Comparative Study of Short-term Electric Load Forecasting: Case Study EVNHCMC

Nguyen Tuan Dung, Tran Thu Ha, Nguyen Thanh Phuong

Abstract - ***Short-term load forecasting (STLF) plays an important role in building business strategies and ensuring reliability and safe operation of any electric power system. There are many different methods used for short-term forecasting, including regression models, time series, neural networks, expert systems, fuzzy logic, machine learning, and statistical algorithms. There are always debates about which algorithms are the best for electric load forecasting. In this paper, we compared the SVR (Support Vector Regression), NN (Neural Network) and RFR (Random Forest Regression) algorithms, based on the dataset of EVNHCMC to find out a suitable STLF for the dataset.***

Keywords: ***STLF; SVR; NN; RFR; Standardized Load Profile.***

I. INTRODUCTION

Load forecasting of power systems is a topic that has been studied over a long period of time, based on the historical data of the load and its influencing factors for constructing the regression curve by using mathematical algorithms [1]. The main issue is how to take advantage of existing historical data to establish predictive models and forecast load values for a given time period in the future. Although the short-term forecasting task has been studied for several decades with different provided models, short-term load forecasting is still a difficult task. The reason is that the accuracy requirements are becoming increasingly high because of the impact of renewable energy development, smart grid and energy efficiency issues, such as the Ho Chi Minh City area.

Short-term load forecasting is usually the forecast of the load for the next day or next week with a cycle of 60/30 minutes. Days are classified in different types, such as: normal day (Tuesday to Friday), Monday, Saturday, Sunday, New Year (solar calendar and lunar calendar) and holidays. Each day type has its own morphological characteristics, therefore the predictive model is set up to reflect these characteristics.

The load profile contains latent factors, which tend to be similar to the previous load model, however, this leads to false predictions of subsequent days if the date is different from the previous day or there is an event that impacts. In addition, seasonal changes, meteorological trends (such as the El Niño, La Niña Phenomenon) are also carefully considered. Therefore, the load, classifying according to its inherent characteristics, is inevitable.

Nguyen Tuan Dung is with Ho Chi Minh Power Company, Vietnam, correspondings author email: dp1526@gmail.com

Tran Thu Ha is with HCMC University of Technology and Education, email: thuha@hcmute.edu.vn

Nguyen Thanh Phuong Ho Chi Minh City University of Technology, email: nt.phuong@hutech.edu.vn

Classical linear methods have been used to predict electrical load forecasting including moving average (MA), exponential smoothing and linear regression [3] - [7]. Today, data mining methods such as the artificial neural network and SVM (Support Vector Machine) [8] - [11] show good results. Therefore, this paper compares the results of some short-term load forecasting methods and proposes data preprocessing procedures and classifying load characteristics.

II. METHODOLOGY

A. Load Characteristics Analysis

Load data of the Electricity Corporation of Vietnam – Ho Chi Minh City (EVNHCMC) from January 01, 2018 to March 08, 2018 is used for constructing the short-term load forecasting algorithm. They include: Power capacity (Pmax, Pmin); Power Consumption (A); Capacity of 60 minute cycle; Temperature (0^oC) of 60 minute cycle; List of holidays in a year.

Observing the load profile of the 7 days of a week in Ho Chi Minh City (Figure 1), we see that the variation between weekdays (from Monday to Friday) is not very much and they have the same load profile type. For the load profile of Saturday, there are changes but to a low degree / but not many, compared to weekdays, mainly the load demand decreases in the evening. Separately, the load profile of a/the Sunday is completely different from the normal days (the load demand is low).

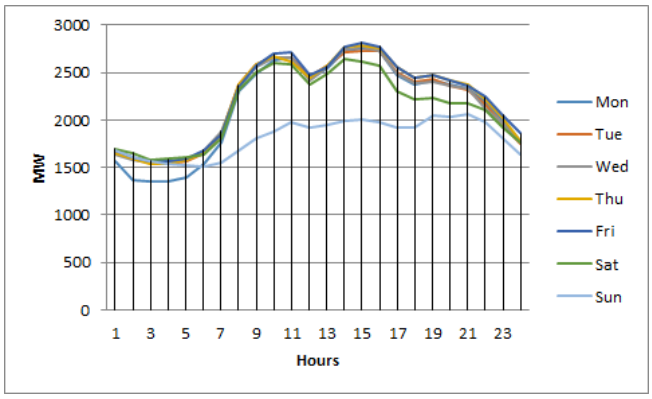


Figure 1. *Load Profiles 07 days*

Load changes from hour to hour of a day. The figure 2 below shows the typical load profile of January. It can be seen that peak loads occur twice a day, one is from 10:00 AM to 12:00 AM and the other is from 2:00 PM to 4:00 PM. And because of this load demand behavior, an hourly feature can be very useful in short-term load forecasting. This is the basis for the standardized load profile.

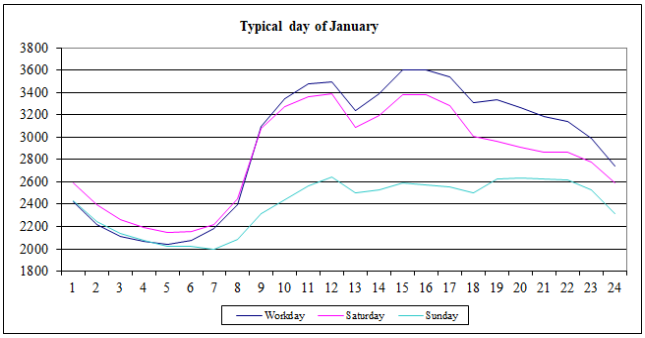


Figure 2. *Load Profiles typical day of January*

Weather variables have a significant impact on load demand. Many weather variables could be considered for load forecasting, such as temperature, wind speed, cloud cover, humidity, but their effects on load demand are different. According to many authors, temperature is a commonly used weather variable in load forecasting. However, it should only be considered with caution because the temperature load forecast is a much more complex issue than the load forecast, and the estimated temperature is used for the input instead of the actual one. However, temperature is not always a good choice for forecasting inputs, which can sometimes reduce the accuracy of forecasting results.

In addition, for Asian countries using the lunar calendar, one of the most unsettling issues is holidays, the Lunar New Year (usually in the last days of January or early February). The forecasting results of this period often have a big error, due to the deviation of the load cycle between solar calendar days and lunar days

B. Support vector regression (SVR)

The feature of SVR is that it provides us with a sparse solution. That is, to build the regression function, we do not need to use all the data points in the training set. The points that contribute to the construction of the regression are called Support Vectors. The layering for a new data point will depend only on the support vectors [12].

The regression function has the formula:

$$y = f(x) = w^T \Phi(x) + b \quad (1)$$

Thus, the goal of SVR training is to find w and b [13] - [15]. For the training set {(x1, t1), (x2, t2), …, (xN, tN)} $\subset R^n \times R$. With a simple regression problem, to find w and b we have to minimize the normalized error function:

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|w\|^2 \quad \text{where } \lambda \text{ is a normalized constant} \quad (2)$$

To get a sparse solution, we will replace the above error function with the ε-insensitive error function. The characteristic of this error function is that if the absolute value of the difference between the predicted value y(x) and the target value is less than ε (with ε > 0) then it is considered the error is zero.

Now, we must minimize the normalized error function:

$$C\sum_{n=1}^{N} E_\varepsilon(y(x_n) - t_n)^2 + \frac{1}{2}\|w\|^2 \quad (3)$$

With $y(x_n) = w^T \Phi(x_n) + b$, C is a normalized constant like λ but is multiplied by an error function instead of $\|w\|^2$.

Using the Lagrange function and the Karush-Kuhn-Tucker condition, we have the equivalent optimization problem:

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \hat{a}_n)(a_m - \hat{a}_m)k(x_n, x_m) - \varepsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)t_n \quad (4)$$

Where k is the kernel function: $k(x, x') = \Phi(x)^T \Phi(x')$. Maximizing with constraints:

- $0 \le a_n \le C$
- $0 \le \hat{a}_n \le C$ (5)
- $\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$

From here, we have the regression function of SVR:

$$y(x) = \sum_{n=1}^{N}(a_n - \hat{a}_n)k(x, x_n) + b \quad (6)$$

Thus, for SVRs using the ε-insensitive error function and the Gaussian kernel function we obtain three parameters: the normalization coefficient C, the parameter γ of the Gaussian kernel function, and the width of the pipe ε [13]. All three parameters affect the prediction accuracy of the model and need to be selected carefully. If C is too large, it will give a priority to the training error, leading to a complex model, it is easy to be overfitting. If C is too small, it gives a priority to the complexity of the model, leading to a too simple model, reducing the prediction accuracy. The meaning of ε is the same. If ε is too large, there are less support vectors, making the model too simple. On the other hand, if ε is too small, there are many supporting vectors, leading to complex models, which are more likely to be overfitting. The γ parameter reflects the correlation between the support vectors and also affects the prediction accuracy of the model.

C. Random Forest Regression (RFR):

Random forest (RFs) is a machine learning method for both classification and regression [16] – [17]. RF is a set of decision trees that grow in randomly selected subspaces of the characteristic space. A single tree in RF is constructed on a subset of the training points and a subset of the features is considered at each node to divide. Furthermore, the trees in the forest are planted to the maximum size and the pruning step is ignored. The classification or regression function is the simple majority voting method or the simple average method.

The general failure of a random forest depends on the accuracy of each member tree of the forest and on the interdependence of the member trees. The random forest algorithm uses unpruned trees to keep the bias component low and uses the randomness to control the low correlation between trees in the forest.

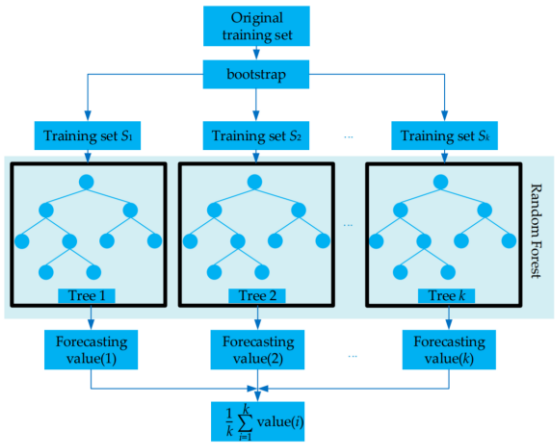


Figure 3. *Random Forest Regression*

Algorithm of RF can be briefly described as follows:

- From the LS data set we have m elements and n variables (feature variables), construct T decision trees independently.
- The t decision tree model is built on the t Bootstrap sample (sampling the returned m elements from the LS set).
- At the inner node, select the random n' variable (n' << n) and compute the best partition based on n' of this variable.
- The tree is built to maximum depth without pruning.

Finishing the process of constructing the T basic model, using the majority voting strategy to classify a new element or the mean value for the regression problem.

D. Research models

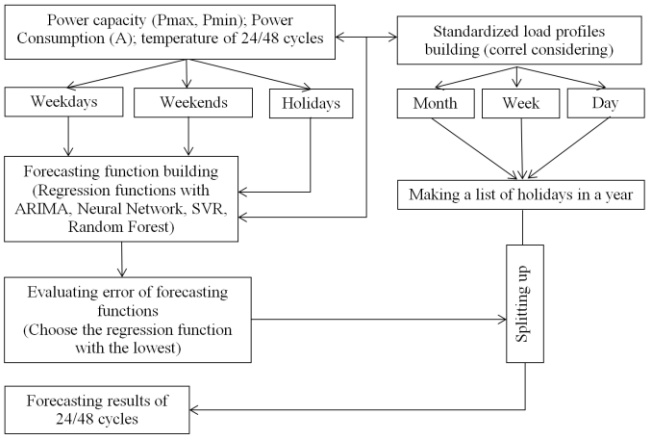


Figure 4. *STLF procedure*

The standardized load profile is constructed by taking the power value collected in the 60-minute cycle divided by its maximum capacity. Through observation, the standardized load profile shows all the load characteristics of each time, season, holidays and Tet. At the same time, the standardized load profile is also an important input parameter for the training of machine learning algorithms such as SVR, NN, RFR, to minimize the influence of temperature variables.

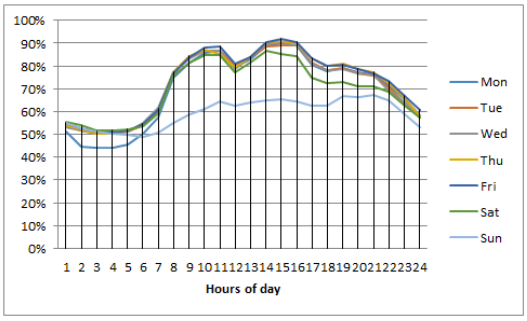


Figure 5. *Standardized Load Profile*

III. RESULTS AND DISCUSSION

The paper uses the data set from January 1, 2014 to March 8, 2018 of EVNHCMC to run the models. After preprocessing, the data set is divided into two sets: the training set and the test set, in which the test set is the last 30 days of the data set. Or split the data set into phases to test the forecast results in different phases and time.

Input data for algorithm training include: power (Pmax/Pmin) in 60 minutes cycle; temperature (max/min) in 60 minutes cycles; standardized load profile of 24 hours a day; list of holidays and Tet in the forecasting year.

To solve the cyclical difference between the solar calendar and the lunar calendar of our country, we use an algorithm to converse a solar day to a lunar day and vice versa. On the basis of coordinating solar days and lunar days, we construct the standardized load profile corresponding to the special days of the lunar days (such as Lunar New Year, Hung Kings Commemoration Day - 10th day of the 3rd lunar month) and special days of solar days (such as New Year, Reunification Day April 30, International Workers' Day May 1, National Day of Vietnam - September 2).

A useful measurement parameter is the mean absolute percentage error (MAPE), which only makes sense if all values are positive. This method is used to evaluate the models.

$$MAPE = \frac{1}{n}\sum 100\frac{\left|Y_t - Y_t^f\right|}{Y_t} \quad (7)$$

The algorithms are programmed in the Mathlab language and the results are exported to Excel for exploiting the forecasting data.

B. SVR Models

It is important to select the correct input parameters to run SVR models: normalized coefficient C, pipe widths ε and Gaussian kernel function. The algorithm uses the same set of input data of the models. Some featured SVR model parameters are proposed:

Model	C - BoxConstraint	ε - Epsilon	KernelFunction
SVR 1	93.421	32.502	Polynomial
SVR 2	449.836	44.983	Gaussian
SVR 3	1	44.983	Linear
SVR 4	100	0.01	Gaussian

C. RFR Models

The Tree Bagger function is used for model building, which is a set of regression trees with different sets of rules for performing non-linear regression. The algorithm builds a total of 20 trees, with a minimum leaf size of 20. The size of the leaf is less than or equal to the tree size to control overfitting phenomena and get high efficiency. The algorithm uses the same set of input data of the models.

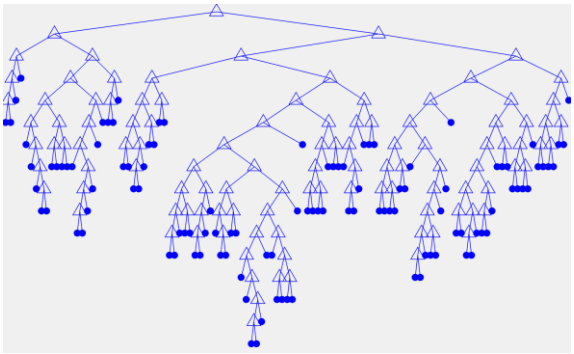


Figure 6. *Random Forest Regression Model (in mathlab)*

D. Model NN

We use the Feedforward Neural Network model with the above input variables and the training data set as above. Network structure of 01 hidden layer with a class size of 10 and Sigmoid trigger function. At the same time, using the Neural network with the network structure of 03 hidden layers, where: the first hidden layer size is 10; The second hidden layer size is 8 and the third hidden layer size is 5;

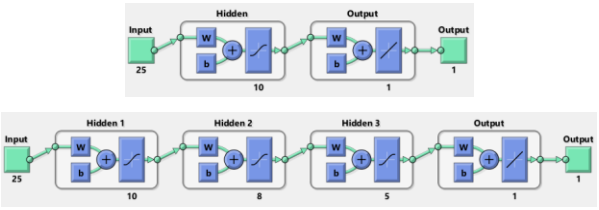


Figure 6. *Neural Network Models*

E. Results and Discussion

- Forecasting results of 07 days in October (no special event):
  - SVR Models:

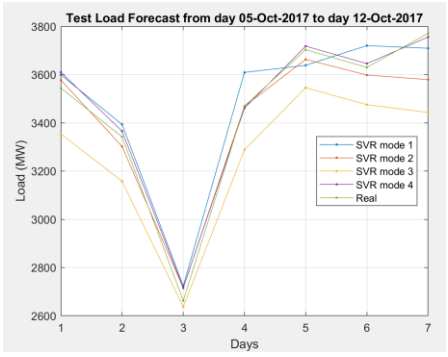


Figure 7. *Result of SVR models*

MAPE SVR1	MAPE SVR2	MAPE SVR3	MAPE SVR4
2,18%	1,60%	4,90%	0,87%

From the results of running the 04 SVR models we can see: corresponding to the settings for the SVR models, model number 3 has a large MAPE error (4.9%). The test result for model 4 has the lowest error (0.87%) and the regression curve is quite consistent with the actual data line.

- NN and RFR models:

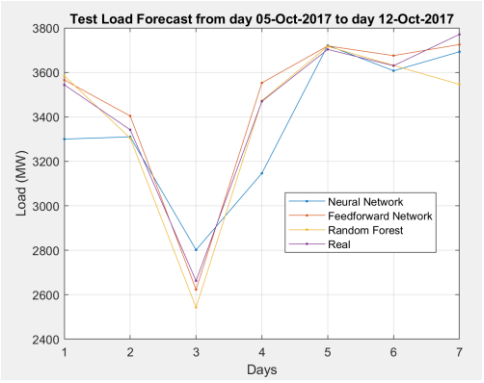


Figure 8. *Result of NN + RFR models*

MAPE NN	MAPE FDN	MAPE RFR
1,32%	3,65%	1,89%

Based on the results of running the NN and RFR models, it can be seen that the regression curves of the NN and RFR models yielded a MAPE error as well (<1.89%), the simulated curves close to with actual load curve.

Comparison of SVR, NN, RFR simulation results shows that the SVR4 model has the best results, with the smallest error. Therefore, the SVR4 model will be selected as the forecasting model for the next days. Accordingly, the result is:

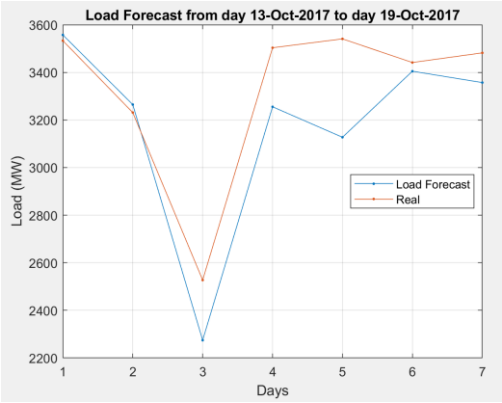


Figure 9. *Result of SVR models*

The results of the SVR4 forecasting model for the days from October 13, 2017 to October 19, 2017 have the lowest error (0.8%). Forecasting results for 24 cycles of 07 days per week from October 13, 2017 to October 19, 2017 as follows:

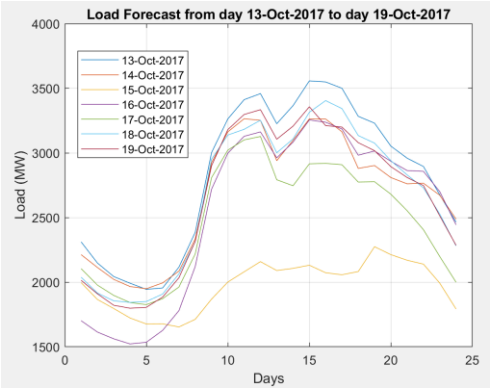


Figure 10. *Result for 24 cycles of 07 days*

- The results of the month in Lunar New Year (from February 7, 2018 to March 8, 2018): run the pilot model to find the model with the lowest error, then run the forecasting result

  - SVR Models:

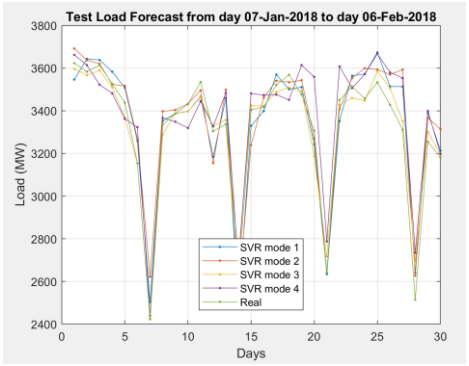


Figure 11. *Result of SVR models*

MAPE SVR1	MAPE SVR2	MAPE SVR3	MAPE SVR4
2,36%	2,58%	1,36%	3,07%

From the results of running the 04 models SVR we can see: corresponding to the settings for the SVR models, model number 4 has a large MAPE error (3.07%). The test results for model 3 has the lowest error (0.87%) and the regression curve is quite consistent with the actual data line.

  - NN and RFR Models:

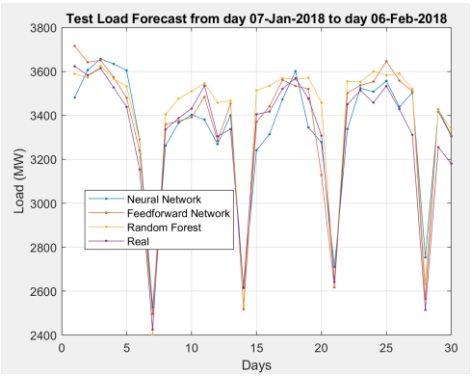


Figure 12. *Result of NN + RFR models*

MAPE NN	MAPE FDN	MAPE RFR
2,19%	2,69%	2,68%

Based on the results of running the NN and RFR models, it can be seen that the regression curves of the NN and RFR models yielded a MAPE error as well (<2.68%), the simulated curves close to with actual load curve.

Comparison of SVR, NN, RFR simulation results shows that the SVR4 model has the best results, with the smallest error. Therefore, the SVR4 model will be selected as the forecasting model for the next days. Accordingly, the result is:

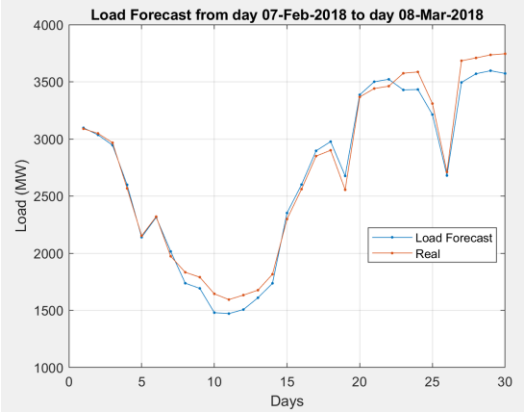


Figure 13. *Result for days February 7, 2018 to March 8, 2018*

The results of the SVR3 forecasting model for days February 7, 2018 to March 8, 2018 show the forecasting curve that matches the actual results recorded. The algorithm has solved the previous problem that is the forecasting result of this phase has a large error, due to deviation of load cycles between solar day and lunar day.

Forecasting results for 24 cycles of 07 days per week from March 1, 2018 to March 8, 2018 as follows:

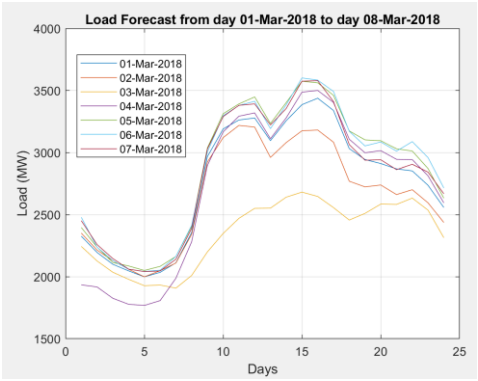


Figure 14. *Result for 24 cycles of 07 days*

From the results, it can be seen that the forecasting load profile are fully in accordance with the above standardized load profile. For example, the yellow curve (March 3, 2018) is the load profile of Sunday; the purple one (March 4, 2018) is the load profile of Monday.

IV. CONCLUSION

Research has shown that the combination of the Support Vector Regression (SVR) algorithm and the standardized load profile has been successful for short-term power load forecasting (STLF). MAPE values less than 1% indicates

that the SVR forecasting model is superior to current machine learning models. In addition, the SVR model has achieved a forecasting error rate of 5% below the allowable error of power companies and a tolerance of 3% for the electricity market. Thus, comparative results indicate that SVR models can be used as tools for forecasting electrical load with satisfactory results.

However, the results obtained in this study clearly show that SVR is very sensitive to the parameters such as: normalization coefficient C, parameter γ of kernel function and the width of tube ε. Therefore, the performance of the SVR gives good results depending largely on the choice of these three parameters. A specific study on the optimum input parameters of the SVR model is needed to further improve future forecasts.

REFERENCES

[1] Arunesh Kumar Singh, Ibraheem, S. Khatoon, Md. Muazzam, D. K. Chaturvedi, "An Overview of Electricity Demand Forecasting Techniques", Network and Complex Systems, National Conference on Emerging Trends in Electrical, Instrumentation & Communication Engineering, Vol.3, No.3, 2013.

[2] A. Y. Saber and A. A. Shareef, Load forecasting of a desert: a computational intelligence approach. in Proc. 2009. ISAP '09. 15th International Conference on intelligent system applications to power systems, 2009, pp.1–7.

[3] N. Amjady, F. Keynia, and H. Zareipour, "Short-term load forecast of microgrids by a new bilevel prediction strategy," *IEEE Transactions on Smart Grid*, vol.1, no.3, pp. 286–294, Dec. 2010.

[4] L. A. S. Romani, A. M. H. de Avila, D. Y. T. Chino, J. Zullo, Jr., R. Chbeir, C. Traina, Jr. and A. J. M. Traina, A new time series mining approach applied to multitemporal remote sensing imagery, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 140–150, Jan. 2013.

[5] Y. Goude, R. Nedellec, and N. Kong, Local short and middle term electricity load forecasting with semi-parametric additive models, *IEEE Transactions on smart grid*, vol. 5, pp. 440–446, Jan. 2014.

[6] M. Khashei and M. Bijari, A novel hybridization of artificial neural networks and ARIMA models for time series forecasting, Applied Soft Computing, vol. 11, pp. 2664–2675, 2011.

[7] L. A. Daz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera, A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile, Atmospheric Environment, vol. 42, pp. 8331– 8340, Nov. 2008

[8] P. Chang, C. Fan, and J. Lin, Monthly electricity demand forecasting based on a weighted evolving fuzzy neural network approach, Electrical Power and Energy Systems, vol. 33, pp. 17–27, Jan. 2011.

[9] A. Khosravi, S. Nahavandi, D. Creighton, and D. Srinivasan, Interval type-2 fuzzy logic systems for load forecasting: A comparative study, *IEEE Transactions on Power Systems*, vol. 27, pp. 1274–1282, Aug. 2012.

[10] D. X. Niu, Y. Wang, and D. D. Wu, Power load forecasting using support vector machine and ant colony optimization, Expert Systems with Applications, vol. 37, pp. 2531–2539, Mar. 2010.

[11] E. Ceperic, V. Ceperic, and A. Baric, A strategy for short-term load forecasting by support vector regression machines, *IEEE Transactions on Power Systems*, vol. 28, pp. 4356–4364, Nov. 2013.

[12] V.Vapnik, 1995, "The nature of statistical learning theory," Springer, NY.

[13] A.J. Smola, B. Schölkopf, Aug. 2004: A Tutorial on Support Vector Regression, Statistics and Computing, Vol. 14, No. 3, pp. 199 – 222.

[14] V. Cherkassky, Y. Ma, 2002: Selection of Meta-parameters for Support Vector Regression, International Conference on Artificial Neural Networks, Madrid, Spain, Aug. pp. 687 – 693.

[15] D. Basak, S. Pal, D.C. Patranabis, Oct. 2007: Support Vector Regression, Neural Information Processing – Letters and Reviews, Vol. 11, No. 10, pp. 203 – 224.

[16] Breiman L.: Random Forests. Machine Learning 45 (1), 5-32 (2001).

[17] Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: Classification and Regression Trees. Chapman & Hall (1984).