Development of an Efficient Electricity Consumption Prediction Model using Machine Learning Techniques

Ghaidaa Hamad Alraddadi[1]
Department of Computer Science
College of Computer, Qassim University
Buraydah 51452, Saudi Arabia

Mohamed Tahar Ben Othman[2]
BIND Research Group, IEEE Senior Member
Department of Computer Science, College of Computer
Qassim University, Buraydah 51452, Saudi Arabia

***Abstract*—Electricity consumption has continued to go up rapidly to follow the rapid growth of the economy. Therefore, detecting anomalies in buildings' energy data is considered one of the most essential techniques to detect anomalous events in buildings. This paper aims to optimize the electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies. Further, as the used dataset is huge and published publicly, many research used part of it based on their needs. In this paper, the dataset is grouped as daily consumption and monthly consumption to compare the network topologies of all other works that used the same dataset with the selected part. The proposed methodology will depend basically on long short-term memory (LSTM) because it is powerful, flexible, and can deal with complex multi-dimensional time-series data. The results of the model can accurately predict the future consumption of individual households in a daily or monthly consumption base, even if the household was not included in the original training set. The proposed daily model achieves Root Mean Square Error (RMSE) value of 0.362 and mean absolute error (MAE) of 19.7%, while the monthly model achieves an RMSE value of 0.376 and MAE of 17.8%. Our model got the lowest accuracy result when compared with other compared network topologies. The lowest RMSE achieved from other topologies is 0.37 and the lowest MAE is 18% where our model achieved RMSE of 0.362 and MAE of 17.8%. Further, the model can detect the anomalies efficiently in both daily electricity consumption data and monthly electricity consumption data. However, the daily electricity consumption readings are way better to detect anomalies than the monthly electricity consumption readings because of the different picks that appear in the daily consumption data.**

Keywords*—Anomalies detection; deep learning; electricity consumption forecasting; LSTM

I. Introduction

Global electricity consumption has grown rapidly faster than the rate of energy consumption where the electricity consumption in both commercial and residential buildings has significantly increased and can account between 20% and 40% in developed countries [1-2]. During 1980-2013, energy consumption went up from 7300 TWh to 22100 TWh. Also, it has grown even faster since the twenty-first century by 1.2 percentage points more than the average annual rise in energy consumption. In 2013, the annual electricity consumption of the world reached 3048 KWh per capita which is up to 42.3% from 1990. In Asia, Bahrain, South Korea, and United Arab Emirates are the top three consumers where they exceeded 10000 KWh [2].

As a result, the energy demand will steadily be increased shortly due to the rise in population, comfort levels of buildings and spending a long time inside buildings. Thus, optimizing energy consumption and the efficiency of energy in buildings is a primary concern for anyone wishing to save energy [1]. Although the extensive modeling techniques that investigate designing buildings with a low level of energy consumption, buildings (especially commercial) often exceed the promised energy-saving design by some anomalous events, such as lighting equipment faults. These anomalous events can reach figures between 2-11% of the total energy consumption in commercial buildings [3].

To this end, detecting anomalies in buildings' energy data is considered one of the most essential techniques to detect anomalous events in buildings. Therefore, metering buildings' electricity provides data that have a significant impact on energy-saving opportunities due to analyzing energy usage, managing energy consumption and, thus, identifying anomalous patterns which if corrected will improve the comfort level of buildings with the least power consumption.

At a household level, the electricity consumption includes a high amount of noise, and the time series data is considered nonlinearly because of the seasonal changes' effects. Thus, this is a motivative of using machine learning field because of its capability to capture the nonlinearities among the time series data. In particular, LSTM models have proven effective in this type of context and in learning complex nonlinear patterns [4]. In this study, the aim is to solve the high volatility and uncertainty of electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies.

Therefore, the proposed approach can deal effectively with a large number of smart meters data. After training and implementing the appropriate parameters, the resulting model will have the ability to forecast the future consumption of households that were not included in the training process. Hence, the resulting model has a great potential to forecast big data accurately. The validity of the proposed approach is tested based on an extensive real-world dataset that contains

thousands of households' consumption in several years so, several seasonal patterns. The used dataset is originally published in [5]. It is a collection of 104057 records of London's households from 2011 to 2014. In addition to the date and time, it consists of a unique household identifier, Acorn group categories, which is a classification of neighborhoods. Furthermore, historical weather data for London area during the same dates of smart meters registers was merged with the original dataset, which can be found in [6]. The weather dataset consists of the date of the day, maximum and minimum temperature of that day, high and low apparent temperature, wind bearing, dewpoint, cloud cover, wind speed, pressure, visibility, humidity, UV index, and moon phase.

The paper is organized as follows. Section II outlines historical selected works performed within electricity consumption forecasting. Then, the model is proposed in Section III. In Section IV, the results of the electricity consumption forecasting with the proposed model are discussed and shown. Finally, Section V concludes the paper and Section IV shows the future work.

II. Related Work

Recently, many machine learning models have contributed very well to estimating energy consumption by prediction models. The following section mentions some of the recently published researches in the areas of predicting energy consumption by using different machine learning approaches.

In references [7-9], they reviewed the state of the art of machine learning models of all kinds of energy systems including demand prediction, cost prediction, energy consumption prediction, load forecasting, etc. They identified the highest popularity models in energy systems which are Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machines (SVM), Extreme Learning Machine (ELM), Wavelet Neural Network (WNN), ensembles, deep learning, and hybrid machine learning models. Nevertheless, the hybrid machine learning models have significantly increased the performance of energy models.

Besides, García-Martín et al. illustrated in [10] the recent approaches used to estimate the energy consumption especially from data mining and neural networks perspectives. They revealed some emerged works, such as NeuralPower and SyNERGY that allow energy evaluation in machine learning. NeuralPower evaluates the energy by building a prediction model to estimate the energy consumption while SyNERGY builds energy estimation models based on integrating tools to the current machine learning suites. Further, they presented some challenges faced by current modeling approaches, such as the rapid changes in hardware, implementation and design of neural networks.

In [11] the authors proposed a hybrid approach that combines ELM with stacked autoencoders (SAE) to predict the energy consumption in buildings. Firstly, they used SAE to extract the features of energy consumption in buildings. Then, for prediction, they employed ELM to get precise prediction results. Their results showed that their proposed approach has the best prediction performance compared to other approaches, such as SVM, multiple linear regression, backpropagation neural network (BPNN) and the generalized radial basis function neural network (GRBFNN). Also, in [12], they used deep extreme learning machine and SVM techniques to predict the energy consumption. Their proposed model obtained an accuracy of 90.70%.

J. Y. Kim and S. B. Cho. [13] proposed an autoencoder model to predict electricity consumption based on LSTM. The proposed model defines a state that represents the demand information, then, the future energy demand is predicted according to this state. The state contains information like the input values features, produced data features and the expected energy consumption. Moreover, this model allows inserting conditions to predict the electricity demand according to these conditions, such as economy or weather information and that is what makes it more efficient compared to other works. In addition to [13] and [11], Y. Jin *et al.* [14] proposed a clustering analysis method to analyze the daily electricity consumption based on an autoencoder algorithm. Their suggested method has a limitation of outlier detection when there are large outlier data.

In [15], they used different machine learning algorithms such as linear regression (LR), DTs, deep neural network (DNN), recurrent neural network (RNN), gated recurrent units (GRUs) and LSTM to evaluate their performance. They forecasted the data based on the ACORN groups in London. Then, they forecasted number of step-ahead like 1, 2, 12, 24, 48. LSTM achieved the lowest MAE compared to other algorithms for all forecasting types where they were 19%, 21.7%, 23.5%, 24.2%, and 25.6%, respectively. On the other hand, forecasting one step ahead has the lowest MAE and the worse was forecasting 48 steps ahead, as shown in Fig. 1. Furthermore, the authors in [16] proposed a machine learning-based ensembled model to improve the electricity consumption prediction. The model combines Cat Boost (CB), Gradient Boost (GB) and Multilayer Perceptron (MLP) algorithms. Moreover, they employed the genetic algorithm to get optimal features to be used for the model. They obtained RMSE of 5.05 and MAE of 3.05.

F. Z. Abera et al. proposed in [17] a method that uses the CLARA clustering technique to group their dataset into three clusters based on the mean of the consumption values. Then, SVM and ANN classifiers are used to predict the appliance that consumes more energy. They proved that ANN and SVM are worthful methods for analyzing and forecasting smart meter data with an accuracy of 99%.

M	Regression Model					
	LR	DT	DNN	RNN	GRU	LSTM
1	26.2%	24.9%	23.5%	22.4%	22.5%	**19.2%**
2	27.6%	27.5%	26.7%	25.0%	27.6%	**21.7%**
12	30.2%	28.1%	27.1%	27.7%	27.3%	**23.5%**
24	30.0%	28.1%	28.3%	28.8%	28.4%	**24.2%**
48	30.7%	29.8%	28.7%	28.1%	27.8%	**25.6%**

Fig. 1. Forecasting Results of [15].

D. H. Nguyen et al. proposed in [18] a machine learning-based approach called iRBF–NN to predict the electricity consumption based on historical consumption and weather data. They demonstrated the relation between weather parameters (temperature, humidity, precipitation, sunshine duration and wind speed) and electricity consumption. They found that the humidity and temperature parameters have the highest relation to electricity consumption. Other parameters such as population and sunshine also affect the electricity consumption, however, according to the hardness of collecting their data and the simplicity of the model they did not consider them. The prediction performance of the proposed approach was good; however, the weather prediction was not accurate. E. Y. Shchetinin. [19] proposed a method to estimate the electricity consumption in commercial and business buildings by using a gradient boosting algorithm (GBM). Their results showed that GBM has the ability to estimate the accuracy of energy consumption prediction more than other machine learning algorithms like random forest and regression. Conversely, X. M. Zhang et al. used in [20] support vector regression to predict residential electricity consumption (rather than commercial consumption) according to their daily consumption and hourly consumption. Their results showed that the MAPE error is 12.78 and 22.01 for daily prediction and hourly prediction, respectively, and that means predicting daily consumption is better than predicting hourly consumption since it mitigates the effect of the randomness of behaviors in hourly family members.

S. Aman et al. proposed in [21] a novel model namely REDUCE (Reduced Electricity Consumption Ensemble) that combines outputs from three base models. These three models are called pre-DR (demand response), in-DR and all-day consumption sequences. Their results showed that the model is strong for buildings that do not follow a strict schedule of electricity consumption and they do not have enough historical demand response data.

On the other hand, Eisses, J. used in [22] three different machine learning techniques to detect anomalies in electricity consumptions data which are k-nearest neighbors (KNN), SVM and ANN. They conclude that ANN has the best accuracy performance with an error of 14%. Furthermore, K. Hollingsworth et al. [23] proposed an application for detecting anomalies in energy. Their application is based on the combination of two types of machine learning algorithms: ARIMA and LSTM. Their application correctly identified the anomalies and provided the time of the incident. Also, it provides higher accuracy by benefiting from both separated models' abilities. In [24], they combine K-means and DNN to identify the anomalies in energy consumption. Firstly, K-means is used to cluster the customers based on their similar electricity consumption behavior. Then, DNN algorithm is used to accurately identify the anomalies of each consumer.

III. Methodology

Long short-term memory (LSTM) is a recurrent neural network (RNN) that is used widely in the deep learning field. RNN has the ability to process any hidden patterns that exist in the data since it takes into consideration the sequential nature of the data. Therefore, it does not feed all the information to the network at once, but it feeds them as a chain structure where one element is processed and then passed to the second element in the sequence. In other words, RNN is recurrent in nature since it implements the same function for each input of the data whilst the output of the current input relies on the previous computation. Once the output is produced, it is copied and sent back into the recurrent network [25-26]. Fig. 2 shows the unrolled recurrent neural network where X_t is the input of the state, h_t is the output of the state and A is the activation function of the state. Firstly, X_0 is taken from the sequence of the input and then outputs h_0. After that, h_0 and X_1 will be input for the next step. Similarly, h_1 from the next step with X_2 will be the input for the next step, etc.

Despite the stability of RNN, it has challenges in practice. It suffers from a well-documented problem called vanishing gradients.

The vanishing gradient is encountered when training a large number of samples since it requires many layers. Therefore, the gradient reduced dramatically because it propagated through the network [27].

LSTM can solve the issue of the vanishing gradient by capturing long-term dependencies. It's a well-known branch of deep learning and gained wide attention to forecasting time series data in recent years.

In this study, the aim is to solve the high volatility and uncertainty of electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies.

A. Data Description and Preprocessing

The used dataset is real data of electricity consumption that was originally published in [5]. In particular, the dataset consists of electricity consumption readings of 5567 London households from November 2011 to February 2014. Readings measured in kWh were taken in half-hourly intervals. The dataset contains date and time, unique household identifier, Acorn group categories, which is a classification of neighborhoods. Furthermore, historical weather data for London area during the same dates of smart meters registers was merged with the original dataset, which can be found in [6]. The weather dataset consists of the date of the day, maximum and minimum temperature of that day, high and low apparent temperature, wind bearing, dewpoint, cloud cover, wind speed, pressure, visibility, humidity, UV index, and moon phase. Since the length of the samples varies from one household to another and it might that one household may have only one reading sample, we must assure that each household contains at least two readings one for each year, therefore, the others are excluded.

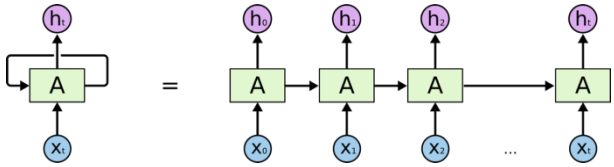

Fig. 2. Recurrent Neural Network.

As clarified earlier, the original electricity consumption data contains 167,932,474 lines of half-hourly readings for 5567 households in, approximately, four years. Nevertheless, there exist some missing consumption values. Therefore, to avoid the undesirable impact on the forecasting model data cleansing process should be done. Firstly, all NULL values are replaced with NAN. Meanwhile, there are 27858 of 0s in the consumption column where we consider them as inconsistent values. Therefore, they are replaced with NAN. After that, all NAN values are imputed by the average of the electricity consumption column. Fortunately, there are no missing or inconsistent values in the weather dataset. From the resulting dataset, 80% of the data are considered as training set and the other 20% for the testing set.

The proposed methodology is general and can deal with high-frequency time series data. However, the dataset is huge, and it contains more than 167 million records even after cleaning. Therefore, to reduce the dimension and volatility, the dataset is aggregated into daily intervals by calculating the mean consumption of each day for each household, as shown in Fig. 3. Moreover, for a comparison purpose, the average monthly electricity consumption is calculated for each household, as shown in Fig. 4. Fig. 3 illustrates the daily electricity consumption sorted by households' id. Although most of the consumption runs on average, there are many picks on different days that have high consumption. On the contrary, there is a clear difference between the monthly consumption in Fig. 4 and the daily consumption in Fig. 3 because when considering monthly consumption, it is mostly going to the average except for a huge pick in the middle. This potentially signs to get better results when using daily measurements for detecting anomalies rather than the monthly consumption because of the detailed data that appear in the daily readings.

B. Proposed Model

Generally, the proposed methodology depends basically on LSTM because, as mentioned earlier, it is powerful, flexible, and has the ability to deal with complex multi-dimensional time-series data.

Further, an important reason behind using LSTM is that its performance is affected by the size of the data. Therefore, the main idea of the proposed model is to train a single model for all the considered data. The model is trained using long history to have the ability to predict new smart meters that were not used in the training process.

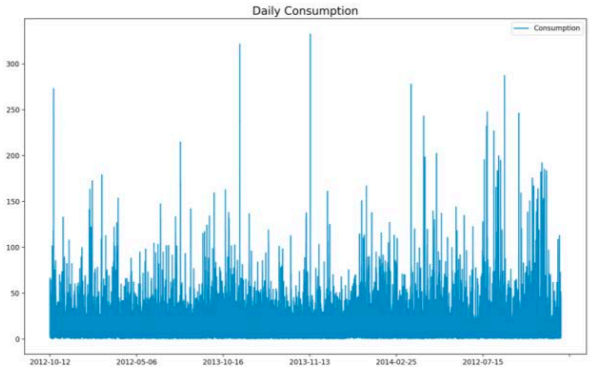

Fig. 3. Average Daily Electricity Consumption of London Households.

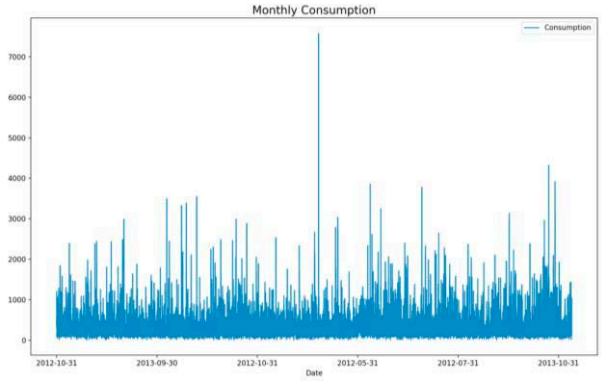

Fig. 4. Average Monthly Electricity Consumption of London Households.

To perform the electricity consumption prediction, firstly we set LSTM to do the time series forecasting with the historical data. Then, for detecting anomalies, the difference between the actual value and the predicted value will be calculated. A consumption is classified as an anomaly if its error is above a selected threshold. The threshold is selected intuitively depending on the density of the errors, as it will be clarified later.

To implement LSTM, the TensorFlow tool is used which is an open-source library developed to handle large datasets, optimization algorithms, and automatic differentiation. First, since the used data is time-series data, 3-dimensional units are defined to be used in the LSTM. The first dimension is the number of samples which is, in this case, the number of daily reading samples of the training set. The second dimension is the time steps that are used to forecast the current day (t) by given historical consumption and weather information at the prior time step, 30 in daily electricity consumption case and 15 in monthly electricity consumption case. The third and last dimension is the features which are the meteorological variables that it is defined earlier (like the weather conditions).

Next, the parameters of the model are adjusted as the following, the model is built with two LSTM layers. The first layer has 100 neurons and the second layer has 32 neurons. The 'relu' activation function is used to convert the output of each unit to be an input for the next layer. Further, a dropout layer is added with a rate of 20% to avoid overfitting. Finally, a dense layer is added with one unit to provide the corresponding forecast.

Immediately by designing the topology of the network, it compiles by defining the optimization algorithm and the loss function. the mean absolute value is selected for the loss function and Adam optimizer as the optimization algorithm. Indeed, all the previous parameters have been chosen after testing multiple network topologies. Once the network has been compiled, the associated weights are fitted which is a very expensive step in the methodology from a computational point of view. Then, the trained model is used to predict one step ahead for all the desired smart meters in the future.

The model needs to identify the number of epochs that indicate the number of passes of the entire training dataset. We select 40 epochs. Every epoch will be divided into a fixed-sized number from the training set, namely batch. In the daily

data, a batch size of 5000 days is selected since it is a big dataset. Hence, every epoch has 425 batches. However, for the monthly data, a batch size of 1000 days is selected, therefore, every epoch has 74 batches.

IV. Results and Discussion

In this section, a summary of the main numerical results obtained from our proposed model in both daily electricity consumption and monthly electricity consumption is provided.

A. Electricity Consumption Forecasting

The forecasting model is designed to work with LSTM on both daily and monthly electricity consumption for each household. To explore the forecasting accuracy, two evaluation metrics are used are the root mean squared error (RMSE) and the mean absolute error (MAE) which can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}\left(Y - \hat{Y}\right)^2} \quad (1) \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y - \hat{Y}i\right| \qquad (2)$$

In general, the smaller values of the performance metrics provide higher forecasting accuracy. Table I presents the forecasting results obtained from the proposed model and compares it with other works presented in [28], [15], [29] and [30], where the same dataset is shared. However, the used dataset is a huge dataset and its size is about 11 GB. Therefore, every research used part of it based on their needs, as illustrated in the second column of Table I. In this work, daily electricity consumption and monthly electricity consumption datasets are used to explore how the performance will be affected when using different network topologies in the same dataset. As it is clear from Table I, the proposed dataset and the dataset used in [28] are the most complicated and longest since we have not excluded any reading from the original dataset. However, in [28], they used only three weather conditions which are humidity, wind speed, and temperature whereas, in this work, all the weather conditions are used without excluding any influencing factor. The lowest RMSE value is achieved from [28] where it is equal to 0.05. On the other hand, the worst RMSE value obtained from [29] where is 3.35. To the research that used MAE performance measure, the lowest MAE was achieved by [16] when forecasting one step-ahead where they got MAE of 19%, whereas the worst MAE was obtained from the same research when forecasting 48 steps-ahead where they got MAE of 25.6%. Although the proposed model does not provide the best result when compared with other works, it got the lowest accuracy result when compared with other compared network topologies, as shown in the 12 and 13 columns in Table I.

To give a graphical representation of the daily electricity consumption model's results, Fig. 5 illustrates the forecasted consumption in comparison with the original consumption in both training and test data. As it is clear from the figure, there is some difference between the original values and the predicted ones. This might be because of the different consumption habits of every household which increase and decrease without a clear pattern, as illustrated earlier in Fig. 3.

Though the proposed model appears as it did not forecast the daily electricity consumption perfectly, Fig. 6 can approve that the electricity consumption pattern forecasted well. In Fig. 6, the plot of test loss drops below training loss which means the model overcomes the overfitting problem and learned perfectly. Moreover, the proposed model achieves RMSE value of 0.362 and MAE of 19.7% which are the lowest values achieved when compared with other network topologies used in Table I.

After implementing the forecasting model for the monthly electricity consumption dataset, Fig. 7 shows the graphical representation of the true monthly electricity consumption and the forecasted values for both the training and testing process.

As it is clear, the graphical representation forecasting result of the monthly electricity consumption is better than the graphical representation of the daily electricity consumption prediction in Fig. 5. This approves the assumption that the daily electricity consumption contains a lot of high and low picks in the consumption which is contrary to the monthly electricity consumption data that run mostly in the average pattern.

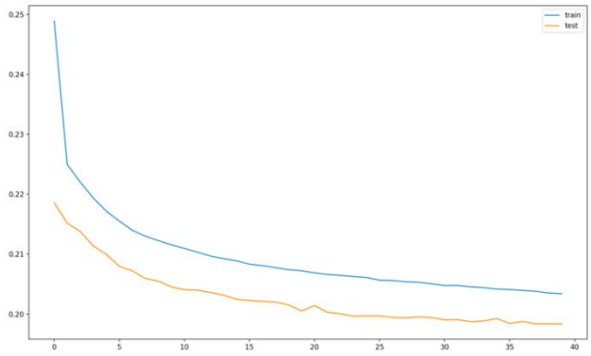

Fig. 5. Original Daily Electricity Consumption Data and Prediction Results of LSTM.

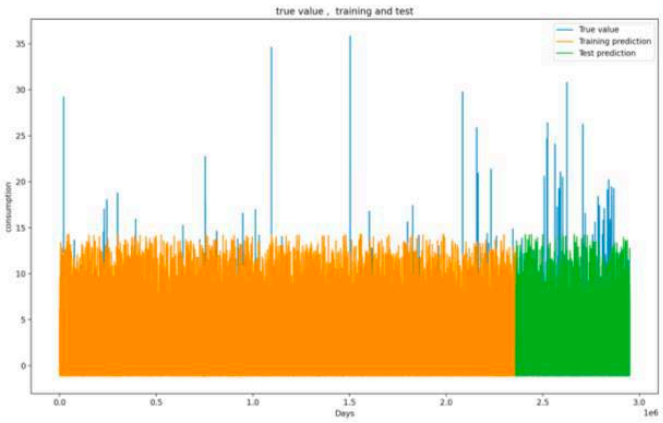

Fig. 6. The Training and Testing Loss of Daily Electricity Consumption Prediction for 40 Epochs.

TABLE I. A COMPARISON OF THE FORECASTING RESULTS OF DIFFERENT NETWORK TOPOLOGIES

Paper	Used data	Batch size	number of epochs	Optimizer	LSTM layers	LSTM units for each layer	Activation function	Loss function	Forecasting type	Performance	Proposed daily dataset performance	Proposed monthly dataset performance
[28]	All buildings with parameters of humidity, wind speed, and temperature.	-	-	Adam	2	32 for each layer	-	MAE	Forecasting one week ahead	RMSE: 0.050	RMSE: 0.492	RMSE: 0.37
[15]	50 households from 16 Acorn of the year of 2013	-	100	-	3	32 for each layer	tanh	-	Forecasting m steps ahead where m = [1, 2, 12, 24, 48]	m: 1 — MAE: 19.2 % — m: 1 — MAE: 21% — m: 1 — MAE: 18.2% m: 2 — MAE: 21.7 % — m: 2 — MAE: 23.4% — m: 2 — MAE: 23.5% m: 12 — MAE: 23.5 % — m: 12 — MAE: 27.3% — m: 12 — MAE: 46.3% m: 24 — MAE: 24.2 % — m: 24 — MAE: 29.94% — m: 24 — MAE: 49% m: 48 — MAE: 25.6 % — m: 48 — MAE: 33.94% — m: 48 — MAE: 46.6%		
[29]	Half hourly consumption data of 500 days for 112 households	1	50	-	4	-	sigmoid	MSE	-	RMSE: 3.35	RMSE: 0.45912	RMSE: 0.509
[30]	Hourly consumption data for 3891 households of the year 2013	1000	40	Adam	2	32 units for the first layer and 16 units for the second layer	tanh	MAE	Forecasting 24-hours ahead	MAE: 0.04	MAE: 0.206	MAE: 0.18
Proposed model	Daily electricity consumption data	5000	40	Adam	2	100 for the first layer and 32 for the second layer	RelU	MAE	Forecasting one day ahead	MAE: 19.7% RMSE: 0.362	-	-
Proposed model	Monthly electricity consumption data	164	40	Adam	2	100 for the first layer and 32 for the second layer	RelU	MAE	Forecasting one day ahead	MAE: 17.8% RMSE: 0.376	-	-

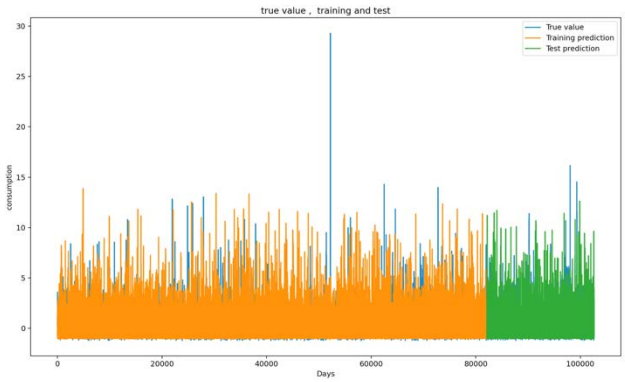

Fig. 7. Original Monthly Electricity Consumption Data and Prediction Results of LSTM.

Further, the loss plot of the daily electricity consumption data and the loss plot of the monthly electricity consumption data drop below training loss which means the model overcomes the overfitting problem and learned well for both datasets, as shown in Fig. 6 and Fig. 8.

Besides, the model of the monthly electricity consumption data achieves RMSE value of 0.376 and MAE of 17.8%. The RMSE value of the monthly electricity consumption is higher than the RMSE value achieved from the daily electricity consumption model. However, The MAE value for the monthly electricity consumption model is lower than the MAE of the daily electricity consumption model.

B. Anomalies Detection

In the previous section, a model has been built to accurately forecast electricity consumption. Now, this model can be used to identify the anomalies in all considered data. The main idea is to forecast the electricity consumption at time t. Then, the difference between the actual value and the predicted value is calculated. A consumption is classified as an anomaly if its error is above a selected threshold. The threshold is selected intuitively depending on the density of the errors, as shown in Fig. 9 and Fig. 10. Fig. 9 illustrates the plots of the mean absolute error values of the daily electricity consumption data. Here, as shown in the figure the density of the errors is around zero and it is going up to 25. The large errors of the daily electricity consumption forecasting occur because of metering the different daily consumption habits of every household.

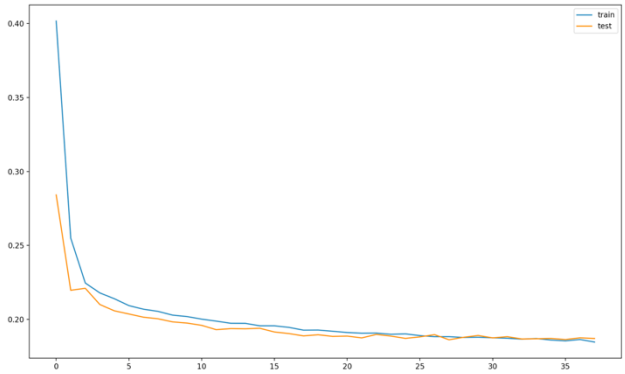

Fig. 8. Training and Testing Loss of Monthly Electricity Consumption for 40 Epochs.

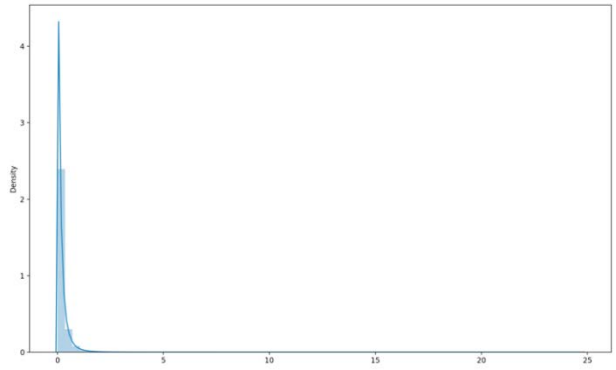

Fig. 9. Mean Absolute Error between the Original Consumption's Value and the Predicted Value of the Daily Electricity Consumption Data.

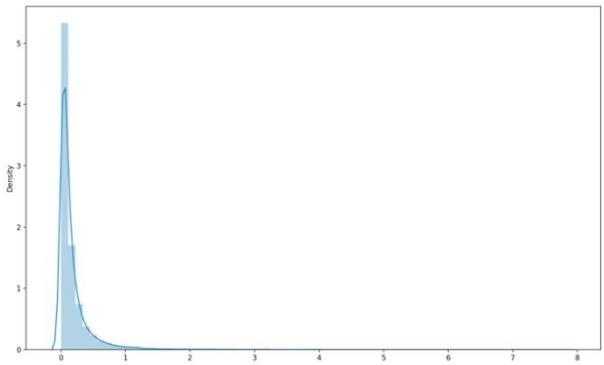

Fig. 10. Mean Absolute Error between the Original Consumption's Value and the Predicted Value of the Monthly Electricity Consumption Data.

Further, the density of the error of the monthly electricity consumption data is around zero; however, it is going up to 8 which is much lower than the errors in the daily consumption forecasting, as illustrated in Fig. 10.

After trying several attempts and adjustments, the proper threshold found is 7 for the daily electricity consumption data and 3 for the monthly electricity consumption data. Fig. 11 and Fig. 12 show the anomalies detected in the daily electricity consumption and the monthly electricity consumption, respectively. The blue line is the consumption, and the red dots are the anomalies.

Generally, the consumption is concerned as an anomaly because of the unexpected trend changes in the data like the sudden increase of the consumption that is different from the normal consumption.

As you can see from Fig. 11 and Fig. 12, it is clear that the model can detect the anomalies efficiently in both daily electricity consumption data and monthly electricity consumption data. However, the daily electricity consumption readings are way better to detect anomalies than the monthly electricity consumption readings because of the different picks that appear in the data. It has been already assumed that in Fig. 3 and Fig. 4 the daily readings will impact the forecasting results and that was true. The daily smart metering has larger errors between the true values and the predicted values than the monthly smart metering which helps detect the anomalies in the data.

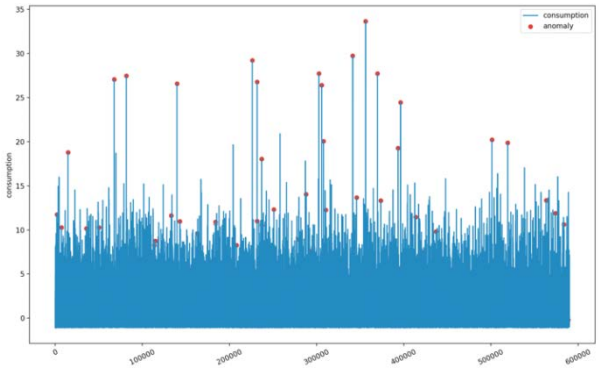

Fig. 11. Detected Anomalies in Daily Electricity Consumption Data.

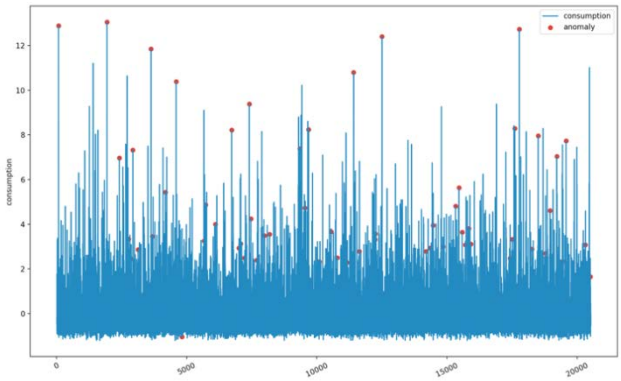

Fig. 12. Detected Anomalies in Monthly Electricity Consumption Data.

To conclude, both daily readings and monthly readings provide similar forecasting performance, however, using daily readings can provide detailed data for households more than the monthly readings. Therefore, using daily metering provide effective results for detecting the anomalies.

V. Conclusion

This work is focusing on detecting anomalies in electricity consumption data by using the long short-term memory (LSTM) approach. The anomalies are identified in two steps: forecasting future consumption and thus anomalies detection. The proposed model is tested using a large real-world dataset with thousands of households segregated into daily consumption and monthly consumption to explore how these may impact the forecasting accuracy of the model. Since the used dataset is huge and published publicly, many research used part of it based on their needs. In this work, we did not exclude any information from the dataset. Instead, the average daily consumption and the average monthly consumption are calculated for comparison purposes.

In conclusion, the proposed model got the lowest accuracy result when compared with other network topologies. The lowest RMSE achieved from other topologies is 0.37 and the lowest MAE is 18% where the proposed model achieved RMSE of 0.362 and MAE of 17.8%. Moreover, both daily and monthly readings have similar forecasting performance; however, the daily readings provide more detailed data for households than the monthly readings. Therefore, using daily metering provides effective results to detect anomalies.

VI. Future Work

In this work, the used dataset is public electricity consumption data. In the future, we aim to collect the electricity consumption data of Saudi Arabia's buildings. Furthermore, the weather information of the collected years will be added. Finally, there is a plan to construct an efficient energy management system that identifies the anomalies in daily real-time buildings' electricity consumption.

Acknowledgment

The authors would like to thank Qassim University for supporting this research.

References

[1] L. Pérez-Lombard, J. Ortiz and C. Pout, “A review on buildings energy consumption information,” *Energy Build*, vol. 40, no. 3, pp. 394-398, 2008.

[2] Z. Liu, “ Global energy development: the reality and challenges,” in *Global Energy Interconnection*, Academic Press, pp. 1–64, 2015.

[3] Y. Heo, R. Choudhary and G. A. Augenbroe, “Calibration of building energy models for retrofit analysis under uncertainty,” *Energy Build*, vol. 42, pp. 550-560, 2012.

[4] K. Bandara, C. Bergmeir, and S. Smyl, “Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach,” arXiv preprint arXiv:1710.03222, 2017.

[5] [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

[6] [Online]. Available: https://www.kaggle.com/jeanmidev/smart-meters-in-london.

[7] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamshirband and A. Varkonyi-Koczy, “State of the art of machine learning models in energy systems, a systematic review,” *Energies*, vol. 12, no. 7, 2019.

[8] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo and P. Chatellier, “Modeling and forecasting building energy consumption: A review of data-driven techniques,” *Sustainable Cities and Society*, vol. 48, pp. 101-533, 2019.

[9] U. Farouk, M. Asante and J. Ben, “A survey of machine learning’s electricity consumption models,” *Communications on Applied Electronics*, vol. 7, no. 21, pp. 6–10, 2018.

[10] E. García-Martín, C. F. Rodrigues, G. Riley and H. Grahn, “Estimation of energy consumption in machine learning,” *Journal of Parallel and Distributed Computing*, vol.134, pp. 75-88, 2019.

[11] C. Li, Z. Ding, D. Zhao, J. Yi and G. Zhang, “Building energy consumption prediction: An extreme deep learning approach,” *Energies*, vol. 10, no. 10, pp. 1–20, 2017.

[12] T. M. Ghazal, S. Noreen, R. Said, M. Khan, S. Siddiqui, S. Abbas, S. Aftab and M. Ahmad, “Energy demand forecasting using fused machine learning approaches,” *Intelligent Automation and Soft Computing*, vol. 31, no. 1, pp. 539–553, 2022.

[13] J. Y. Kim and S. B. Cho, “Electric energy consumption prediction by deep learning with state explainable autoencoder,” *Energies*, vol. 12, no. 4, 2019.

[14] Y. Jin, D. Yan, X. Zhang, M. Han, X. Kang, J. An and H. Sun, “District household electricity consumption pattern analysis based on auto-encoder algorithm,” *IOP Conference Series: Materials Science and Engineering*, vol. 609, no. 7, pp. 072-028, 2019.

[15] P. A. Schirmer, I. Mporas and I. Potamitis, “Evaluation of regression algorithms in residential energy consumption prediction,” in *3rd European Conference on Electrical Engineering and Computer Science, EECS 2019*. pp. 22–25, 2019.

[16] P. W. Khan and Y. C. Byun, “Adaptive error curve learning ensemble model for improving energy consumption forecasting,” *Computer, Material and Continua*, vol. 69, no. 2, pp. 1893–1913, 2021.

[17] F. Z. Abera and V. Khedkar, “Machine learning approach electric appliance consumption and peak demand forecasting of residential

customers using smart meter data," *Wireless Personal Communications*, vol. 111, no. 1, pp. 65–82, 2020.

[18] A. H. Neto and F. A. S. Fiorelli, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," *Energy Build.*, vol. 40, no. 12, pp. 2169–2176, 2008.

[19] E. Y. Shchetinin, "Modeling the energy consumption of smart buildings using artificial intelligence," *CEUR Workshop Proc.*, vol. 2407, pp. 130–140, 2019.

[20] X. M. Zhang, K. Grolinger and M. A. M. Capretz, "Forecasting residential energy consumption using support vector regressions," in *Proc. IEEE Inter. Conf. Mach. Learn. Appl.*, pp. 1–10, 2018.

[21] S. Aman, C. Chelmis, and V. K. Prasanna, "Learning to REDUCE: A reduced electricity consumption prediction ensemble," *AAAI Work. - Technical Report*, vol. WS-16-01-, pp. 204–210, 2016.

[22] J. Eisses, "Anomaly detection in electricity consumption data of buildings using predictive models," *University of Amsterdam*, pp. 1-7, 2014.

[23] K. Hollingsworth, K. Rouse, J. Cho, A. Harris, M. Sartipi, S. Sozer, B. Enevoldson, "Energy anomaly detection with forecasting and deep learning," in *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data* , pp. 4921–4925, 2019.

[24] A. Maamar and K. Benahmed, "A hybrid model for anomalies detection in ami system combining k-means clustering and deep neural network," *Computer, Material and Continua*, vol. 60, no. 1, pp. 15–39, 2019.

[25] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132-306, 2020.

[26] R. DiPietro and G. D. Hager, "Deep learning: RNNs and LSTM," *Handbook of Medical Image Computing and Computer Assisted Intervention*. pp. 503–519, 2019.

[27] S. Hochreiterf, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 6, no. 2, pp. 107-116, 1998.

[28] T. T. Q. Nguyen, T. P. T. Tran, V. Debusschere, C. Bobineau and R. Rigo-Mariani, "Comparing high accurate regression models for short-term load forecasting in smart buildings," in *Proc. (Industrial Electron. Conf.*, pp. 1962–1967, 2020.

[29] D. Kaur, R. Kumar, N. Kumar and M. Guizani, "Smart grid energy management using RNN-LSTM: A deep learning-based approach," in *proc. 2019 IEEE Global Communications Conference*, pp. 1-6, 2019.

[30] A. M. Alonso, F. J. Nogales and C. Ruiz, "A single scalable lstm model for short-term forecasting of massive electricity time series," *Energies*, vol. 13, no. 20, 2020.