978-1-7281-4452-8/20/$31.00 ©2020 IEEE

Short Term Load Forecasting on PRECON Dataset

Ahmad Nadeem
Computer Science Depertment
Lahore University of Management Sciences
Lahore, Pakistan
ahmad.nadeem@lums.edu.pk

Naveed Arshad
Computer Science Depertment
Lahore University of Management Sciences
Lahore, Pakistan
naveedarshad@lums.edu.pk

***Abstract*—As the electricity market is growing, the need for accurate Short Term Load Forecasting (STLF) is increasing. Electrical grid operators require STLF to plan schedules for power generation plants. With the introduction of intermittent renewable resources, the stakes are now even higher. Developed countries have been fortunate in this regard as most of the research on STLF focused on these countries and developed highly accurate models. There is now a need to focus on developing countries as these are substantial energy markets with thriving economies and high population growth rates. With the 43^{rd} largest economy by GDP and 6^{th} largest nation by population, Pakistan is one such country. As the energy demand of Pakistan is increasing, there is a need to understand the energy demand patterns of its citizens better. PRECON is an electricity consumption dataset of residential buildings in Pakistan that can help in this regard. In this paper, we present preliminary results of applying STLF techniques on PRECON. These initial results show that Multiple Linear Regression and Support Vector Regression perform better than Artificial Neural Network ambient temperature and autoregressive attributes as input variables. The results also discuss various performance metrics, such as ME, RMSE, and MPE. The results show a unique phenomenon, load shedding, not experienced in developed countries.**

***Index Terms*—STLF, PRECON, Machine Learning**

I. Introduction

Dependence on electricity is rapidly growing with the introduction of new and advanced technologies. One of the ways to measure productivity and rate of development of a country is to calculate the amount of electricity it consumes [1]. An increase in electricity consumption at an alarming rate has become a challenge for distribution companies. The government enforces Electricity Distribution Performance Standards (EDPS) [2] to ensure quality and convenience to its citizens. EDPS states that the Planned Power Supply Interruptions should be limited and controlled, and companies should maintain steady voltage and frequency levels with minimum variations. The power sector requires accurate Short Term Load Forecasting (STLF) to comply with EDPS. Forecasting the consumer load in advance allows distribution companies to be prepared for future events.

STLF is crucial for the power system operator. With varying load demand, the operator controls electricity generation in such a manner that the voltage and frequency of the power grid comply with the limits defined in EDPS. Grid-tied power generation plants have to be notified in advance on when to start generating electricity, as it takes several minutes for the process to initialize. The time elapsed from being notified to the beginning of the generation is called ramp rate [3]. Generation plants with shorter ramp rates effectively cost more. For an efficient system, the operator has an immense responsibility in both optimizing the cost and in providing a high-quality supply of electricity to its consumers.

978-1-5386-5541-2/18/$31.00 ©2018 IEEE

It is unfeasible to accommodate the growing electricity demand using thermal power plants. Such a generation requires the utilization of fossil fuels, which themselves are depleting rapidly. Thermal power plants also cause pollution, further creating environmental issues and affecting the health of the population and natural habitats. Efficient alternatives are renewable resources. Wind and solar power are the leading renewable resources to generate electricity. STLF is vital for the high penetration of these intermittent renewable resources within the power grid. The system operator requires both the accurate forecast of these renewable resources [4], as well as the precise load forecasts to make impeccable decisions. Without accurate STLF, these renewable resources will become an inconvenience rather than being beneficial.

Short Term Load Forecasting(STLF) is an essential part of a Smart Grid. Without any accurate forecasts, smart grids are unable to make error-free decisions in time. The automatic control of generation resources, smart charging of storage batteries, and net metering with the utility are operations performed by the smart grid, which require accurate load forecasts. The smart grid also requires STLF for Demand Response (DR) operations [5]. In the DR management system, the demand of a consumer is altered to match the generation with respect to electricity price changes, generation-demand deficit, and conservation of electricity to avoid excessive use of electricity.

STLF is also crucial for the exchange of electrical energy between prosumers - entities that produce and consume electricity. With an accurate forecast, smart grids can decide the volume of energy which needs to be shared and the price for delivering electrical energy in times when it is excess. A new technology called Soft Load Shedding(SLS) [6] has been recently introduced. SLS is applied when there is a collective deficit of energy among interconnected nodes of a smart grid. The algorithm uses STLF to plan the distribution of electrical energy resources in such a manner where all the consumers

978-1-7281-4452-8/20/$31.00 ©2020 IEEE

are treated fairly. SLS further employs that consumers have flexible loads, which can be shifted to non-peak hours to manage the deficit. Consumers are ranked using clustering techniques such as K-means, and each cluster of consumers is assigned a quota of energy to be consumed for each time interval.

Most of the research is initiated in developed countries to implement these new techniques and technologies. However, developing countries are lagging in this area. Due to the lack of resources and several inherent problems within developing countries, researchers are unable to study and analyze the compatibility of current smart grids with consumer behaviors and patterns of electricity consumers. One major problem is the absence of a comprehensive electricity consumption dataset being publicly available. However, recently, a new dataset that monitors the electricity demand of households in Pakistan is published, which is called Pakistan Residential Electricity Consumption (PRECON) dataset. Pakistan is a developing country located in South Asia. It is the 6^{th} largest country by population [7] and is the 43^{rd} largest country by GDP [8]. With a flourishing economy and rapidly increasing economy, the electricity demand of Pakistan is increasing rapidly. In this paper, we use various techniques to forecast electricity consumption of households featured in PRECON. A linear regression model, a support vector machine, and an artificial neural network are used for such purpose. Utilizing daily and weekly lags, hour label, and a weather indicator such as apparent ambient temperature, satisfactory results were obtained with some limitations.

II. Literature Review

Short Term Load Forecasting (STLF) is a well-studied field in literature. The authors of [9] first defined STLF in 1987 and discussed the importance of its integration in the Energy Management System (EMS). This paper also discusses the seasonality of the load demand and external factors such as weather, time, and economy. One of the first techniques discussed for STLF includes ARIMA and its variants. The ARIMA model exploits the variation of the load demand due to changes in seasons. In [10], a comparison of Fuzzy Logic (FL), Neural Networks(NN), an Auto-Regressive (AR) model are provided. By simulating the three models on electricity load data from Texas, it was found that the performance of NN and FL is far superior to AR. This study was performed in 1996.

STLF has always been an area of interest in the department of electrical engineering for several years. However, with the advancement in machine learning, STLF is no more an exclusive problem to electrical engineering. In [11], a classic example of a neural network application to STLF is presented. Forecast of the load pattern for the next 24-hour is presented by using electricity consumption data of industrial customers from Korean Utility companies.

Residential Load Forecasting using Canadian household's data as a test case is presented in [12]. LSTM recurrent neural network was used to produce accurate forecasts. However, it was concluded that the consumption pattern of a single customer is too volatile to be forecasted. The model presented also uses appliance-based load patterns to increase accuracy. Another such study [13], provides evidence that using seasonal factors such as weekly lags and hour of the day have strong prediction power. The paper used electricity consumption data from the Australian distribution of 27 households.

Most of the studies found in literature use electricity consumption from developed countries as test cases to justify the performances of their developed models. In [14], authors used electricity load data from households within the UK. Similarly, [15]–[17] used to load data from Korea, China, and Ireland, respectively, to demonstrate the forecasting power of various techniques and machine learning algorithms.

However, there is a significant vacuum in the literature for studies that focus on electricity consumption in developing countries. PRECON is an effort to fill this gap. STLF is a relatively new topic for developing countries, as the region has been ignored in the past. This paper provides a preliminary effort to forecast the residential load demand from the PRECON dataset.

III. Methodology

Evolving power grids in developing countries demand an accurate short term load forecasting. Predicting the electric load of households is a complex problem that requires multiple models and carefully selected independent variables. Three of the commonly used models, such as linear regression, support vector machine, and artificial neural network, are used in this paper. The independent variables used to train these models are the ones commonly used for load forecasting. The following section provides a brief description of the various aspects of the developed models.

A. Dataset

PRECON is a household electricity consumption dataset generated by installing smart meters in 42 households within Lahore, Pakistan. This unique dataset consists of year-long information at the sampling rate of a minute. The whole procedure for collection and preprocessing of the dataset is explained in [18], and the dataset is available at [19]. The dataset covers a diverse set of households. Figure 1, shows the monthly boxplot of average electricity demand by all households. The median electricity demand in January till March is less than 1 kW, which increase in the following months until August when it peaks at 2.5 kW. The electricity demand starts reducing again in the last months of the year.

B. Dependent Variables

The PRECON dataset contains the total electricity usage of a household for every minute for a year, which is then averaged out to be converted to hourly values. Minute interval forecasts are not convenient for smart grids as most systems require hourly forecasts. The majority of the weather services also provide hourly forecasts, which are further utilized in the models. Moreover, minute interval time series is reduced to

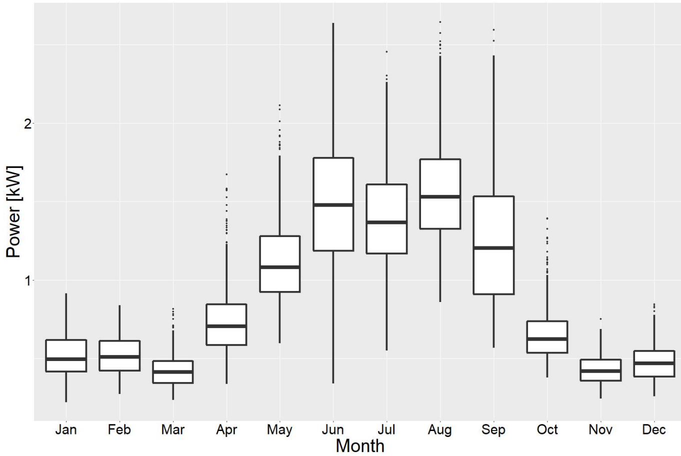

Fig. 1: Average electricity demand distribution for each month for all households

hourly values to decrease the random error in fitted models. The variation in minute interval load demand due to various activities within households are also smoothed out in hourly time series.

C. Independent Variables

One of the most successful approaches in forecasting time series is autoregression. In this technique, the lag values of the dependent variable are used as input for the model in calculating the upcoming values. The selection of the lag values depends upon the patterns and seasonality within the dataset. In the case of electric load forecasting, it is intuitive to include a day lag, which is the value of the same hour from the previous day. As a result of this assumption, the load of a household at any hour of the day depends upon the electric load of the household on the same hour of the previous day.

Another lag value used for training the models is week lag. Similar to day lag, weekly lag highlights the fact that households have a pattern of consuming electricity that repeats every week. The main reason for incorporating weekly lag was to differentiate between daily and weekly occurring routines and habits.

One of the best explanatory variables used for the training models can be the hour lag. Hour lag states that electricity consumption at any hour depends upon the electricity consumption in the immediately preceding hour. Such a phenomenon is known as Time Series Momentum [20]. However, including hour lag is infeasible under the assumption that the system will already have the electricity consumption observation for the current hour required to forecast the next hour value. An hour label parameter was used instead, to suffice for this shortcoming. Hour label informs the model regarding the hour it has to forecast. The model is trained to differentiate between each hour of the day using this parameter without using hour lag.

On the other hand, the weather has a significant influence on electricity consumption in households. Citizens of Lahore consume more electricity in summer as the temperature rises to 48°C [21]. Households use various cooling appliances such as fans, air coolers, and air conditioners. However, incorporating only the temperature as a dependent variable is inefficient. Humidity and wind speed also has a high impact on the overall temperature that people feel. In training the models, a combination of temperature and humidity is used to calculate Apparent Temperature(AT) [22], which is a better representative of the actual temperature that people feel. AT has a high correlation of 0.8 with average electricity consumption. This high correlation is also shown in figure 2; the changing pattern of AT corresponds to the pattern observed in electricity demand of PRECON households shown in figure 1.

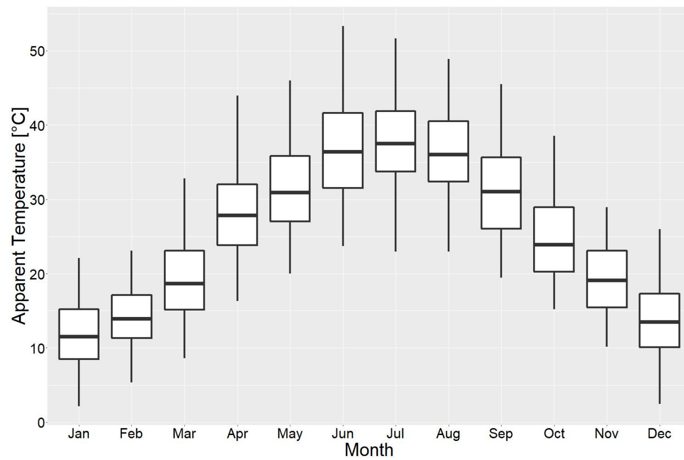

Fig. 2: Boxplot showing distribution of apparent temperature in each month for the city of Lahore

D. Multiple Linear Regression Model

Multiple Linear Regression (MLR) is a simplistic approach in finding the relationship between a dependent variable such as load demand of a household and the various independent variables. The model used in this study employs Ordinary Least Squares(OLS), which predicts the unknown β coefficients of dependent variables by minimizing the sum of squares of errors. The model can be represented as:

$$\begin{aligned} Load_t = \beta_o + \beta_1 Load_{t-24} + \beta_2 Load_{t-168} \\ + Apparent\,Temperature_t + \beta_3 Hour_t \end{aligned} \tag{1}$$

Where $Load_t$ is the electricity load demand of a household at time t, subscript $t-24$ and $t-168$ shows the day and week lags. $Hour_t$ is the hour label for the hour at time t.

E. Support Vector Regression

Support Vector Regression(SVR) is used for modeling continuous numeric variables. It is derived from Support Vector Machines (SVM), which was primarily developed for classification problems but can also be used for regression. In this paper, the SVR of Epsilon Regression (ϵ-regression) type and Radial Basis Function(RBF) kernel is used. The ϵ-regression allows the tweaking of ϵ hyperparameter, which controls the sensitivity of errors within the model. Using the RBF kernel allows tuning of the model using hyperparameter γ controlling the smoothness and complexity of the model.

F. Artificial Neural Network

Artificial Neural Network(ANN) has recently gained a great deal of popularity for solving both classification and regression problems. ANN consists of artificial neurons which make up layers within the network. A typical ANN used for regression has multiple input neurons and a single output neuron with several hidden layers. The connection between these layers are called edges and is assigned a specific weight. The ANN model uses several supervised examples to learn the dataset and is later used to label the unseen data. In this study, ANN used for forecasting the PRECON dataset contains a single hidden layer with eight neurons. One of the samples ANN can be observed in figure 3.

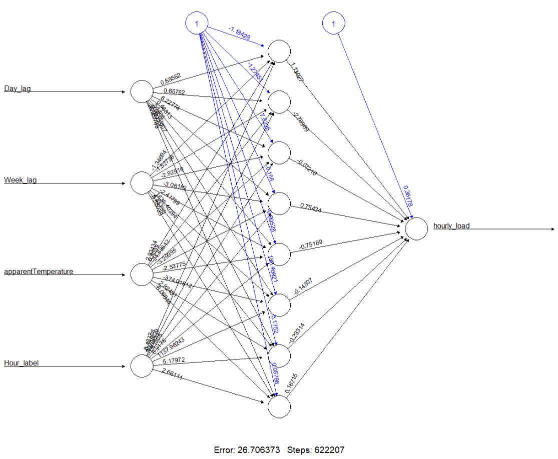

Fig. 3: Trained neural network used for STLF

G. Model Training

By using all the four dependent variables, i.e., hour and week lag, apparent temperature, and hour label, all three models above were trained for every household separately. PRECON contains the total usage of the households in kW for a year from 1st June 2018 till 31st May 2019. For training the model, the information regarding electricity consumption of the first 11 months was used, and the last month was used as test data. All models had 8016 data points to learn the electricity consumption of each household.

IV. Results & Discussion

Various techniques are used to calculate the performance of individual models for each household. Error metrics commonly used for this purpose include Mean Error (ME), Root Mean Square Error (RMSE), and Mean Percentage Error (MPE). ME is an excellent metric to confirm whether the model is under or overestimating, whereas RMSE is useful to understand the spread of the errors. However, both ME

	ME	$RMSE$	MPE
Minimum	-0.17	0.22	-20421.29
1_{st} Quartile	-0.01	0.33	-208.22
Median	0.03	0.48	-96.38
Mean	0.04	0.53	-854.62
3_{rd} Quartile	0.07	0.62	-43.62
Maximum	0.28	1.46	4.19

(a) Multiple Linear Regression Model

	ME	$RMSE$	MPE
Minimum	-0.11	0.13	-9277.45
1_{st} Quartile	0.02	0.34	-201.38
Median	0.09	0.48	-58.88
Mean	0.18	0.54	-469.70
3_{rd} Quartile	0.20	0.65	-24.68
Maximum	0.62	1.45	17.16

(b) Support Vector Regression

	ME	$RMSE$	MPE
Minimum	-2.77	0.25	-56397.65
1_{st} Quartile	-0.39	0.50	-186.46
Median	0.09	0.69	-133.83
Mean	-0.09	0.88	-1889.23
3_{rd} Quartile	0.39	1.00	-18.09
Maximum	0.82	3.46	52.03

(c) Artificial Neural Network

TABLE I: Performance of each model

and RMSE lack a comparison between the error and the actual value. MPE compares the error to the actual amount and provides a better evaluation of the model performance. Equations below show the formulas used to calculate the error metrics:

$$ME = \frac{1}{n}\sum_{t=1}^{n} Actual_t - Forecasted_t \quad (2)$$

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

$$MPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{Actual_t - Forecasted_t}{Actual_t} \quad (4)$$

Table I shows the summary of errors in each model for all the households. Evaluating the results showed that the linear and support vector regression model with a range of ME of 0.45 and 0.73 respectively performed better than the neural networks with the ME range of 3.59. For this observation, the range is defined as the difference between the maximum and minimum value. The same is evident from the range of RMSE.

However, MPE shows very unusual results. This is because when at least at a single point, the actual value is very small, the percentage error becomes large. As outliers inherently highly influence the mean, MPE shows a significant error. In other words, MPE is an inefficient metric if the actual values to be forecasted approaches zero. Three of the houses identified with such data are House 12, 37, and 42. Figure 4 shows the actual and forecasted data by the three models for these three households. As can be observed, the actual values approach zero at several points for all these three houses, which explains the odd values of MPE. The load demand of a household in developed countries hardly approaches zero as some appliances always provide the baseload. However, that is not the case with PRECON. For some hours of the day, electricity supply to the households is cut off by the utility due to overload and power deficits. This phenomenon is called load-shedding, and it is quite common in developing countries. These results show that there is a need to model the load-shedding aspect of the electric grid to get better STLF accuracy.

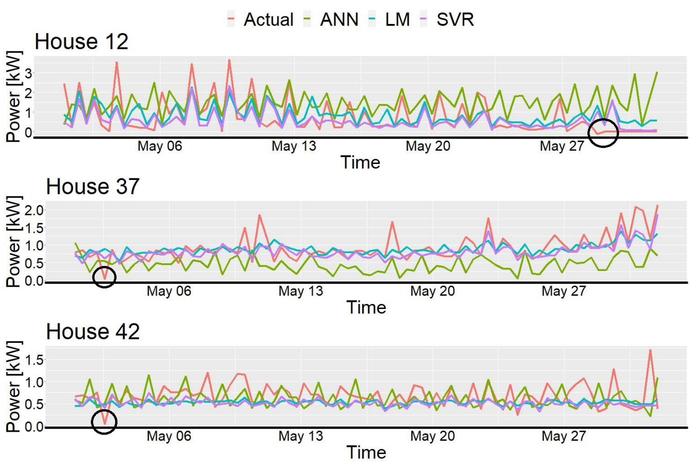

Fig. 4: Actual and forecasted demand for three households. The encircled area shows the events when the actual demand reached zero due to loadshedding

Figure 5 shows boxplot for model error and actual electricity consumption for the forecasting month by averaging the values at each hour of the day for every household involved. As the first plot shows the error, an accurate forecast should have short boxplots centered at zero. The boxplots created by ANN are long and below the zero lines for most of the day, which means that ANN is overestimating with a significant variation. LMR and SVM, on the other hand, are closer to zero lines with shorter boxplots. However, all three of the models show similar patterns in their error boxplots in comparison to the actual electricity consumption boxplots. Even after using the hour label as input, all three models were unable to learn the pattern thoroughly.

The results obtained in this study are not satisfactory, with much room for improvement. Forecasting day-ahead electricity consumption of an individual household is a tedious task [23], [24]. In this paper, only external attributes that influence electricity consumption are used. The models designed in this study can not anticipate the usage of individual home appliances in a household. The lifestyle of residents of a household also needs to be integrated into the model to get a more accurate forecast. Unexpected events such as leaving the household or returning at unexpected times by a resident can profoundly influence the forecast. To increase the accuracy of the forecast, these internal influences of electricity consumption need to be documented and analyzed. Other than this, modeling load-shedding and other cultural and religious events will increase the accuracy of the forecasting model.

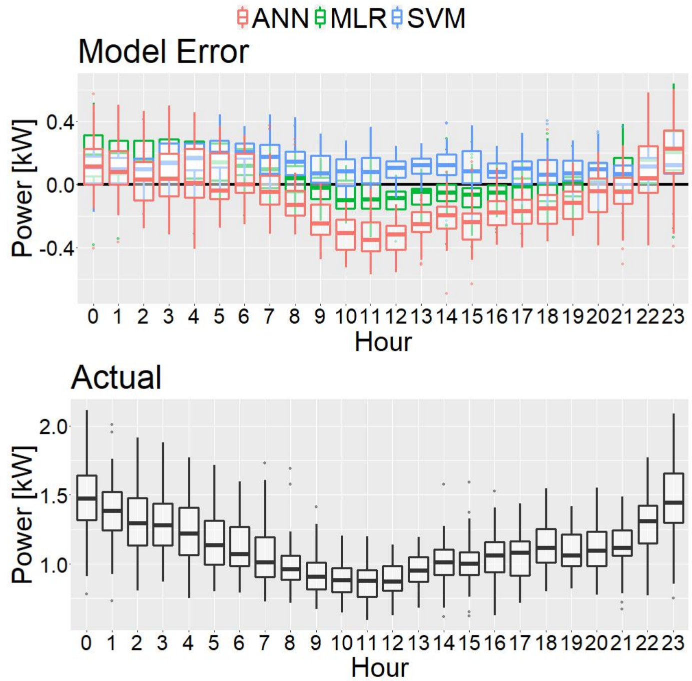

Fig. 5: Model error and the actual load demand distribution

The data used for training the models was less than a year, which is relatively small for creating highly accurate forecasting models. With data available for a longer duration, these models ought to produce better results. PRECON is a relatively new dataset. As the research on the dataset will develop with time, our understanding of the electricity consumer behavior in this new locality i.e., a developing country, will evolve, which will result in improved forecast results.

V. Conclusion

The aim of this study was two folds. First, to create an initial bar for any future study to produce even better results and second, to highlight the challenges involved in STLF in developing countries. The forecast results presented in this paper will encourage researchers to create even better forecasting models. As shown in this paper, to get an accurate forecast, machine learning algorithms can not only rely on external factors. Input from inside the household by monitoring individual appliances, room occupancy, and daily routines of the residents will also increase the accuracy. We need to reevaluate our approach to STLF for developing countries because of its unique consumer behaviors such as cultural and religious holidays and power grid characteristics such as loadshedding, which is a rare phenomenon in developed countries.

References

[1] C. for Global Development. (2016) Electricity consumption and development indicators. [Online]. Available: cgdev.org/media/electricity-consumption-and-development-indicators

[2] N. E. P. R. Authority. (2005) Electricity distribution performance standards (edps). [Online]. Available: https://nepra.org.pk/Legislation/Rules/Performance%20Standards%20(Distribution)%20Rules%202005.pdf

[3] A. J. Svoboda, C.-L. Tseng, C.-a. Li, and R. B. Johnson, "Short-term resource scheduling with ramp constraints [power generation scheduling]," *IEEE Transactions on Power Systems*, vol. 12, no. 1, pp. 77–83, 1997.

[4] R. H. Inman, H. T. Pedro, and C. F. Coimbra, "Solar forecasting methods for renewable energy integration," *Progress in energy and combustion science*, vol. 39, no. 6, pp. 535–576, 2013.

[5] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE transactions on industrial informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[6] T. Aslam and N. Arshad, "Soft load shedding: An efficient approach to manage electricity demand in a renewable rich distribution system." in *SMARTGREENS*, 2018, pp. 101–107.

[7] M. Basit, S. Sajjad, M. I. Khan, A. Ali, and S. K. Kurshid, "Spatio-temporal trends of urban population in pakistan," *Asian Journal of Multidisciplinary Studies*, vol. 6, p. 8, 2018.

[8] S. Times. (2018) Projected gdp ranking (2019-2023). [Online]. Available: statisticstimes.com/economy/projected-world-gdp-ranking.php

[9] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558–1573, 1987.

[10] K. Liu, S. Subbarayan, R. Shoults, M. Manry, C. Kwan, F. Lewis, and J. Naccarino, "Comparison of very short-term load forecasting techniques," *IEEE Transactions on power systems*, vol. 11, no. 2, pp. 877–882, 1996.

[11] S. Ryu, J. Noh, and H. Kim, "Deep neural network based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, 2017.

[12] W. Kong, Z. Y. Dong, D. J. Hill, F. Luo, and Y. Xu, "Short-term residential load forecasting based on resident behaviour learning," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 1087–1088, 2017.

[13] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Applied Energy*, vol. 205, pp. 654–669, 2017.

[14] B. Stephen, X. Tang, P. R. Harvey, S. Galloway, and K. I. Jennett, "Incorporating practice theory in sub-profile models for short term aggregated residential load forecasting," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1591–1598, 2015.

[15] H. Son and C. Kim, "Short-term forecasting of electricity demand for the residential sector using weather and social variables," *Resources, conservation and recycling*, vol. 123, pp. 200–207, 2017.

[16] X. Fu, X.-J. Zeng, P. Feng, and X. Cai, "Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in china," *Energy*, vol. 165, pp. 76–89, 2018.

[17] A. Kell, A. S. McGough, and M. Forshaw, "Segmenting residential smart meter data for short-term load forecasting," in *Proceedings of the Ninth International Conference on Future Energy Systems*. ACM, 2018, pp. 91–96.

[18] A. Nadeem and N. Arshad, "Precon: Pakistan residential electricity consumption dataset," in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. ACM, 2019, pp. 52–57.

[19] EIG. (2019) Precon dataset. [Online]. Available: http://web.lums.edu.pk/ eig/precon.html

[20] T. J. Moskowitz, Y. H. Ooi, and L. H. Pedersen, "Time series momentum," *Journal of financial economics*, vol. 104, no. 2, pp. 228–250, 2012.

[21] M. Zahid and G. Rasul, "Frequency of extreme temperature and precipitation events in pakistan 1965–2009," *Science International*, vol. 23, no. 4, pp. 313–9, 2011.

[22] R. G. Steadman, "A universal scale of apparent temperature," *Journal of Climate and Applied Meteorology*, vol. 23, no. 12, pp. 1674–1687, 1984.

[23] S. Humeau, T. K. Wijaya, M. Vasirani, and K. Aberer, "Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households," in *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*. IEEE, 2013, pp. 1–6.

[24] R. Sevlian and R. Rajagopal, "Short term electricity load forecasting on varying levels of aggregation," *arXiv preprint arXiv:1404.0058*, 2014.