether or not certain attributes have effects on the accuracy of the prediction [2], [5]. To show that feature selection has a significant effect on prediction accuracy, this research ran a sample model with parameters c=125 g=0.001 e=0.01 p=0.0045, kernel Radial Basis Function along with the model's architecture with two data sets. The first data set was filtered by feature selection and the second data set was one that did not undergo feature selection. Table III shows that the SVRM model with a data set that underwent feature selection with a MAPE value of 4.09% performs better than the model without feature selection with 6.77% MAPE value. The MAPE value generated in the two datasets is the average of the daily MAPE values of the validation set used in this study. This study confirms that feature selection does increase accuracy.

TABLE III. Model with Feature Selection vs. Model Without Feature Selection

Model	MAPE Values
Model with Feature Selection	4.09%
Model without Feature Selection	6.77%

Table IV summarizes the correlation of KW_DEL to time, month, year, day type, date type and day to systematically choose features for the implemented SVRM model. Time attribute has a 0.52901261, 0.56725912 and 0.39739378 respectively as correlation values to Pearson's, Spearman's and Kendall's tests. It is shown that the time attribute has the highest correlation to KW_DEL while month, year, day type, day, and date type have a relatively low correlation to KW_DEL. Performing the three types of correlation based approach resulted to the selection of time attribute as the only attribute that can affect the predictive variable KW_DEL. All three approaches have shown that time has the highest correlation to KW_DEL and thus a valid candidate as a feature for the SVRM model while the rest of the attributes show a low correlation to KW_DEL.

TABLE IV. SUMMARY OF THE CORRELATION BETWEEN KW_DEL AND OTHER ATTRIBUTES

	Pearson	Spearman	Kendall
Time	0.52901261	0.56725912	0.39739378
Month	-0.270584464	-0.2463835254	-0.1805927803
Year	-0.0675271275	-0.0749241004	-0.0615430231
Day type	-0.028027669	-0.0489527428	-0.0402100228
Date type	-0.021399675	-0.027798819	-0.022834086
Day	-0.0183968342	-0.0434596375	-0.0312288852

B. SVRM Model Implementation Results

The developed software allows the user to set the parameters in training the SVRM model such as the cost value of C, the epsilon function, width of epsilon function and the gamma function for the radial basis function kernel. In addition, the user can also select which kernel to use for training the SVRM. As shown in Fig. 1, these options will add flexibility for the users and will enable the support vector regression machines to be retrained in the future if there will be possible significant changes on the behavior of the daily load consumption.

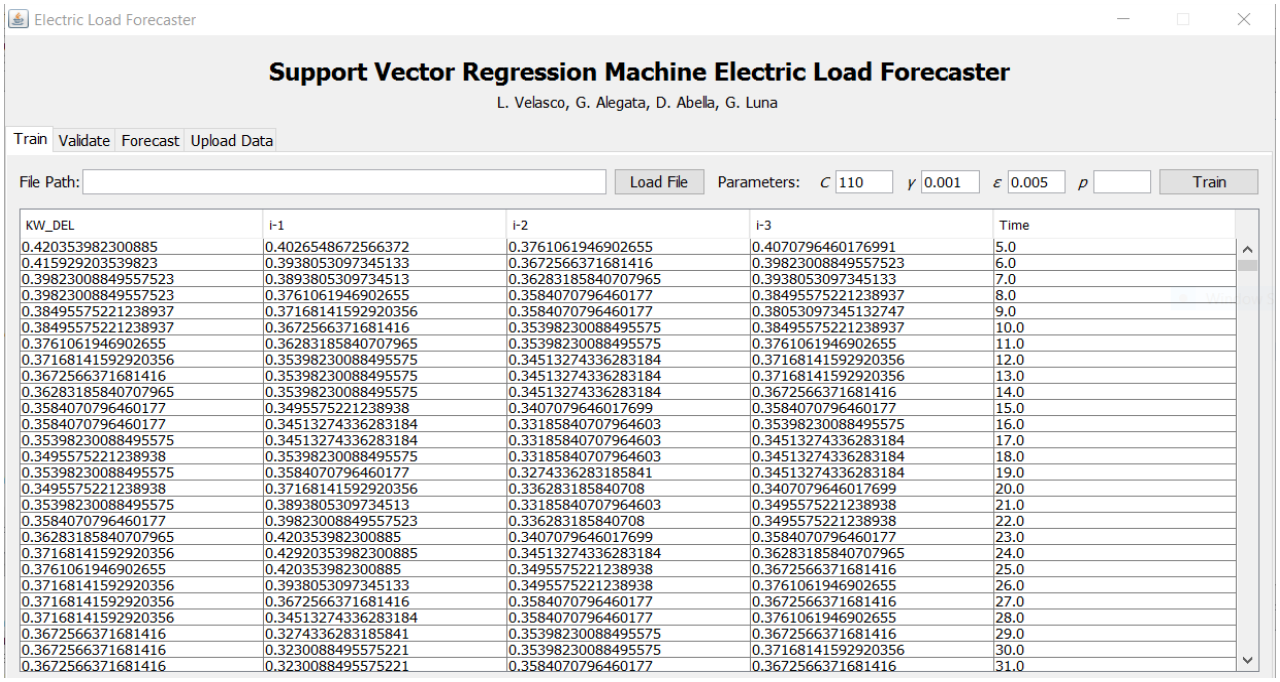

Fig. 1. The input interface of the load prediction software.

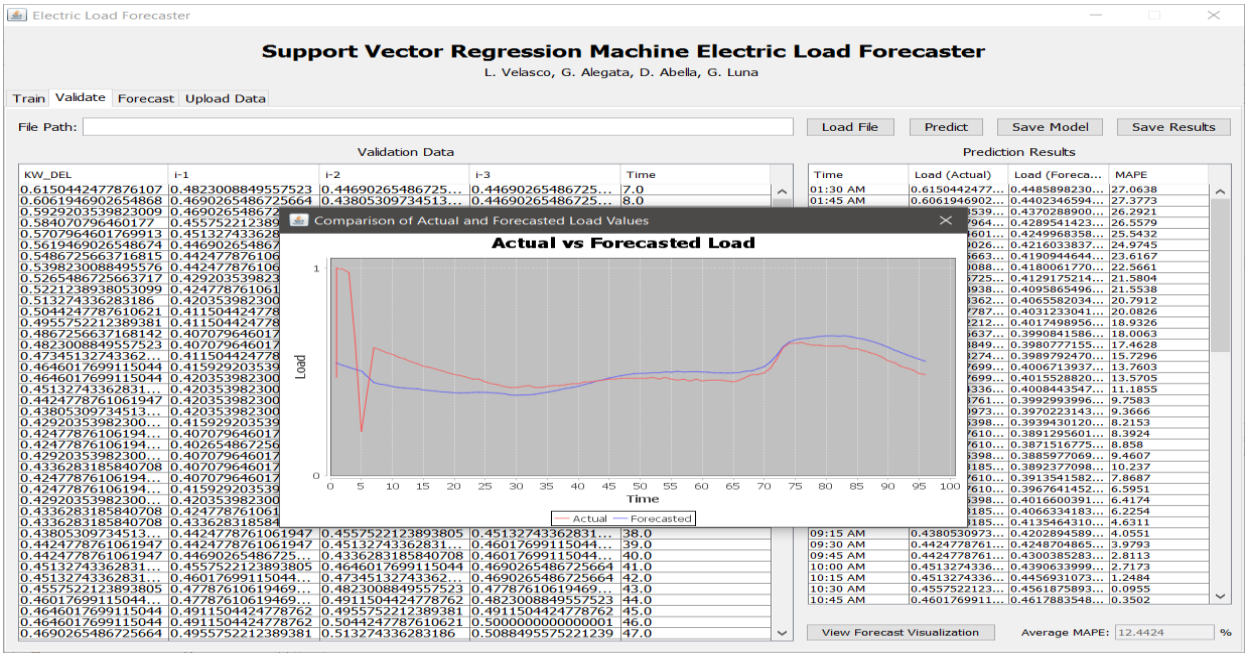

Fig. 2. The output interface of the load prediction software.

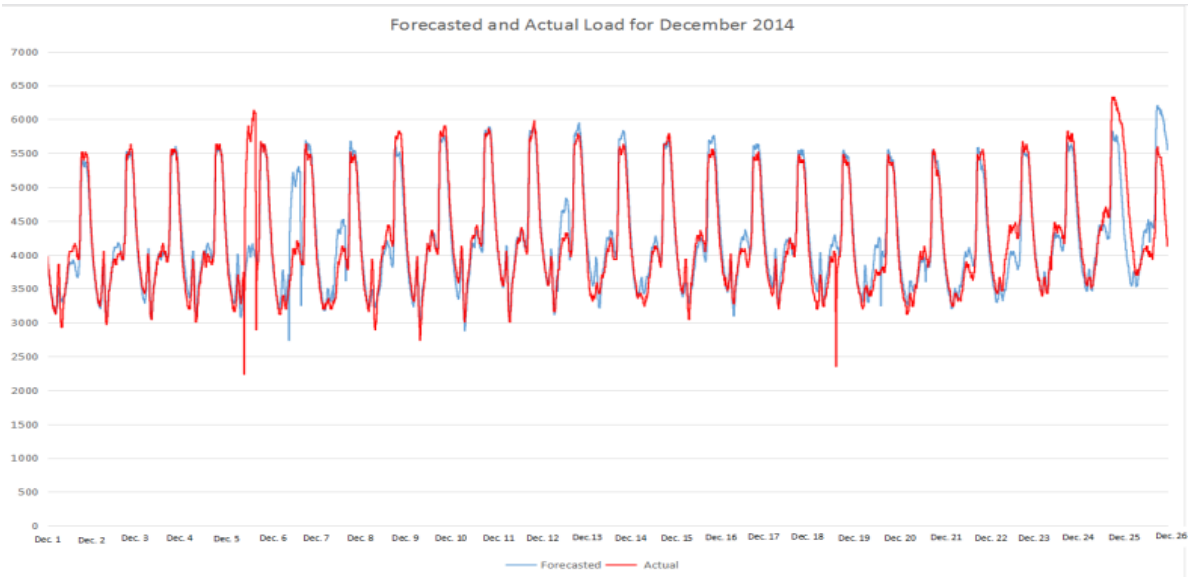

Fig. 3. Actual load vs. predicted load.

The model was trained based on the architecture using the function called svmTrain(). After training the data, the developed system automatically validated the trained data and forecasted results. A table and a graph were generated to compare the actual values and the predicted values as shown in Fig. 2.

The model achieved an average of 4.03% for the 25 days of December while the lowest MAPE achieved for a day was 1.59% achieved in the prediction for Dec. 11, 2014 load. As shown in Fig. 3, the highest MAPE was 11.29% taken on December 25, 2014.

The developed SVRM model, taking past electric load data and time as the only attributes, was not optimized enough to adjust for KW_DEL prediction for Christmas day since holiday and special events were not considered in training the model. If Christmas day was excluded in validation set, the average MAPE for the 24 remaining days of December 2014 would be 3.73%, which is significantly lower, instead of 4.03%. Table V shows a portion of MAPE values of the predicted days for one week. While the implemented SVRM model produces a MAPE of 4.03%, a study produced an SVRM model with MAPE of 3.67% [2]. But it is worth noting that the study used attributes day, date type, electricity price, and daily temperature. Another study which also used the same architecture produced a MAPE of 2.86% and used only the attributes day and date while this study only used attributes of past KW_DEL and time [6]. Also, comparing with other load forecasting studies which does not use SVRM showed that the model generated in this study performs at par, if not, better with other models created using different methods. This is a study which used ANN in load forecasting for a Nigerian electrical power system and yielded an average error of 2.54% and lowest error in a day of 1.73% [17]. The study used the load values of the previous hour, day and week, the day of the week and the hour of the day as network inputs. While another study on day-ahead load forecasting which utilized ANN on this study's datasets only yielded 2.40% and 2.80% as the lowest and highest MAPE respectively on a week's forecast [16].

TABLE V. SUMMARY OF THE CORRELATION BETWEEN KW_DEL AND OTHER ATTRIBUTES

DATE	MAPE	DATE	MAPE
Dec-09	2.54%	Dec-13	3.44%
Dec-10	3.68%	Dec-14	2.01%
Dec-11	1.58%	Dec-15	2.57%
Dec-12	3.44%		

Moreover, validation was also done by comparing the performance of the SVRM model with local forecasting standards. Having a lowest error of 1.59% places the prediction accuracy of the model well under the 5% acceptable tolerance rate of the locality's electricity spot market and according to a study, 2%-3% error range for a day-ahead prediction is considered normal and will be considered as reliable [18]. Thus, the forecasting performance of the SVRM model is within the acceptable error and considered as accurate and reliable. Below illustrates the results of the developed model which confirms the ability of the model to forecast a day-ahead load within the acceptable error.

IV. CONCLUSION AND RECOMMENDATIONS

This paper proposed a short term electric load forecasting strategy using SVRM by executing data preparation and by implementing an SVRM model in LIBSVM. An application software with features to scale, train, validate, forecast and visualize results of the data was developed using Java. Data preparation which is composed of data representation and correlation-based feature selection paved the way for the implementation of an SVRM model with Radial Basis Function kernel. Having parameters of C = 110, g = 0.001, e = 0.01 and p = 0.005 implemented in an architecture, the implemented model produced the lowest MAPE of 1.48% in day ahead load prediction and an average of 4.44% for the prediction in December 2014.

Based on the findings of the study, the researchers would like to recommend for further studies in utilizing SVRM to

expand the pool of knowledge on electric load forecasting using. It is recommended to explore different methods in selecting kernels as this will help in establishing a reliable SVRM process in electric load forecasting. Studies focused on ensuring that kernels would fit a given electric load data can be conducted as this will increase potential in the accuracy of prediction using SVRM models. The researchers also recommend for the conduct of performing grid-search for parameters and an automated system that performs parameter selection that are suited and optimized for SVRM as this will also help in improving the accuracy of prediction.

This study aims to contribute to the technologies of developing an electric load forecasting model and an electric load forecasting software that could aid power utility companies in their decision-making, electric load planning and load power utilization. The results generated and obtained in this study clearly suggests that with proper data representation, feature selection, and SVRM model implementation, SVRM is a viable forecasting technique for a day - ahead electric load forecasting system.

ACKNOWLEDGMENT

The authors would like to thank the support of the MSU-IIT Office of the Vice Chancellor for Research and Extension and PRISM-Premiere Research Institute in Sciences and Mathematics for their assistance in this study.

REFERENCES

[1] E. E. Elattar, J. Goulermas, and Q. H. Wu, Electric load forecasting based on locally weighted support vector regression. In Proc. IEEE Transactions on Systems, Man and Cybernetics C, Vol. 40, No. 4, pp. 438–447 (2010).

[2] B. E. Türkay, and D. Demren, Electrical load forecasting using support vector machines. In Proc. 7th International Conference on Electrical and Electronics Engineering (ELECO), pp. I-49-I-53, (Bursa, 2011).

[3] L. C. P. Velasco, P. B. Bokingkito, and J. T. Vistal, Week-Ahead Load Forecasting using Multilayer Perceptron Neural Network for a Power Utility. In Proc. 17th Conference of the Science Council of Asia, (Manila, 2017)

[4] E. Almeshaiei, and H. Soltan, A methodology for electric power load forecasting, Alexandria Engineering Journal, 50(2), pp. 137-144 (2011)

[5] B. Schölkopf, and A. J. Smola, A. J. (2002), Support vector machines and kernel algorithms. The Handbook of Brain Theory and Neural Networks, (MIT Press, 2002).

[6] S. Ostojin, F. Kulić, G. Švenda, and R. Bibić, Short-term electrical load forecasting using support vector machines. Computers and simulation in Modern Science Vol I; Mathematics and Computers in Science Engineering, (WSEAS Press, A Series of Reference Books and Textbooks, 2008)

[7] E. Ceperic, V. Ceperic, and A. Baric, A strategy for short-term load forecasting by support vector regression machines. In Proc. IEEE Transactions on Power Systems, Vol. 28, No. 4, pp. 4356–4364, (2013).

[8] L. C. P. Velasco, P. N. C. Palahang, C. R. Villezas, and J. A. A. Dagaang, Performance Analysis of Different Combination of Training Algorithms and Activation Functions in Predicting the Next Day Electric Load. In Proc. of 16th Philippine Computing Science Congress 2016, Computing Society of the Philippines, (Puerto Princesa, 2016)

[9] M. A. Hall, L. A. Smith, Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In Proc. 12th Int'l Florida Artificial Intelligence Research Society Conference, pp. 235-239, (Florida, 1999).

[10] V. Kumar and S. Minz, Feature Selection: A Literature Review. Smart Computing Review, Korea Academia-Industrial Cooperation Society, Vol. 4, No. 3, pp. 211-229, (2014).

[11] A. Jain and B. Satish, Clustering based Short Term Load Forecasting using Support Vector Machines. In Proc. 2009 IEEE Bucharest PowerTech, pp. 1-8. (Bucharest, 2009)

[12] M. Sarhani, A. El Afia, Electric load forecasting using hybrid machine learning approach incorporating feature selection. In Proc. International Conference on Big Data Cloud and Applications pp. 1-7, (2015).

[13] M. Haindl, P. Somol, D. Ververidis and C. Kotropoulos, Feature Selection Based on Mutual Correlation. Progress in Pattern Recognition, Image Analysis and Applications (Springer, 2006).

[14] C. Chang and C. Lin, LIBSVM - A Library for Support Vector Machines, Retrieved April 17, 2017, from National Taiwan University: https://www.csie.ntu.edu.tw/~cjlin/libsvm, (2011).

[15] J. M. Espinoza, J. Suykens, R. Belmans, and B. D. Moor, Electric load forecasting, In Proc. IEEE Control Systems, 27(5), pp. 43–57 (2007).

[16] L. C. P. Velasco, C. R. Villezas, P. N. C. Palahang, and J. A. A. Dagaang, Next day electric load forecasting using Artificial Neural Networks. In Proc. of the IEEE International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), (Cebu, 2015).

[17] G. A. Adepoju, S. O. A. Ogunjuyigbe, and K. O. Alawode, Application of Neural Network to Load Forecasting in Nigerian Electrical Power System. The Pacific Journal of Science and Technology, 8(1), pp. 68-72, (2007).

[18] H. Soliman , K. Sudan, and A. Mishra, A smart forest-fire early detection sensory system: another approach of utilizing wireless sensor and neural networks, IEEE Sensors, pp.1900-1904, (2010).