Estimation of Hourly Utility Usage Using Machine Learning

Albert Wong
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-0669-4352

Chunyin Chiu
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-5932-539

Abigail Abdulgapul
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-7285-1096

Mirza Nomaan Beg
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-6769-7151

Youry Khmelevsky
Computer Science
Okanagan College
Kelowna BC, Canada
ORCID: 0000-0002-6837-3490

Joe Mahony
Harris SmartWorks
Ottawa Ontario, Canada
JMahony@harriscomputer.com

2022 IEEE International Systems Conference (SysCon) | 978-1-6654-3992-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/SysCon53536.2022.9773816

***Abstract*—The COVID-19 pandemic has had a major impact on the usage of various utilities. To assess the impact, this research explores the (baseline) estimation of hourly utility usage if the pandemic did not happen. Using usage data from Harris SmartWorks, various machine learning algorithms are implemented to show that they are effective in modelling hourly usage patterns, calendar effects, as well as "lingering" effects of the exogenous factors and produce accurate results.**

***Index Terms*—utility usage, time series, machine learning, deep learning applications, big data**

I. Introduction

The COVID-19 pandemic is impacting personal, family, and business environments in general and utility usage in particular. For some municipalities and cities across North America, residential utility usage increased during the lockdown [1], [2]. The higher energy consumption reflects the increased use of computing such as videos streaming and conferencing due to work-from-home and learning-from-home activities, as well as other stay-at-home activities such as food preparation. A study last year has indicated that residential refrigerators are working overtime due to the increased storage of warm leftovers being placed in the appliance [3].

As providers of critical infrastructure, the utility industry plans for many foreseeable hazards, but it is less likely that health emergencies, such as the COVID-19 crisis, are planned for. There is a need to support utility companies in having the appropriate data available for continuity plans that are adaptable to fully address the fast-moving and unknown variables of an outbreak such as COVID-19. Therefore, it is important to quantify the impact of COVID-19 on utility usage to fulfill this business need.

In this paper, we present the approach to measuring the impact of COVID-19 by estimating, using historical data up to March 2021, utility usage during the pandemic period if COVID-19 did not happen. These baseline estimates could then be compared with the corresponding actual usages during the pandemic periods to arrive at estimates of the pandemic impact. Simply put, the impact of COVID-19 on usage could be estimated as the difference between the actual usage and the estimated (baseline) usage derived from the historical data before the pandemic.

The approach of estimating the baseline usage using historical data was implemented through a number of machine learning models and on several utility data sets in electricity, water, gas, and steam. Given that results from these data sets were similar, we will, in this paper, focus on the work with one particular electricity data set in the United States.

Note that the work presented here was conducted as part of the applied research and capstone projects at Okanagan and Langara Colleges by faculty and students with support from industry [4]–[6], [6]–[9], [9]–[21], [21]–[24].

II. Literature Review

In the past, many studies use the traditional modelling tools, such as Autoregressive Integrated Moving Average (ARIMA) models, to produce predictions for a time series. This approach has been applied in many fields such as medicine, climate research, and energy consumption research [25]–[29]. These studies used trends and patterns of the time series of interest to produce predictions for the future. Some articles also used the combination of a time series model with other statistical techniques, for example, integrating the time series models with smoothing techniques, for the development of forecasts [30], [31].

Apart from the traditional time series analysis, the employment of a regression model is another option [28], [29]. A regression model allowed researchers to take relevant predictors of the forecast values into account. However, it is challenging, from a modelling standpoint, to incorporate historical trends

Supported by Harris Utilities (Harris) and NSERC Grant "A novel approach to COVID-19 Impact Analysis and Reporting for Utilities, 2020-2021.

and patterns of the predicted variable as independent variables in a regression setting.

Recently, researchers have started to use various machine learning algorithms for time series analysis. For example, the Support Vector Regression (SVR) algorithm was used to forecast individual electricity consumption [32]. A number of researchers also used ensemble methods, especially the boosting algorithms, to create predictive models with a high level of accuracy using time series data [27], [30].

In addition, the multiple layer perceptron (MLP) algorithm is commonly used in this area [27], [30], [33], [34]. MLP can explore non-linear associations between numerical or categorical predictors and the variable of interest, but it cannot "learn" directly the autocorrelation pattern of the dependent variable over a period of time. In this regard, in a recent study a transformation technique was used on the calendar data so that a MLP model can take patterns over time into consideration and therefore incorporate the calendar effect [34].

It is also quite common to use a Long Short-Term Memory (LSTM) algorithm [27], [35], [36]. As a type of Recurrent Neural Network (RNN), its characteristic of feedback connection allowed it to process data in sequence, a key feature of a time series data set.

III. Methodology

As mentioned, the impact of COVID-19 on energy consumption can be evaluated as the difference in utility usage under the COVID-19 environment (actual and observed) and the usage that we would see if the COVID-19 pandemic did not happen (not observable and to be estimated.) This difference can be estimated by comparing the actual utility usage under the pandemic and the usage estimates if the pandemic did not exist, the so-called baseline usage. The baseline usage is not observable but can be estimated using data collected before the pandemic. It is the main objective in this research.

A statistical time series method such as ARIMA is a reasonable approach in generating estimates for the baseline usage. However, other exogenous factors, such as temperature and relative humidity, should be considered as these factors also affect energy usage [37], [38]. As well, machine learning models are considered better candidates in modelling usage for their ability to allow for explicit specification of the usage patterns and other calendar effects such as holidays and weekends. Other features, such as maximum or minimum temperature for the day, could also be generated and specified from the exogenous factors so as to specify the possible "lingering" effect of the exogenous factors.

A. Machine Learning Models for Utility Usage Forecasts

This research considered the following machine learning algorithms for predicting hourly electricity usage: random forest regression (RFR), artificial neural nets (ANN), and support vector regression (SVR). As usual, we experimented with different combinations of features and hyper-parameters for each algorithm so as to maximize accuracy. Long-Short Term Memory (LSTM), along with the traditional multiple layer perceptron (MLP), were the two implemented as the competing ANN algorithms.

With the exception of the SVR algorithm, results from these algorithms are stochastic in nature due to the probabilistic routines (the re-sampling process in RFR and the gradient descent process in ANN) used within them. To ascertain the stability of the predictive power of the models developed using these algorithms, each model developed was run ten times. The mean of the performance metrics over the ten runs were then used for the evaluation of the models.

B. Description of the Data Set

The data set used in this research is collected by Harris SmartWorks. Harris SmartWorks manages the data generated hourly by millions of utility meters for approximately 30 utility providers in Canada as well as over 100 in North America and worldwide. In this applied research project, this usage and operational data with other environmental data was used to estimate the baseline usage.

The data set in this research is about electricity usage in a part of United States from May 28, 2018 to March 10, 2020. This data set originally has an hourly time stamp with usage and was augmented by the corresponding hourly data on temperature and relative humidity from the same geographic area.

Data collected from May 28, 2018 to June 30, 2019 was used as the training data set for the development of the ML models. The remaining data, from July 1, 2019 to March 10, 2020, was used as the testing data set to evaluate the estimation accuracy of the models.

C. Feature Engineering

As in a previous study, a transformation technique could be used to get the ML models to learn the calendar effect [34]. In this research, instead of using this transformation technique, a set of indicator variables for different calendar effects were created: the month of the year (12 one-up variables), day of the week (7 one-up variables), hour of the day (24 one-up variables), whether the day is a weekend (Saturday or Sunday; 1 one-up variable), and whether the day is a holiday (1 one-up variable).

To model the "lingering" impact of temperature on utility usage, we used the recorded temperature in the data set to create the average, lowest, and highest temperature of the day as well as the average, lowest, and highest temperature of the previous day. These variables are then used as part of the feature set in the development of the ML models.

D. Scaling

To ensure uniform unit among all numerical features, we use either the standardization (refer to Equation 1) or the Min-Max normalization (refer to Equation 2) technique.

$$X_{standardized} = \frac{X - E(X)}{SD(X)} \tag{1}$$

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

E. Performance Metrics

Two traditional performance metrics for estimation accuracy, Root Mean Square Error and Mean Average Percent Error, were used in this research. In addition, a third metric was developed by the project team during the research to better reflect the needs in measuring estimation accuracy for utility usage.

1) Root Mean Square Error and Mean Absolute Percent Error: The root means square error (RMSE) is commonly used in evaluating the predictive or estimation performance of models. Equation 3 shows the calculation of this metric.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)} \tag{3}$$

The mean absolute percent error (MAPE), which measures the average percentage of absolute error to the actual value, is also used in this project. Equation 4 shows the calculation of MAPE.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i|} \tag{4}$$

2) Total Absolute Error Percentage: Besides the two traditional performance metrics, a new metric, called the Total Absolute Error Percentage (TAEP) was developed by the project team to measure accuracy of the estimates comparing to the actuals, taking into account the average magnitude of the usage. This metric is therefore applicable as a performance metric for different time series data sets with different measurements and magnitudes. The TAEP metric is calculated as follows.

$$TAEP = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{\sum_{i=1}^{n} Y_i} \tag{5}$$

We used the TAEP metric as a primary metric for model comparison.

IV. Results

Together with the application of the usual feature engineering and hyper-parameter tuning techniques, various models were developed using the algorithms described above. Table I shows the best models developed for each type of algorithm and the hyper-parameters used. Models developed under LSTM were deemed to be inferior based on the performance metrics used and therefore omitted here.

In our experiments, we found that models with standardized numerical features performed better than those with normalized ones. Therefore, numerical features of the models shown here were all standardized.

The MLP model has the smallest average TAEP metric over ten runs. It is considered as the best of these three models even though it is less "consistent" with largest standard deviation. However, the upper range of the TAEP metric for this model is likely to be lower than that of the RFR or SVR which is constant at 6.05. The same can be said using the RMSE and MAPE metrics. Figure 1 graphically depicts the forecasting result (rolled up in days) of the MLP model on the testing data set. The estimated usage (blue line) follows the patterns of the actual usage well (orange dashed line). It does show, however, the challenges it has on estimating peak usage.

TABLE I
Performance of the Best Forecasting Model Under the Three Machine Learning Algorithms

Machine Learning Algorithms	**TAEP over 10 runs**	**MAPE over 10 runs**	**RMSE over 10 runs**
Random Forest Regression (Trees: 100)	Mean: 6.25 SD: 0.021	Mean: 6.95 SD: 0.03	Mean: 3909 SD: 18.19
Multiple Layer Perceptron (2 Hidden Layers: 100-32, relu)	Mean: 5.79 SD: 0.205	Mean: 6.54 SD: 0.17	Mean: 3564.19 SD: 92.13
Support Vector Regression (Kernel: rbf, Gamma: Scale, C: 1)	6.05	6.91	3762.68

* SVR is not stochastic in nature, there is no variation among ten runs

The TAEP metric and the MAPE are not scale sensitive, that is, these metrics are not affected by the unit or the scale of the actual and predicted values. Among all models shown in Table I, the mean and the standard deviation of TAEP metric is smaller than those of the MAPE and are deemed to be more appropriate for comparing utility usage estimates. For this reason, the TAEP metrics is implemented as the primary performance metric in Harris SmartWorks' utility usage prediction systems.

V. Conclusion and Future Work

In this research, we demonstrate the promise of using machine learning algorithms to generate accurate forecasts on a hourly utility usage time series incorporating relevant exogenous factors such as temperature and humidity. Comparing to traditional time series forecasting models, the ML models allow for the explicit modelling of hourly usage pattern as well as other calendar effects such as holidays and weekends. Engineered features, such as maximum or minimum temperature for the day, could easily be inserted to model the possible "lingering" effect of the exogenous factors. These advantages were explored in this research.

Fig. 1 shows the challenge in estimating peak usage which is critical in energy supply planning and acquisition. Performance metrics that accentuate estimation accuracy for peak usage would be desirable for the development of hourly usage prediction models.

Acknowledgment

We thank our research student assistants and students in capstone projects at Okanagan and Langara Colleges who assisted in the research. We also thank the reviewers of

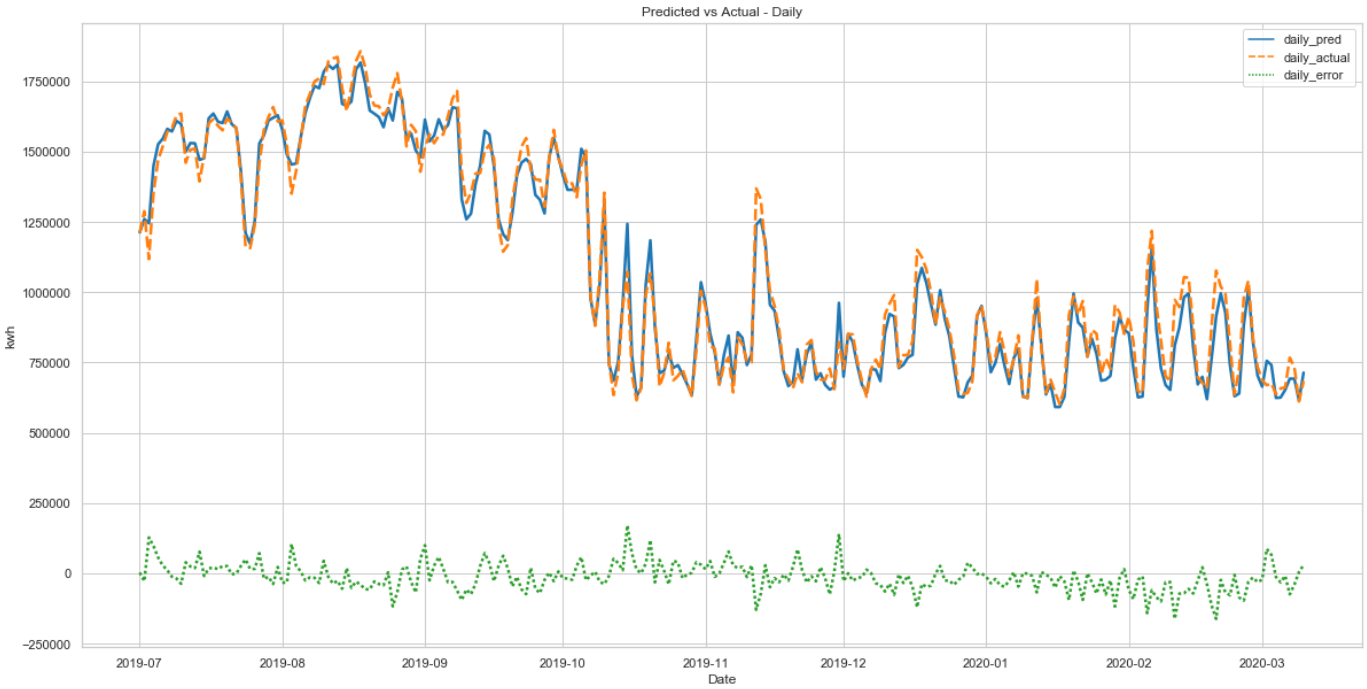

Fig. 1. Actual versus Forecasting Summarized by Day Under the MLP Model

this paper for their thoughtful comments and suggestions for improvement.

REFERENCES

[1] H. Cooley, "How the Coronavirus Pandemic is Affecting Water Demand," 2020. [Online]. Available: https://pacinst.org/how-the-coronavirus-pandemic-is-affecting-water-demand/

[2] B. Marohl and O. Comstock, "U.S. energy consumption in April 2020 fell to its lowest level in more than 30 years." [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=44556

[3] S. Hinson, "COVID-19 is changing residential electricity demand," Renewable Energy World, 4 2020.

[4] Y. Khmelevsky and V. Voytenko, "Cloud computing infrastructure prototype for university education and research," in *Computing*. ACM Press, 2010, pp. 1–5. [Online]. Available: https://doi.org/10.1145/1806512.1806524

[5] Y. Khmelevsky, V. Ustimenko, G. Hains, C. Kluka, E. Ozan, and D. Syrotovsky, "International collaboration in SW engineering research projects," in *Proceedings of the 16th Western Canadian Conference on Computing Education - WCCCE '11*, 2011.

[6] G. Hains, C. Li, Y. Khmelevsky, B. Potter, J. Gaston, A. Jankovic, S. Boateng, and W. Lee, "Generating a Real-Time Algorithmic Trading System Prototype from Customized UML Models (a case study)," no. 1, pp. 1–14, 2012.

[7] Y. Khmelevsky, M. Rinard, and S. Sidiroglou-Douskos, "A Source-to-source Transformation Tool for Error Fixing," in *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, ser. CASCON '13. Riverton, NJ, USA: IBM Corp., 2013, pp. 147–160. [Online]. Available: http://dl.acm.org/citation.cfm?id=2555523.2555540

[8] T. Alstad, J. R. Dunkin, R. Bartlett, A. Needham, G. Hains, and Y. Khmelevsky, "Minecraft computer game simulation and network performance analysis," *Second International Conferences on Computer Graphics, Visualization, Computer Vision, and Game Technology {(VisioGame 2014)}*, 11 2014.

[9] G. Hains, C. Li, N. Wilkinson, J. Redly, and Y. Khmelevsky, "Performance analysis of the parallel code execution for an algorithmic trading system, generated from UML models by end users," in *Parallel Computing Technologies (PARCOMPTECH), 2015 National Conference on*. IEEE, 2015, pp. 1–10. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84935436915&partnerID=tZOtx3y1

[10] T. Alstad, J. R. Dunkin, S. Detlor, B. French, H. Caswell, Z. Ouimet, Y. Khmelevsky, and G. Hains, "Game Network Traffic Emulation by a Custom Bot." *2015 IEEE International Systems Conference (SysCon 2015) Proceedings*, pp. 675–680, 4 2015.

[11] Y. Khmelevsky and V. Voytenko, "Hybrid Cloud Computing Infrastructure in Academia." in *WCCCE 2015 - the 20th Western Canadian Conference on Computing Education, At May 8-9, 2015.* Vancouver Island University (VIU), Nanaimo, British Columbia, Canada., 2015.

[12] G. Hains, Y. Khmelevsky, R. Bartlett, and A. Needham, "Game private networks performance: Analytical models for very-large scale simulation," in *2016 IEEE International Conference on Cybercrime and Computer Forensic, ICCCF 2016*, 2016.

[13] Z. Ouimet, H. Caswell, Y. Khmelevsky, R. Bartlett, and A. Needham, "Game servers deployment automation case study," in *2016 Annual IEEE Systems Conference (SysCon)*, 2016, pp. 1–7.

[14] D. Atkinson, N. McDonald, and Y. Khmelevsky, "Reporting personal and corporate data for secure storage in cloud," in *2016 IEEE International Conference on Cybercrime and Computer Forensic, ICCCF 2016*, 2016.

[15] N. McDonald, D. Atkinson, Y. Khmelevsky, and S. McMillan, "Sport wearable biometric data encrypted emulation and storage in cloud," in *Canadian Conference on Electrical and Computer Engineering*, 2016.

[16] Y. Khmelevsky, "Ten Years of Capstone Projects at Okanagan College: A Retrospective Analysis," in *Proceedings of the 21st Western Canadian Conference on Computing Education*. New York, NY, USA: ACM, 2016, pp. 7:1–7:6. [Online]. Available: http://doi.acm.org/10.1145/2910925.2910949

[17] G. Hains, Y. Khmelevsky, R. Bartlett, and A. Needham, "Game private networks performance: From geolocation to latency to user experience," in *11th Annual IEEE International Systems Conference, SysCon 2017 - Proceedings*, 2017.

[18] B. Ward, Y. Khmelevsky, G. Hains, R. Bartlett, A. Needham, and T. Sutherland, "Gaming network delays investigation and collection of very large-scale data sets," in *11th Annual IEEE International Systems Conference, SysCon 2017 - Proceedings*, 2017.

[19] Y. Khmelevsky, K. Chidlow, K. Sugihara, and K. Zhang, "Engaging and Motivating Students Through Programming Competitions and GIS Applied Research Projects," *Proceedings of the 22nd Western Canadian Conference on Computing Education*, 5 2017. [Online]. Available: http://dx.doi.org/10.1145/3085585.3088491

[20] M. Cocar, R. Harris, and Y. Khmelevsky, "Utilizing Minecraft bots to optimize game server performance and deployment," in *Canadian Conference on Electrical and Computer Engineering*, 2017.
[21] G. Hains, C. Mazur, J. Ayers, J. Humphrey, Y. Khmelevsky, and T. Sutherland, "The WTFast's Gamers Private Network (GPN®) Performance Evaluation Results," in *2020 IEEE International Systems Conference (SysCon)*. IEEE, 2020, pp. 1–6.
[22] C. Mazur, J. Ayers, J. Humphrey, G. Hains, and Y. Khmelevsky, "Machine Learning Prediction of Gamer's Private Networks (GPN®S)," in *Proceedings of the Future Technologies Conference*. Springer, 2020, pp. 107–123.
[23] A. Wong, C. Chiu, G. Hains, J. Behnke, Y. Khmelevsky, and C. Mazur, "Network Latency Classification for Computer Games," in *The IEEE International Conference on Recent Advances in Systems Science and Engineering (submitted)*, 2021.
[24] A. Wong, C. Chiu, G. Hains, J. Humphrey, Y. Khmelevsky, C. Mazur, and H. Fuhrmann, "Gamers Private Network Performance Forecasting - From Raw Data to the Data Warehouse with Machine Learning and Neural Nets," 2021. [Online]. Available: https://arxiv.org/abs/2107.00998
[25] H. A. N. Hejase and A. H. Assi, "Time-Series Regression Model for Prediction of Mean Daily Global Solar Radiation in Al-Ain, UAE," *ISRN Renewable Energy*, vol. 2012, pp. 1–11, 4 2012.
[26] S. Katara, A. Faisal, and G. M. Engmann, "A Time Series Analysis of Electricity Demand in Tamale , Ghana," *International Journal of Statistics and Applications*, vol. 4, no. 6, pp. 269–275, 2014.
[27] S. Kaushik, A. Choudhury, P. K. Sheron, N. Dasgupta, S. Natarajan, L. A. Pickett, and V. Dutt, "AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures," *Frontiers in Big Data*, vol. 3, p. 4, 3 2020. [Online]. Available: www.frontiersin.org
[28] C. Peña-Guzmán and J. Rey, "Forecasting residential electric power consumption for Bogotá Colombia using regression models," in *Energy Reports*, vol. 6. Elsevier Ltd, 2 2020, pp. 561–566.
[29] I. Shah, H. Iftikhar, S. Ali, and D. Wang, "Short-term electricity demand forecasting using components estimation technique," *Energies*, vol. 12, no. 13, p. 2532, 7 2019. [Online]. Available: www.mdpi.com/journal/energies
[30] M. Kalimoldayev, A. Drozdenko, I. Koplyk, T. Marinich, A. Abdildayeva, and T. Zhukabayeva, "Analysis of modern approaches for the prediction of electric energy consumption," pp. 350–361, 1 2020.
[31] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Wiley Online Library*, vol. 3, no. 1, pp. 62–76, 1 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/2475-8876.12135
[32] X. Zhang, K. Grolinger, and M. A. M. Capretz, "Forecasting Residential Energy Consumption Using Support Vector Regressions," in *Proceedings of the IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA*. New York, NY, USA: IEEE Press, 2018, pp. 17–18.
[33] J. S. McMenamin and F. A. Monforte, "Short term energy forecasting with neural networks," *The energy journal*, vol. 19, no. 4, p. 5, 1998.
[34] J. Moon, S. Park, S. Rho, and E. Hwang, "A comparative analysis of artificial neural network architectures for building energy consumption forecasting," *International Journal of Distributed Sensor Networks*, vol. 15, no. 9, p. 1550147719877616, 9 2019. [Online]. Available: https://us.sagepub.com/en-us/nam/open-access-at-sage
[35] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, 2019.
[36] K. Yan, W. Li, Z. Ji, M. Qi, and Y. Du, "A hybrid LSTM neural network for energy consumption forecasting of individual households," *Ieee Access*, vol. 7, pp. 157 633–157 642, 2019.
[37] M. Ali, M. J. Iqbal, and M. Sharif, "Relationship between extreme temperature and electricity demand in Pakistan," *International Journal of Energy and Environmental Engineering*, vol. 4, no. 1, pp. 1–7, 9 2013. [Online]. Available: http://www.journal-ijeee.com/content/4/1/36
[38] I. Staffell and S. Pfenninger, "The increasing impact of weather on electricity supply and demand," *Energy*, vol. 145, pp. 65–78, 2 2018.