Natural and Applied Sciences International Journal (NASIJ)
eISSN: 2788-4619 (online)
https://www.ideapublishers.org/index.php/nasij
Vol. 4, No. 1 (January-June), 2023, pp. 164-176
https://doi.org/10.47264/idea.nasij/4.1.11 Research Article

Multivariate time series forecasting for electricity consumption using machine learning methods

Hadiqa Basit[1] | Nadia Mushtaq[1] | Shakila Bashir*[1] | Angela Melgarejo Morales[2]

1. Department of Statistics, Forman Christian College (A Chartered University), Lahore, Pakistan.
2. Institute of Geophysics, National Autonomous University of Mexico, Morelia, Michoacan, Mexico.

* Corresponding Author Email: shakilabashir@fccollege.edu.pk

Received: 20-Apr-2023 | Revised: 25-Jun-2023 | Accepted: 26-Jun-2023 | Published: 30-Jun-2023

Abstract

Multivariate time-series forecasting plays a crucial role in many real-world applications. Recently, multiple works have tried to predict multivariate time series. In this paper, different aspects of electricity consumption within a household-based in Lahore real data have been used to make one-hour-ahead forecasts for overall usage. In this study, various Neural Networks (NNs) such as the Long Short-Term Memory (LSTM) network, Recurrent Neural Network (RNN) and the Gated Recurrent Unit (GRU) network are used with varying numbers of hidden layers to make multivariate time series analysis and predictions. This study aims to express a clear and precise method for multivariate time series. The models make predictions based on data sets and are trained on past data. Their performance is evaluated using root mean squared error. Their performance was compared, and results are given for the one-hour-ahead forecasts for electricity consumption using machine learning models. In the dynamic field of forecasting electricity use, the study further investigates the possible integration of real data to improve the prediction capacities of machine learning models using Python software. The results show that the RNN performs better than the other two models for the given data.

Keywords: Neural Networks, Long Short-Term Memory, Recurrent Neural Network, Gated Recurrent Unit, Multivariate Time Series, real data, electricity consumption.

How to Cite:
Basit, H., Mushtaq, N., Bashir, S., & Morales, A. M. (2023). Multivariate time series forecasting for electricity consumption using machine learning methods. *Natural and Applied Sciences International Journal (NASIJ), 4*(1), 164-176. https://doi.org/10.47264/idea.nasij/4.1.11

1. Introduction

The importance of predicting electricity consumption stems from the problem of limited electricity resources in Pakistan, prompting the need to study electricity usage within an average household. Using forecasting methods to predict the electricity needs of an average household will allow for better allotment of resources e.g. saving energy when the forecasts show a decline in usage and preparing for an increase in consumption to avoid a shortage, which subsequently leads to load shedding that affects the quality of life for citizens. Multivariate time series forecasting is the most important method for addressing the challenges brought on by the unpredictable electricity consumption pattern. Smyl (2020) proposed that the hybrid machine learning method produces more accurate forecasts than those generated by either pure statistical or pure machine learning approaches. Smith (2019) discussed machine learning applications in electricity consumption forecasting.

Forecasting has not been restricted to the overall usage within the household; rather, consumption within different parts of the house has been factored in to allow the prediction of overall consumption in light of how electricity is being used to make a more informed forecast. For this purpose, it was necessary to opt for multivariate procedures. Traditional univariate time series forecasting techniques frequently fail to capture the complex relationships and interplay between the several elements influencing the amount of electricity consumed. Explainable machine learning algorithms have not yet been used in mid-term aggregated load forecasting research. Yaprakdal and Ansoy (2023) examined the applications of Existing Explainable Artificial Intelligence (XAI) studies that have concentrated on short-term load forecasting at the building level. The objective is to improve prediction reliability and precision by considering various relevant characteristics. This will allow stakeholders to make well-informed decisions on electricity consumption strategy and resource allocation. The study by Lee (2022) aimed to predict power usage by utilizing time series data and various artificial intelligence and metaheuristic techniques.

Neural Networks (NNs) are algorithms that take training data as input to learn from them and produce an output. In the context of multivariate time series forecasting, NNs take time series data as the input and produce forecasts as the output. In a discussion of Multivariate Exponential Smoothing Long Short-Term Memory (MES-LSTM) on multiple aggregated coronavirus disease of 2019 (COVID-19) morbidity datasets, Mathonsi and van Zyl (2022) demonstrated the consistency and significance of the hybrid technique.

This article aims to draw insight into the usage of electricity consumption of real data collected from June 2018 to 2019 as the basis for modelling electricity consumption using machine learning. The objective is to identify which method, together with the ideal input variables and parameter combinations, performs better than others in specific electricity consumption. We developed predictive models for multivariate time series data of electricity usage such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and Vector Auto Regressive (VAR). The machine learning models could be useful in benchmarking the electricity consumption of households and identifying opportunities to improve energy efficiency.

This paper is divided into the following sections: Section 2 shows the literature review, Section 3 describes the nature of the data and how it was preprocessed for forecasting, Section 4

explains the theory behind the various algorithms being used and the performance measure, Section 5 details the application of the algorithms on the data and discusses the results, Section 6 makes comparisons of the predictions using the performance measure and Section 7 concludes the paper.

2. Literature review

Several studies have also used machine learning techniques to forecast electricity consumption. Bezzar et al. (2022) use the Extreme Gradient Boosting Algorithm (XGBoost) to forecast univariate time series data, and Le *et al.* (2020) employ Long Short-Term Memory (LSTM) and the k-means clustering algorithm to forecast multivariate time series data. In addition, smart meters are used to collect data. Kim *et al.* (2023) used smart meter data for univariate time series forecasting. Decomposition techniques for multivariate time series forecasting for electricity consumption in Pakistan were discussed by Iftikhar *et al.* (2023). Goel *et al.* (2016) performed a multivariate time series modelling with real flight data using VARs and LSTMs and drew comparisons.

Wan *et al.* (2019) use deep learning models based on Recurrent Neural Networks (RNNs) and convolutional neural network CNN to conduct multivariate time series prediction for PM 2.5 levels in Beijing. Gonzalez-Vidal. (2019) carry out multivariate time series forecasting for energy data obtained through smart monitoring in which the random forest algorithm is used for 1-step-ahead, 2-step-ahead and 3-step-ahead forecasts. Ruiz *et al.* (2021) compare various multivariate time series algorithms to assess which one solves the Multivariate Time Series Classification (MTSC) problem the best. Che *et al*. (2018) develop a deep learning model based on GRU for multivariate time series data with missing values. Kanchymalay *et al.* (2017) discussed machine learning techniques, such as Support Vector Regression (SVR) and Sequential Minimal Optimization, which are used to create predictions for crude palm oil pricing.

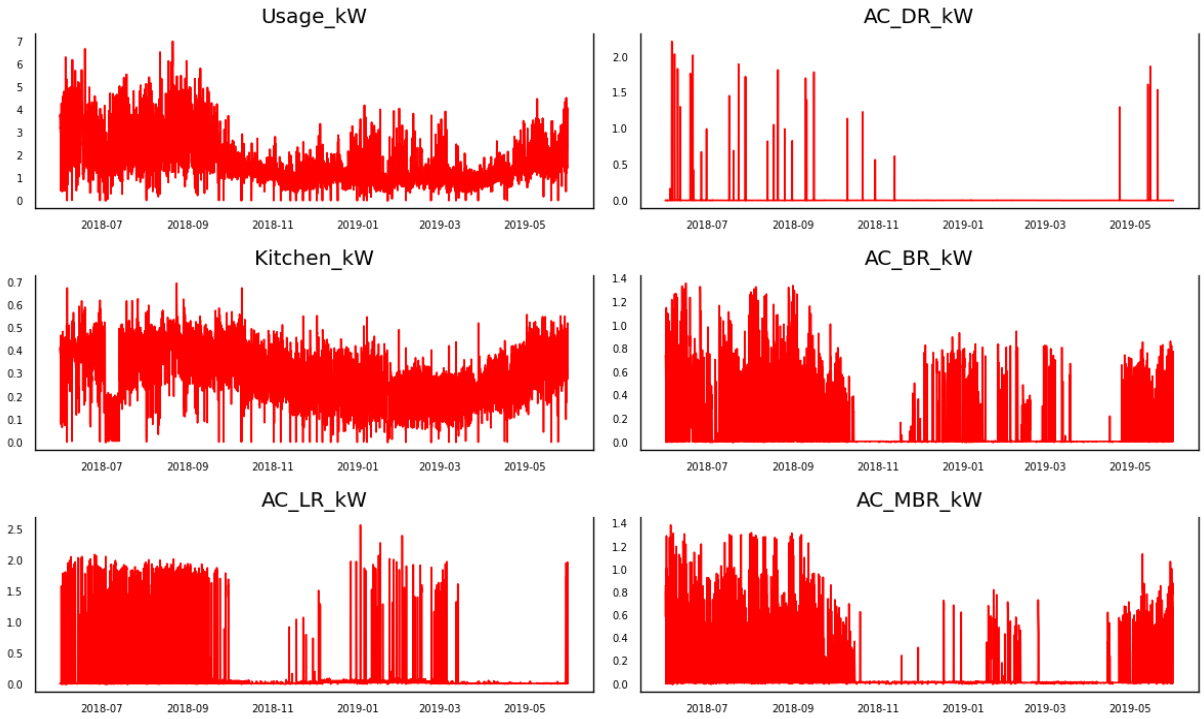
Chen and Sun (2022) propose a Bayesian multidimensional time series prediction framework. Sagheer and Kotb (2019) improve upon the existing LSTM model by using an unsupervised learning approach and developing a pre-trained LSTM-based Stacked Autoencoder (LSTM-SAE). Mishra *et al.* (2023) covered the topic of multivariate time series short-term forecasting utilizing cumulative coronavirus data. Sharma *et al.* (2023) examined the use of deep ensemble learning methods in analyzing and predicting COVID-19 multivariate data. From the review of the available literature, we can conclude that multivariate time series forecasting of electrical consumption can be accurately achieved using machine learning approaches.

3. Data description

3.1. Data collection

This data was published on the Open Data website (Nadeem & Arshad, 2019). It uses a smart meter to monitor the minute-to-minute electricity consumption of the overall usage (Usage_kW), the usage in the kitchen (Kitchen_kW), and the electricity consumed by air conditioners in the drawing room (AC_DR_kW), bedroom (AC_BR_kW), living room (AC_LR_kW) and the main bedroom (AC_MBR_kW) as shown in Figure 1. The data was collected starting from 6/1/2018 to 5/31/2019.

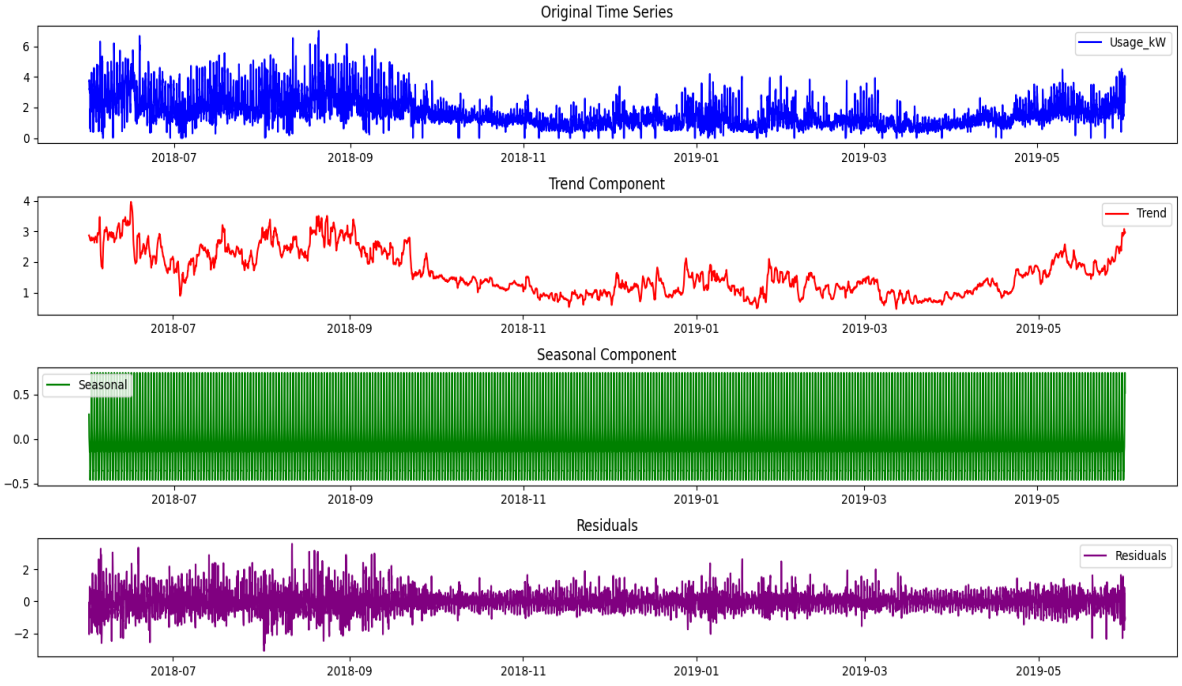
Figure 1: Data description



3.2. Data decomposition

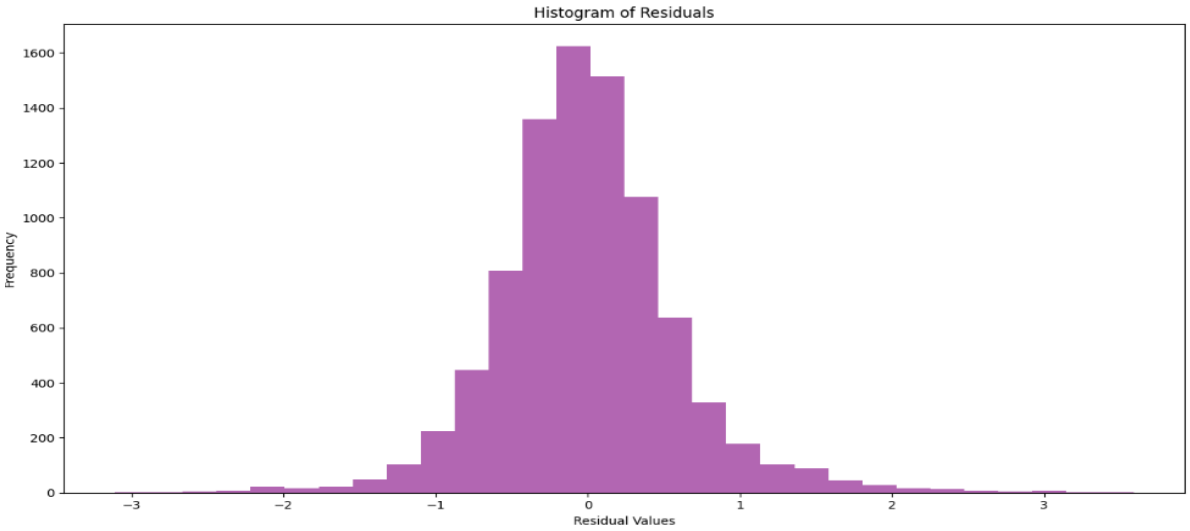
Since the variable of interest is the overall electricity usage (represented by Usage_kW), the time series data was decomposed to understand its nature better. The results are shown in Figure 2.

Figure 2: Decomposition of data sets



From Figure 2, it can be seen that the data is stationary and does not have a seasonal component. Below, Figure 3 shows the residual plots having a Gaussian distribution.

Figure 3: Residual plot



3.3. Data pre-processing

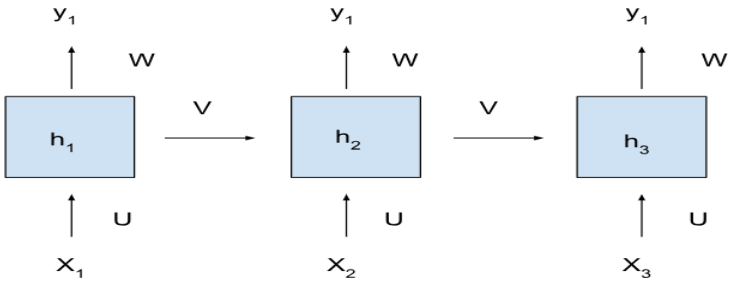
Hourly averages were taken for the entire data set to change the frequency of the data from minutes to hours. Next, to use the data for Neural Networks (NNs), the data was transformed using the sigmoid function, i.e., it was scaled between 0 and 1. The data was then transformed into a supervised learning set where the lag was set to 1 so that the consumption value for the previous hour was used to predict the value for the next hour. In other words, the algorithm gives one-step-ahead predictions. Finally, the data was divided into training and testing sets where the hourly data for 310 days was used to predict the data for the remaining 55 days, which makes up the test set.

4. Theoretical framework

4.1. Recurrent Neural Network (RNN)

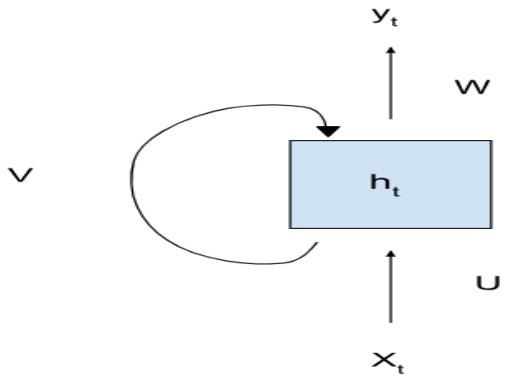
RNN is a type of Artificial Neural Network (ANN) that processes sequential data. The directionality of the inputs does not matter in this type of NN, nor does the variable sequence length. RNNs have loops in their architecture that allow them to carry time series information across time steps. The unfolded version of the loop is given in Figure 4 below.

Figure 4: RNN algorithm



Here, h_1, h_2, and h_3 are hidden states at the 1st, 2nd and 3rd timestamps, respectively. The X's and y's are the inputs and outputs at each step. U, W and V are linear transformations done on the values at each step. U converts the matrix into an array, and W is an activation function; in this paper, it is a sigmoid function that gives information about the next value, and V multiplies the loss into the next step. The folded version of the loop is shown in Figure 5.

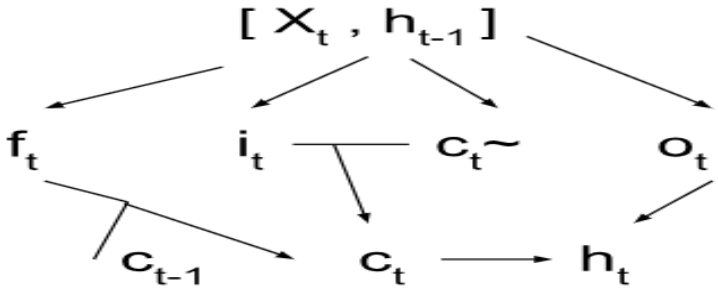
Figure 5: Folded version of the loop



4.2. Long Short-Term Memory (LSTM)

The LSTM model is a complicated form of RNN that is particularly useful since it retains information over an extended period of time. The loop architecture of LSTM is given as: one reason why LSTM may be preferred over RNN is due to the fact that as the values of V multiply, either of two things can happen. The first would be that if the values of V are less than 1, V will vanish; secondly, if the values are more significant than 1, V will explode, making it hard for past values of RNN to impact future predictions. LSTM retains information over a long period of time by introducing gates that select or forget information that impacts the predictions. Figure 6 explains how the gates work.

Figure 6: LSTM algorithm



Here, X_t represents the input at a given time step, h_{t-1} is the output of the previous time step, c_t is a cell state which helps retain information over an extended period and both of these values are fed into the four gates, f_t serves as the forget gate which decided what to forget and keep from the last state by squashing values using the sigmoid function, $\tilde{c_t}$ is the new candidate cell state that squashes information between -1 and 1 using the hyperbolic tangent function, i_t is the input gate which decides what to keep and forget from the candidate cell state and o_t is the output gate which decides what to use from the last cell state. This is the underlying framework behind LSTM.

4.3. Gated Recurrent Unit (GRU)

The Gated Recruitment Unit (GRU) model is a simpler form of the LSTM as it has the same functionality of retaining the information over a long period of time, however, it uses fewer parameters.

4.4. Vector Auto Regressive (VAR)

The VAR model is able to use multiple variables to predict the variable of interest. The mathematical form is given in Equation (1).

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t \quad (1)$$

Where, y_t is a p-dimensional vector of variables at time t, c is the intercept, A_1, A_2, …, A_p are coefficient matrices of lagged variables, and e_t is a p-dimensional vector of white noise error terms at time t.

4.5. Performance measure

The Root Mean Squared Error (RMSE) will be used to check the accuracy of the model. The RMSE calculates the square root of the average sum of square distance between the actual and the predicted values. The formula is given in Equation (2).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

Where, n is the number of data points, y_i represents the actual values and $\hat{y}_i$ represents the predicted value.

5. Application

This section will demonstrate the application of the models discussed above. The epochs were set to 50 for all NNs with batch size 72, and Dense was set to 1. The model summary for each showed that with the same amount of data, each model trained the following parameters: GRU 8,751, RNN 2,901, and LSTM 11,451. We can see that RNN has the lowest number of parameters and LSTM has the highest number. This is understandable as LSTM is the complicated version of the original RNN, and GRU is the simplified version of LSTM. The predictions for each model are given in Figure 7, Figure 8, and Figure 9. The actual values and predictions of all models have been developed and used, which allows us to see how each model performs.

Also, in Table-1, Table-2 and Table-3, the RMSEs for different input lengths (5, 10 and 20) with varying numbers of hidden units (32, 64, 128) are given. This shows the performance of the models when different numbers of past values are taken to predict the electricity usage for the next hour. The numbers in Table-1, Table-2, and Table-3 show the performance of the machine learning models for different combinations of input length and hidden units. The lower values of the resulting tables indicate better performance.

Figure 7: Shows the actual and predicted values using RNN

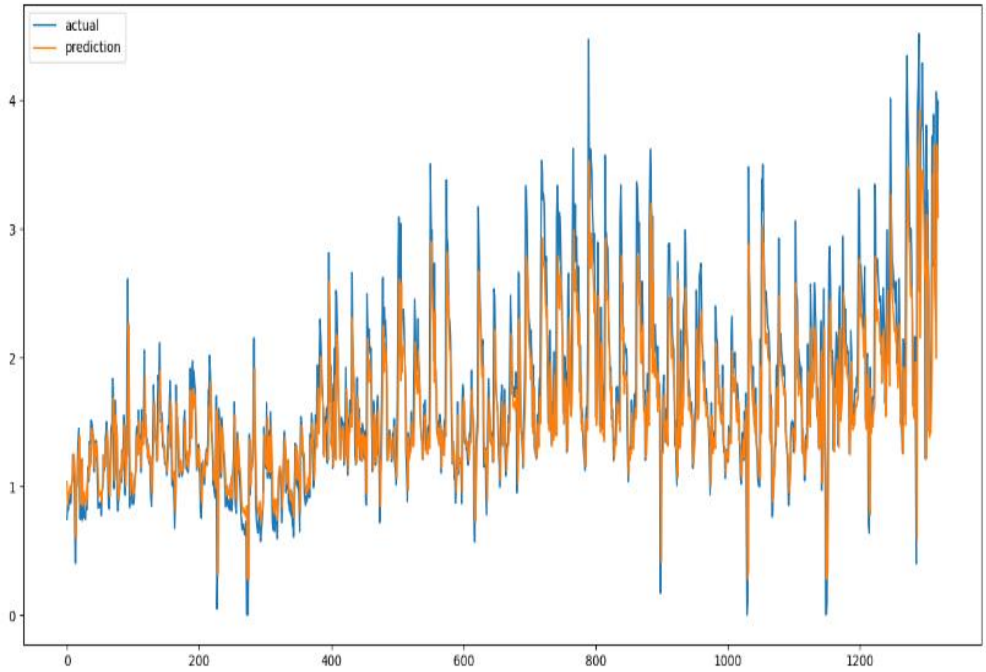


Table-1: Output of RNN

Input	Hidden Units		
Length	32	64	128
5	0.419	0.417	0.439
10	0.416	0.427	0.425
20	0.445	0.419	0.417

Figure 8: Shows the actual and predicted by using LSTM

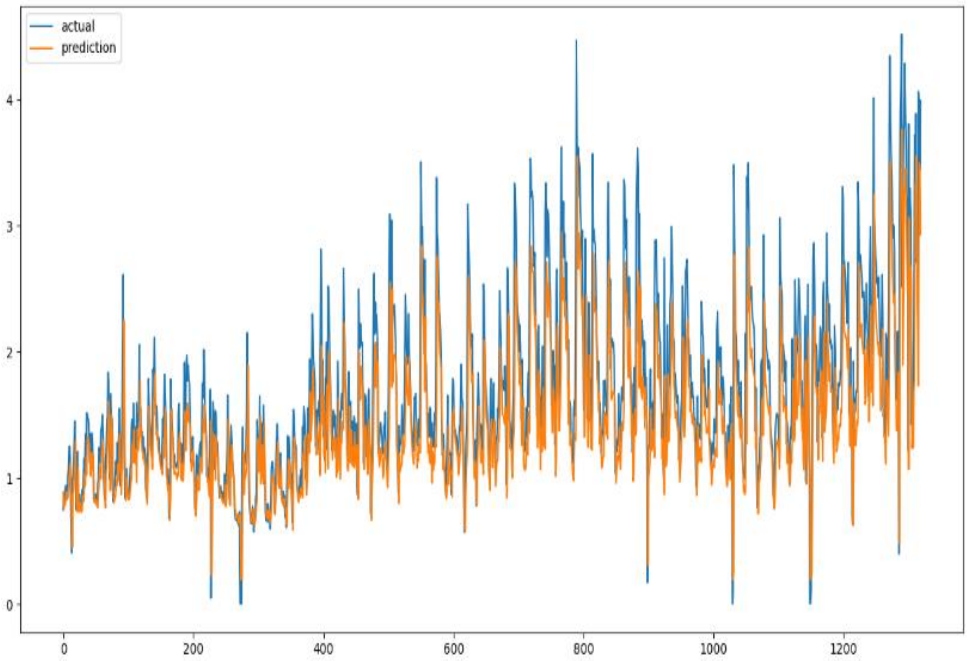


Table-2: Output of LSTM

Input	Hidden Units		
Length	32	64	128
5	0.486	0.489	0.496
10	0.488	0.492	0.497
20	0.485	0.491	0.497

Figure 9: Shows the actual and prediction by using GRU

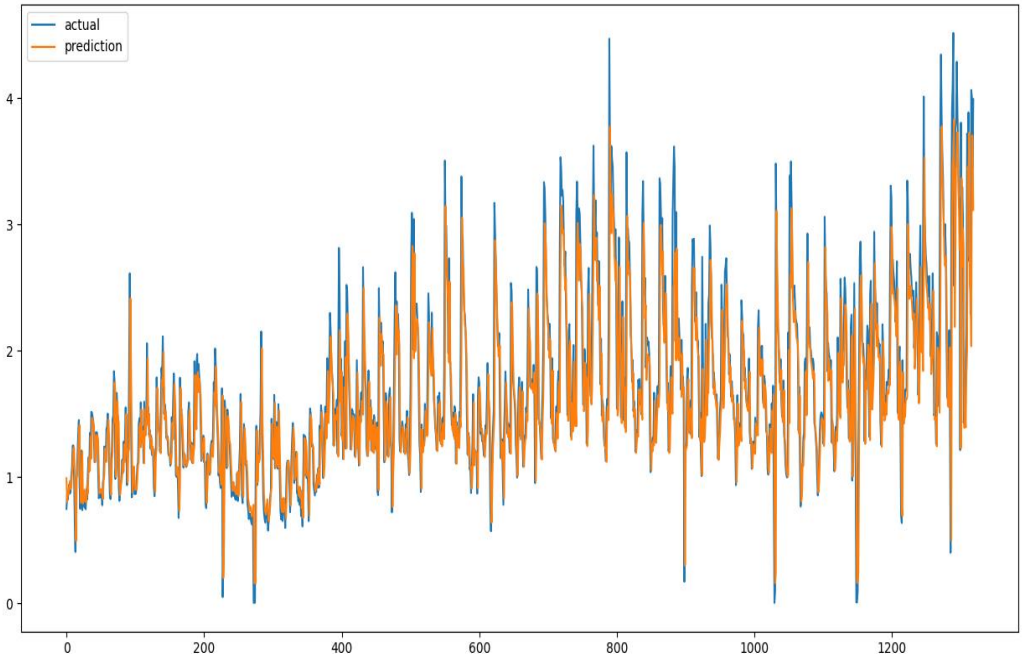


Table-3: Output of GRU

Input Length	Hidden Units		
	32	64	128
5	0.419	0.426	0.424
10	0.426	0.424	0.424
20	0.417	0.430	0.422

6. Results and discussion

After cleaning and pre-processing the data, the houses selected are divided into train and test subsets, respectively. The models are trained using the abovementioned parameters, and the finalized models are used for predicting the test subset. Accurately projecting future power consumption is essential for effective energy management, cost savings, and environmental sustainability, given the rising energy demand. It is important to remember that forecasting electricity consumption is a challenging process that calls for careful consideration of several variables, including seasonality, time of day, and weather. To make accurate forecasts, it is essential to choose suitable models.

Ideally, a predictive should have a low RMSE, and from Table-1, Table-2 and Table-3, we can now compare the accuracy of their performance by looking at the RMSE.

Table-4: Best RMSE of the models

	RNN	LSTM	GRU
RMSE	0.416	0.485	0.417

Table-4 considers RMSE values for RNN, LSTM, and GRU of RNN. RMSE is used to measure prediction accuracy. So, the RMSE of RNN 0.416 is less than among other models and is more accurate in predicting future electricity consumption.

The comparison of the models can be effectively done using RMSE. We can see that the results and predicted values of the models RNN are more accurate and perform well. The results presented in this study and comparing results show that the RNN algorithm is the best model for forecasting electricity consumption.

7. Conclusion

In conclusion of the study, multivariate time series forecasting for electricity consumption has shown promise when advanced machine learning models—namely, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) are applied. An understanding of these models' individual performances can be gained by assessing them using Root Mean Squared Error (RMSE) measurements. The results showed that GRU and RNN are more accurate than other methods. Further work can be done by doing a multidimensional analysis of all the houses in the PRECON dataset, including the overall usage, usage within different parts of the house, and metadata. This paper is based on the idea that knowledge of energy consumption within a household is important. This can help distribution companies understand the needs of their customers better and provide customized tariff rates and better plans for their diverse consumer base. The results show significant findings that several parameters can be predicted with RMSE of RNN, LSTM and GRU as 0.416, 0.485 and 0.417 respectively. As the study progresses, the models can be extended to include commercial and industrial consumers, allowing newer and better energy optimization methods in developing countries. Overall, the results of RNN perform better for future predictions, highlighting the promise of machine learning models for multivariate time series forecasting of electricity usage. These findings support sustainable practices and well-informed decision-making in the dynamic field of electricity consumption. Future studies should concentrate on utilizing machine learning models to predict the usage of electricity in other cities and factory areas.

Declaration of conflict of interest

The author(s) declared no potential conflicts of interest(s) with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Hadiqa Basit https://orcid.org/0009-0006-5876-7900

Nadia Mushtaq https://orcid.org/0000-0002-0652-0029

Shakila Bashir https://orcid.org/0000-0003-4701-6977

References

Bezzar, N. E.-H, Laimeche, L., Meraoumia, A., & Houam, L. (2022). Data analysis-based time series forecast for managing household electricity consumption. *Demonstratio Mathematica, 55*(1), 900-921. https://doi.org/10.1515/dema-2022-0176

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports, 8*(1), 6085. https://doi.org/10.1038/s41598-018-24271-9

Chen, X., & Sun, L. (2021). Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4659-4673. https://doi.org/10.1109/TPAMI.2021.3066551

Goel, H., Melnyk, I., Oza, N., Mathews, B., Banerjee, A. (2016). Multivariate aviation time series modeling: VARs vs. LSTMs. https://goelhardik.github.io/images/Multivariate_Aviation_Time_Series_Modeling_VARs_vs_LSTMs.pdf

Gonzalez-Vidal, A., Jimenez, F., Gomez-Skarmeta, A. F., (2019). A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy and Buildings, 196*, 71-82, https://doi.org/10.1016/j.enbuild.2019.05.021

Iftikhar, H., Bibi, N., Rodrigues, P., Lopez-Gonzales, J., (2023). Multiple novel decomposition techniques for time series forecasting: Application to monthly forecasting of electricity consumption in Pakistan. *Energies, 16*(6), 2579. https://doi.org/10.3390/en16062579

Kanchymalay, K., Salim, N., Sukprasert, A., Krishnan, R., & Hashim, U. R. A. (2017, August). Multivariate time series forecasting of crude palm oil price using machine learning techniques. *IOP Conference Series: Materials Science and Engineering*, 226(1), 012117. https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012117/meta

Kim, H., Park, S., Kim, S. (2023). Time-series clustering and forecasting household electricity demand using smart meter data. *Energy Reports, 9*, 4111-4121, https://doi.org/10.1016/j.egyr.2023.03.042

Le, T., Vo, M., Kieu, T., Hwang, E., Rho, S., & Baik, S. (2020). Multiple electric energy consumption forecasting using cluster-based strategy for transfer learning in smart building. *Sensors, 20*, 2668. http://dx.doi.org/10.3390/s20092668

Lee, M. H. L., Ser, Y. C., Selvachandran, G., Thong, P. H., Cuong, L., Son, L. H., ... & Gerogiannis, V. C. (2022). A comparative study of forecasting electricity consumption using machine learning models. *Mathematics, 10*(8), 1329. https://doi.org/10.3390/math10081329

Mathonsi, T., & van Zyl, T. L. (2021). A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. *Forecasting, 4*(1), 1-25. https://doi.org/10.3390/forecast4010001

Mishra, S., Singh, T., Kumar, M., & Satakshi. (2023). Multivariate time series short term forecasting using cumulative data of coronavirus. *Evolving Systems*, 1-18. https://doi.org/10.1007/s12530-023-09509-w

Nadeem, A., & Arshad, N. (2019). PRECON: Pakistan residential electricity consumption dataset. *e-Energy 19*, 52-57. https://doi.org/10.1145/3307772.3328317

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75-85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Smith, J. A. (2019). Machine learning applications in electricity consumption forecasting. *Energy Forecasting Journal, 12*(3), 45-60. https://doi.org/10.1234/energyforecasting.2019.123456

Sharma, S., Gupta, Y. K., & Mishra, A. K. (2023). Analysis and prediction of COVID-19 multivariate data using deep ensemble learning methods. *International Journal of Environmental Research and Public Health, 20*(11), 5943. https://doi.org/10.3390/ijerph20115943

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification back off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery, 35*, 401–449. https://doi.org/10.1007/s10618-020-00727-3

Sagheer, A., Kotb, M. (2019) Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. *Scientific Reports, 9*, 19038. https://doi.org/10.1038/s41598-019-55320-6

Wan, R., Shuping, M., Wang, J., Liu, M., & Yang, F. (2019). Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics, 8*(8), 876. https://doi.org/10.3390/electronics8080876

Yaprakdal, F., & Arısoy, M. V. (2023). A multivariate time series analysis of electrical load forecasting based on a hybrid feature selection approach and explainable deep learning. *Applied Sciences, 13*(23), 12946. https://doi.org/10.3390/app132312946