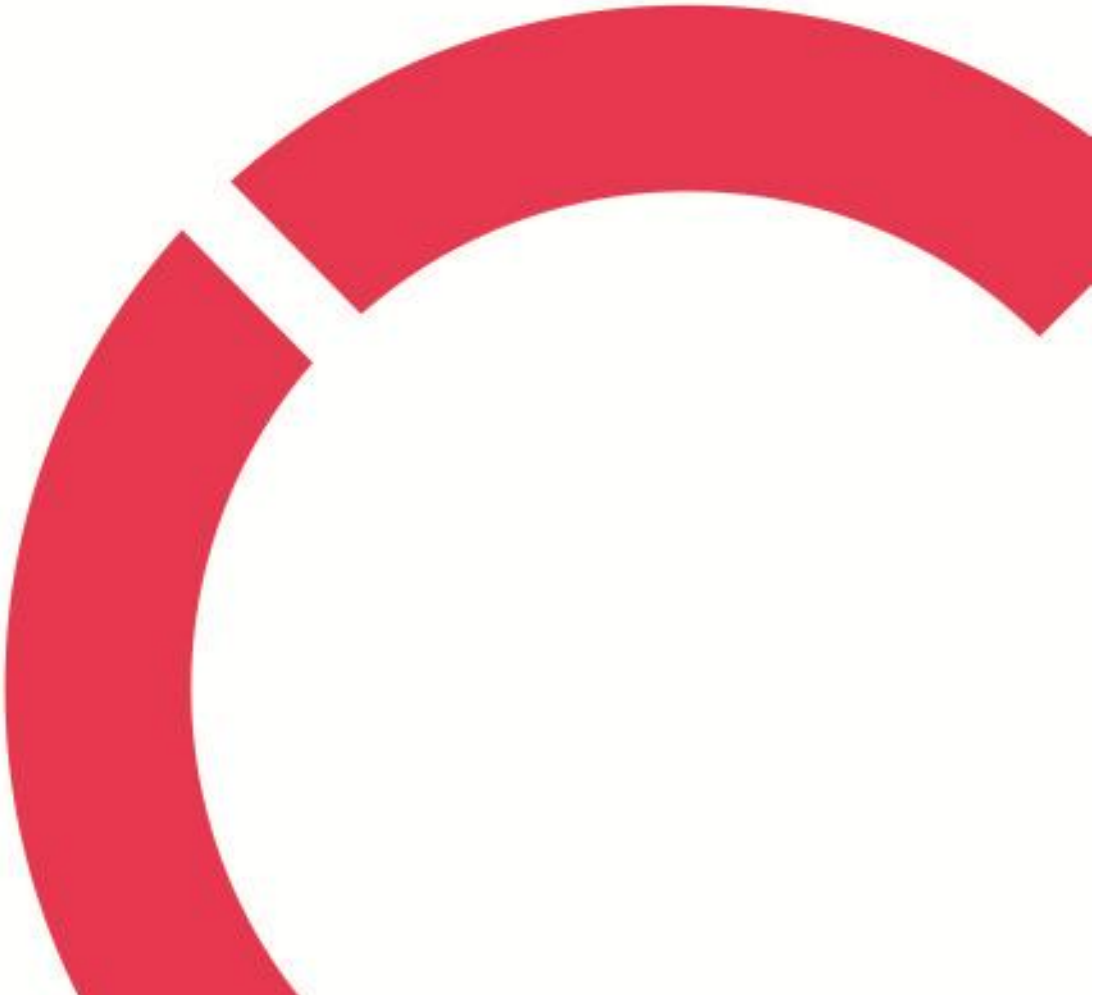
Saurav Amatya

ENHANCED ENERGY FORECASTING FOR VIRTUAL POWER PLANTS

Leveraging Machine Learning for Improved Efficiency

CENTRIA UNIVERSITY OF APPLIED SCIENCES
Bachelor of Engineering, Information Technology
May 2025



ABSTRACT

Centria University of Applied Sciences	**Date** May 2025	**Author** Saurav Amatya
Degree programme Bachelor of Engineering, Information Technology		
Name of thesis ENHANCED ENERGY FORECASTING FOR VIRTUAL POWER PLANTS. Leveraging Machine Learning for Improved Efficiency		
Centria supervisor Panu Wirkkala		**Pages** 21+3
Instructor representing commissioning institution or company Fabian Sander		

The research examines the changing convergence of energy resources and machine learning with particular focus on enhancing energy forecasting for Virtual Power Plants (VPPs). This literature review discusses through the existing solutions regarding energy forecasting and paves the way for extensive examination of data analytics challenges for a biogas-powered electricity system. The study investigates electricity consumption and production patterns with regards to the changes in weather conditions and spot electricity prices. The study highlights the challenges related to data collection and preprocessing which in turn is the foundation of predictive forecasting modelling. This paper outlines practical solutions that highlights the importance of various statistical and machine learning models that improve forecasting accuracy. The findings provide the potential of data-driven solutions in minimizing reliance on grid electricity, maximizing biogas electricity usage and reducing overall energy costs. The analysis concludes by advocating for intelligent, automated solutions that balance energy consumption and cost-effectiveness in VPPs.

Key words
Biogas electricity, data-driven solutions, energy forecasting, machine learning, virtual power plants

CONCEPT DEFINITIONS

DER

(Distributed Energy Resources) is small-scale energy resources usually situated near sites of electricity use.

EDA

(Exploratory Data Analysis) is a method of analysing the data to comprehend its main characteristics.

IOT

(Internet of Things) is a network of interconnected devices that communicate with each other over the internet.

API

(Application Programming Interface) is set of protocols and tools that allow different software application to communicate with each other.

LSTM

(Long Short-Term Memory) is a type of neural network that can remember past data, used in time-series analysis.

XGBoost

(Extreme Gradient Boosting) is a machine learning method that makes predictions by combining multiple models.

ABSTRACT
CONCEPT DEFINITIONS
CONTENTS

FIGURES

TABLES

1 INTRODUCTION

Virtual Power Plants (VPPs) refer to an integrated energy management system that combines and optimizes distributed energy sources (DERs) such as biogas, solar or wind power. VPPs play a key role in innovation for the future of electricity generation and management due to their capability of balancing the operation of multiple energy sources. Further, they ascertain to balance the demand and supply of electricity at optimal level in real time.

In Finland, a country that heavily relies on electricity, it is essential to make the best use of available energy resources. Biogas, produced by process of anaerobic digestion of animal manure, is used to generate clean electricity that provides a sustainable and environmentally friendly energy source. However, the electricity production from biogas is prone to fluctuations due to changes in weather conditions and manure availability. Its consumption is also not stable as it depends on weather conditions. Thus, there is a discrepancy which leads to inefficient use of energy ultimately leading to economic drawbacks. To solve this problem, a data-driven solution is required that utilizes the data collected in the biogas plant as well as the weather data.

The objective of this research is to improve energy efficiency for biogas-based Virtual Power Plant (VPP). The research investigates the key requirements in data collection, data preprocessing and data cleaning techniques as it provides a foundation for the predictive modelling development. Further, it provides the analysis on how to take data-driven approach and apply machine learning techniques to enhance the electricity forecasting.

2 VIRTUAL POWER PLANT

In the context of growing integration of renewable energy, virtual power plants (VPPs) are essential for improving energy security and resilience. They play an important role particularly in increasing renewable energy integration and managing distributed energy resources (DERs). This speciality facilitates the possibility of adjusting supply and demand dynamically despite fluctuations. The various risks and threats related to the inconsistent functioning of renewable energy sources pose a challenge in integration to the grid. As the world moves further towards renewable energy sources, new possibilities and energy security challenges have arrived. Nevertheless, the challenges related to operation of diverse DERs can be reduced by the comprehension of real time data and advanced analytics. The ability to adapt to energy fluctuations helps mitigate challenges posed by the variability of renewable energy but also decreases the dependence on conventional fossil-based energy. (Kaiss, Wan, Gebbran, Unsihuay Vila & Dragičević 2025,3.)

Statistical models like ARIMA (Autoregressive Integrated Moving Average) have been used in traditional forecasting systems. However, as shown by Cadenas & Rivera (2010), ARIMA is not accurate dealing with nonlinear patterns that comes with renewable energy systems innately. ARIMA achieves MAPE values larger than 15% in wind speed prediction tasks. These drawbacks are particularly evident in biogas applications, where feedstock variability introduces non-stationary dynamics (Kozhakhmet, S., Lyazat, S., Siladi, G., Gulbakyt, K., & Maksatbek, K. , 2020).

With the advances in machine learning (ML) and deep learning (DL), alternative approaches have been created that are able to capture these complexities. Gulzat, Lyazat, Siladi, Gulbakyt, and Maksatbek (2020) emphasize the better performance of machine learning models such as Random Forest and XGBoost compared to traditional linear models for energy forecasting tasks. The increase in performance is because of their ability to model variable interactions and non-linear relationships. Gulzat et al. (2020) stated that machine learning has transformed energy forecasting by addressing nonlinear relationships and temporal complexities. Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber (1997), possess gate mechanisms (input, forget, output) to capture long-term dependencies in time-series data. LSTM models can reduce RMSE by 23% compared to ARIMA in renewable energy prediction. (Zhang et al., 2020). Meanwhile, XGBoost's gradient-boosted trees dominate structured forecasting tasks, winning 17/30 competitions in the 2023 IEEE Energy Challenge

(Chen & Guestrin, 2016). The model uses histogram-based optimization to minimize the training time and applies feature importance scoring to achieve better results. (Chen et al., 2019).

3 METHODOLOGY

This chapter outlines the methodology used to develop and evaluate machine learning models for forecasting electricity consumption. It details the data collection process, preprocessing steps, model selection and evaluation techniques. The subsequent sections provide a through explanation of each phase, from data preparation to model optimization and performance assessment. The methodology follows a systematic approach that begins with exploratory data analysis to understand consumption patterns and identify key temporal features. The data preprocessing includes normalization, outlier analysis, creating lag features to capture time-dependent relationships in the data. The chapter also presents various machine learning algorithms to conduct forecasting that includes traditional regression model, ensemble methods such as XGBoost and Random Forest, and deep learning approach with LSTM networks. The subsequent sections provide a thorough explanation of each phase from data preparation to model optimization and performance evaluation.

3.1 Data Collection Process

The data collect process involved multiple sources to gather comprehensive information on biogas plant operations, electricity markets and weather conditions. The primary data sources included, namely, biogas plant, electricity market data and meteorological data. Biogas electricity production and on-site farm electricity consumption data were collected at 15-minute intervals with Fingrid's API connected to the internet. The spot electricity prices and day-ahead forecasts for Finland were obtained from ENTSO-E (European Network of Transmission System Operators for Electricity) open data which was fetched from API. The metrological data such as temperature, wind speed and cloudiness were sourced from the Finnish Metrological Institute (FMI) open data for every hour.

The collected data had time series nature which consisted of the electricity consumption data for every 15 minutes with date and time information. The data was stored in cloud database for proper handling of data and to maintain data integrity. This multi-faceted data collection approach aligns with best practices in biogas research, as it captures the interplay between plant operations, market conditions, and environmental factors. The high-frequency data collection, particularly for electricity production and consumption, allows for detailed analysis of plant performance and energy demand patterns. The dataset collected had twenty months of consumption data. The dataset utilized for this research comprises three primary fields: Unit, representing the measurement of electricity consumption in kilowatt-

hours (kWh); Date, recorded in Coordinated Universal Time (UTC, Zulu format); and Value, expressed as a floating-point number.

3.2 Data Preprocessing

Data preprocessing is a critical step in preparing time series data for analysis and modelling in biogas production forecasting. This process involves several key steps to ensure data quality, consistency, and suitability for predictive modelling. The first step of data processing is the data cleaning. The data collected needs to be combined and it is important to understand the data types, formats and key fields in the data sources before the data cleaning process. Further researchers create memos that document emerging interpretations of the data with insights and analytical significance. The next step is coding where descriptive labels are provided to data segments based on business interests (Lester, Cho, & Lochmiller 2020).

The raw dataset was in CSV format, where the field delimiter was a semicolon (;) instead of a comma (,). The data was parsed in an appropriate format to correctly load the data. Additionally, the “Value” field used a comma as the decimal separator, which was standardized to a period (.) to ensure proper numerical interpretation. The Date field was originally in Coordinated Universal Time (UTC); it was subsequently converted to Eastern European Time (EET) to align with the local context of energy consumption patterns. During the preprocessing phase, the dataset was checked for missing or null values. No missing entries were found in any of the fields; therefore, no imputation methods such as interpolation or forward-filling were necessary. The dataset was visually inspected and statistically analyzed using boxplots, Interquartile Range and z-scores method to detect potential outliers in electricity consumption values. As shown in Figure 1, the red dots indicate detected outliers while the blue dots represent consumption values. No significant anomalies were identified that could adversely affect model training hence, no outlier removal was performed.

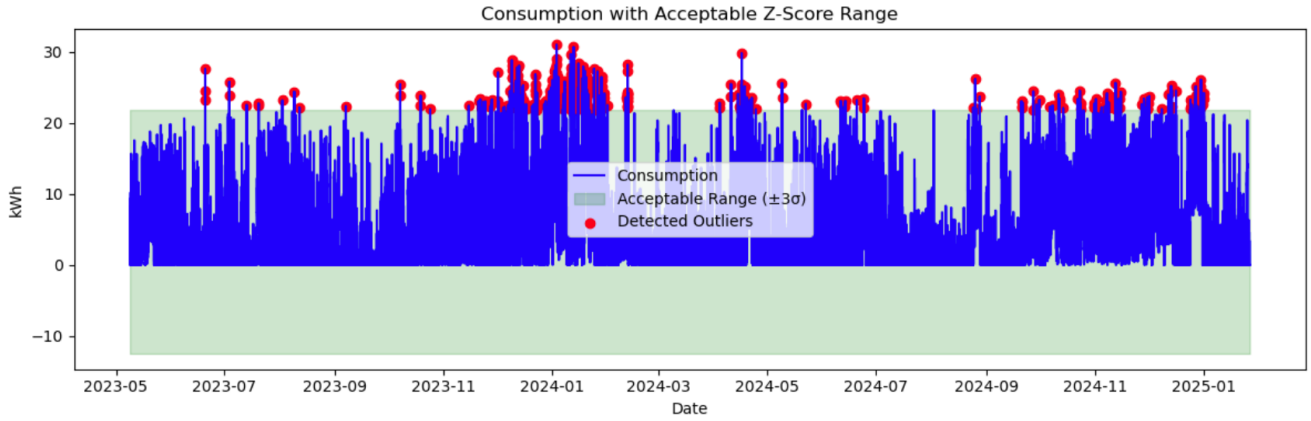


FIGURE 1. Electricity consumption patterns and anomaly detection

3.3 Predictive Modelling

Predictive modelling refers to creation and execution of the use of learning models to recognise patterns in data and making predictions out of them. Machine learning plays an important role in pattern recognition in data to gain the knowledge of information and make predictions with mathematical algorithms. Common machine learning techniques include decision trees like random forests and neural networks such as LSTM that is widely used to solve classification and prediction problems. The selection of the technique is based on the problem statement itself and how this research wants the solutions. The researcher needs to make decisions on selection based on accuracy and efficiency and cost effectiveness.

(Gulzat, Lyazat, Siladi, Gulbakyt, Maksatbek 2020, 300)

3.4 Model Evaluation and Optimization

In the application of energy forecasting, effective model evaluation and optimization is crucial for improving time series forecasting accuracy in the application of energy forecasting. Spatially pooled verification is one of the evaluation approaches where models are unfairly penalised for small spatial misalignments. This technique enables a definitive evaluation of forecasting accuracy as it involves evaluation of multiple forecasts over several spatial regions. A machine learning-based forecasting model, GenCast has demonstrated better performance than traditional ensemble methods. In order to optimize time-series models, selection of the appropriate algorithms, fine tuning the hyperparameters and incorporation of probabilistic metrics are required to ensure efficiency and reliability. (Price, I., Sanchez-Gonzalez, A., Alet, F. et al, 2025.*)*

After the predictive models were developed, they were evaluated based on performance metrics such as MSE, RMSE, MAE, R^2. These metrics are crucial for understanding how well the model generalized from the training data and its prediction compared to test data. The performance is considered better when the values of MAE, MSE, RMSE are lower. The R^2 error is a measure between 0 and 1 and the performance is considered better if the values are higher. The comparison of these metrics for different models allows identification of the best model with the most accurate predictions for the electricity consumption data.

The models made use of base hyperparameters, and its results were observed. Further, hyperparameter tuning was implemented to find out the hyperparameters that gives lower error and better predictions. GridsearchCv was implemented to optimize the hyperparameters for the XGBoost and random forest models. However, for the LSTM model, keras tuner was applied. Additionally, feature engineering, including the creation of lag features, rolling statistics, and time-based attributes, was added to identify the feature importance and select the most appropriate features to improve model performance. The models were then compared based on their evaluation metrics to identify the most effective model for electricity consumption forecasting.

4 IMPLEMENTATION AND RESULTS

This chapter presents the results of the implementation of the methodology described in the previous chapter. The implementation includes data analysis, model training, and evaluation, focusing on the performance of the selected machine learning models—Random Forest, XGBoost, and LSTM—on the electricity consumption forecasting task. Additionally, the chapter highlights the insights derived from the analysis and the effectiveness of the models in meeting the forecasting objectives.

4.1 Data Analysis Insights

During the early exploratory data analysis (EDA) phase, external variables, weather data, such as temperature, wind speed, and cloud coverage were considered as potential features. As illustrated in Figure 2, the electricity consumption varies across different months of the year that demonstrates seasonal patterns. From the analysis of the electricity consumption data, it was found that the winter months (October to January) exhibit the highest levels of energy usage. This period likely corresponds to increased heating demands associated with lower temperatures. Conversely, the spring and fall months (February to September) generally show a reduction in energy consumption, with March recording the lowest point of the year. This decline can be attributed to the transitional nature of these months, where milder temperatures result in decreased heating and cooling needs.

The data analysis of electricity consumption revealed distinct temporal patterns that can inform the forecasting model. Hourly analysis identified peak consumption times in the afternoon, particularly between 1 PM and 3 PM, with lower consumption observed during early morning and late-night hours. This suggests that energy demand is primarily driven by daytime activities. Weekly analysis indicated consistent consumption across the days, with slight peaks on Thursday and minimal variations between weekdays and weekends. This stability suggests that energy usage remains largely unaffected by the day of the week, likely due to continuous household or business routines. The insights gained from the analysis of temporal data can be further used as knowledge for machine learning models.

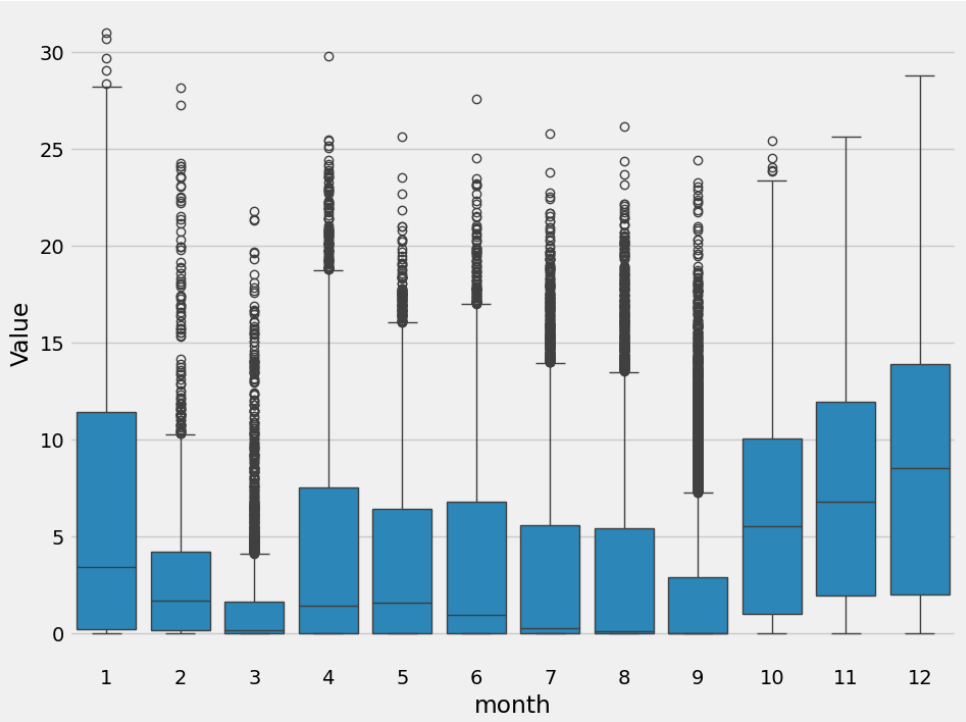


FIGURE 2. Energy consumption over month

However, correlation analysis showed that these weather-related variables had minimal influence on the electricity consumption patterns in the dataset. As demonstrated in Figure 3, the correlation heatmap revealed weak correlation between electricity consumption and weather parameters such as wind speed and average temperature. The highest correlation value observed was -0.25 with temperature. Since the primary focus of this research was to evaluate and compare the forecasting capabilities of machine learning models rather than conduct a comprehensive feature analysis, weather data was excluded to maintain focus and avoid diluting the scope of the study. Future research could explore multi-variable models if weather-dependency is more evident in other datasets or regions.

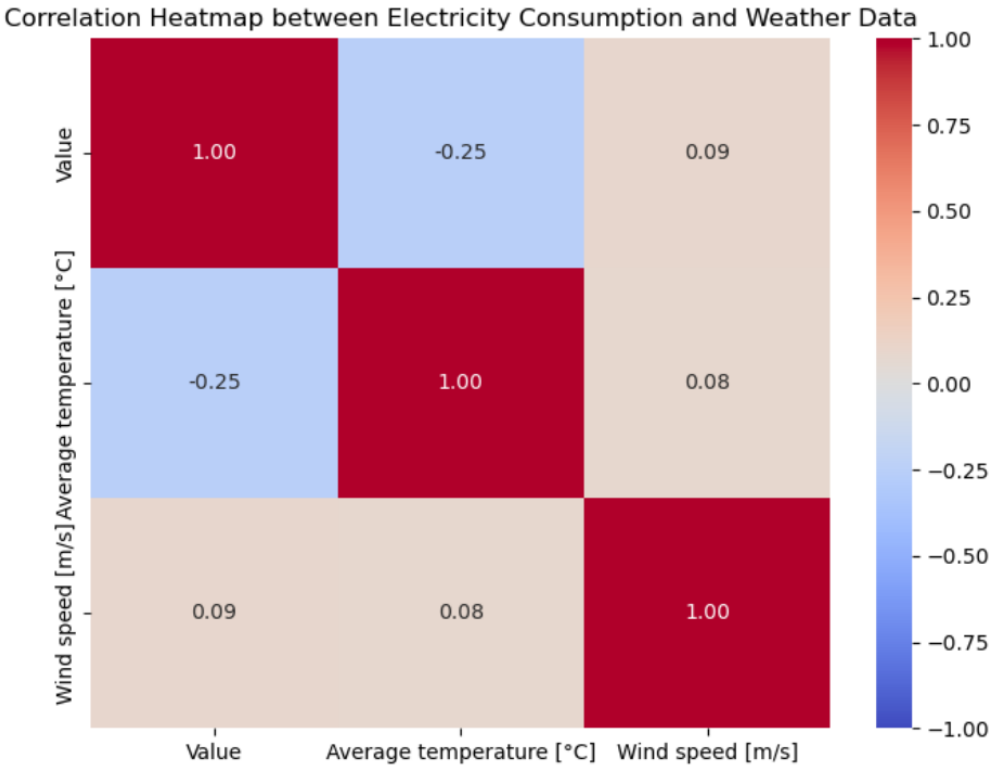


FIGURE 3. Correlation heatmap between electricity consumption and weather data

4.2 Model Implementation

This section presents the machine learning models used for the prediction of electricity consumption. The models used are base model which is naïve regression model with only one feature. Subsequently, more advanced ensemble models such as XGBoost, random forest and LSTM were implemented. The MinMaxScaler function of scikit-learn helped normalizing the target value before feeding into the models. Further, models were created using Python libraries XGBoost and Keras-Tensorflow. The data was divided into training and test with 70-80% as training and the rest as test data. The test data was later compared to the model's predicted value. This provided the evaluation for the performance of machine learning.

4.2.1 Base Model

A base model was created in the beginning model with weather data as independent variable and the consumption value as dependent variable. During correlation analysis, it was observed that the weather data such as temperature, cloud cover, and windspeed had quite a low correlation to the electricity consumption. Amongst them the temperature had the highest correlation of -0.25. Although the correlation is not very significant, a naive model was created based on temperature feature of data. By utilizing

temperature as the sole predictor variable, this model provided a benchmark against which the performance of more advanced models, such as Random Forest, XGBoost, and LSTM, could be assessed, ensuring that any improvements in predictive accuracy were due to the model's enhanced complexity rather than the inclusion of additional features.

After the baseline model, more sophisticated models were constructed which enabled a more comprehensive evaluation of each model's ability to capture the intricate temporal and contextual patterns inherent in electricity consumption data. The comparison of the performance of other models with the baseline allowed for the assessment of their relative efficacy and identification of potential areas for further refinement.

4.2.2 XGBoost

In this model, the dataset was split into training and testing sets using a chronological split, with 70% of the data allocated for training and 30% for testing. Specifically, the training set comprised data up to October 1, 2024, while the test set contained data from this date onwards. The split is visually represented in Figure 4 where blue colour represents training set while red colour represents training.

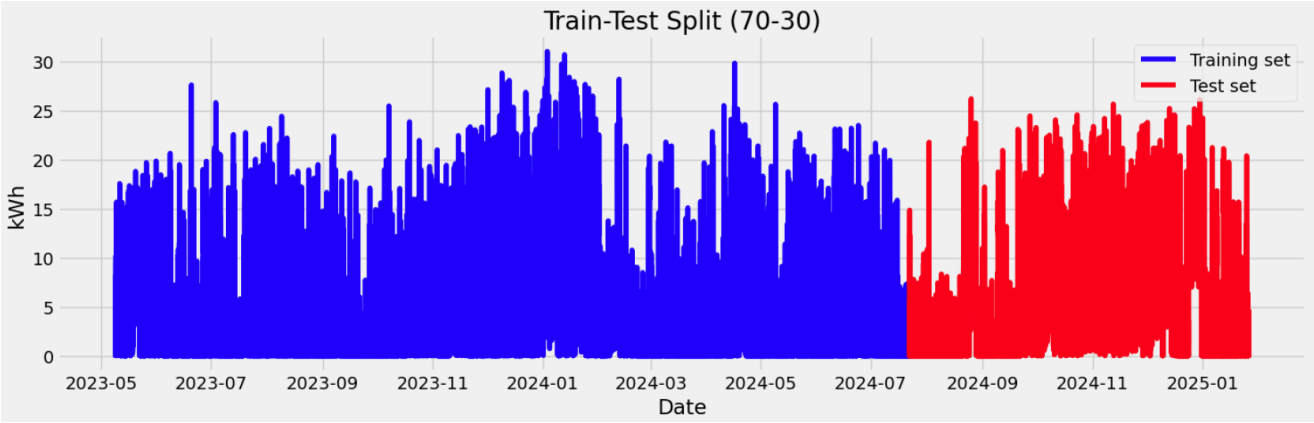


FIGURE 4. Temporal split of data for model training and testing

The data was enriched with several engineered features to capture temporal patterns and trends. Features included time-based variables such as the hour of the day, day of the week, month, and year, as well as cyclical features using sine and cosine transformations. Additional features included lag values (e.g., 1-hour and 6-hour lags) and rolling statistics such as moving averages, standard deviations, and medians over different time windows (e.g., 3, 6, 12, 24, 48 hours). The function implemented these transformations to both the training and test datasets.

The feature set for model training included a combination of these temporal and rolling statistics. The target variable for the regression task was the Value column, representing the electricity consumption. Two iterations of model training were conducted using XGBoost, a gradient-boosting framework that excels at handling structured data. In the first trial, a conservative approach was adopted with a small learning rate and early stopping to avoid overfitting. The hyperparameters were tuned again yielding better model performance in the second attempt. The hyperparameters that were retuned optimally were maximum depth, subsample ration and regularization parameters.

The features were created based on the temporal information of time-series data. The features were created from the date time information itself such as day of the week, month, quarter, year, hour of the day and weekend indicators. The purpose of these features was to understand the data better and to know if there is some pattern based on these features. The other features included lag features, sine features, cosine features and rolling features. Figure 5 shows the importance value of features, and it was observed that the feature with the highest importance is “lag_1” which represents the lag value at 1 hour. This indicated that the forecast significantly depends on the previous time step. The other features that were observed to have high importance were rolling features particularly for 3 hours.

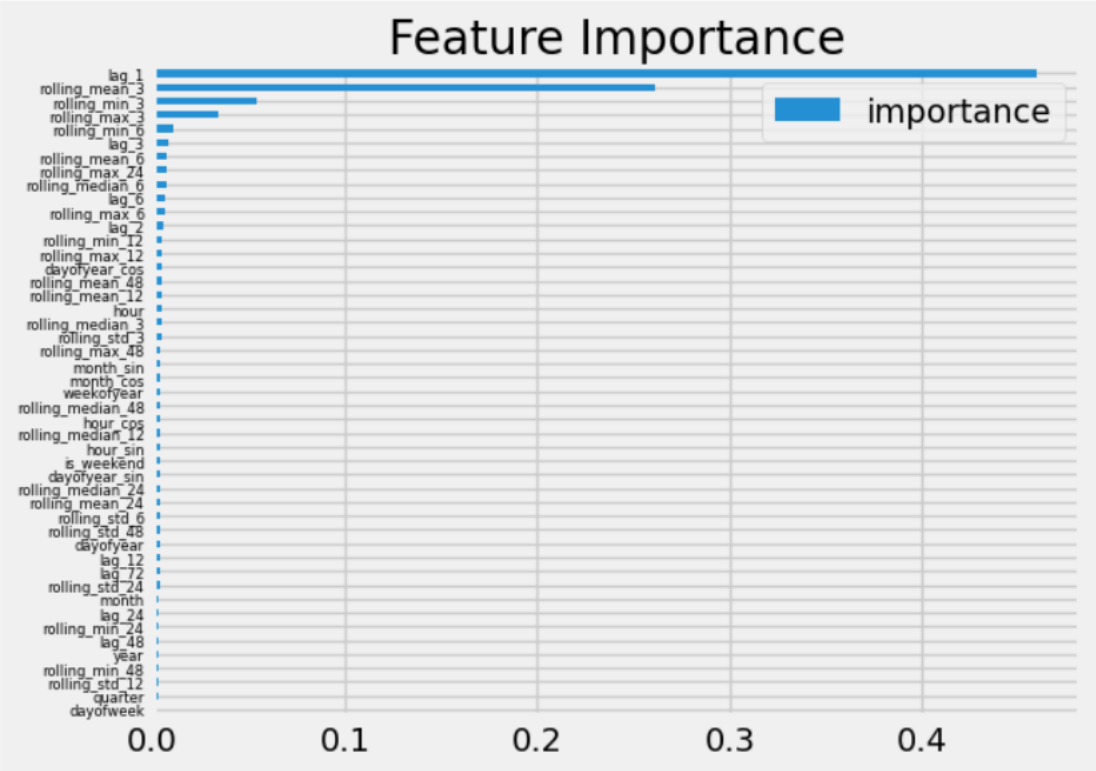


FIGURE 5. Features and their importance

The models were developed and evaluated in two phases. During the first phase, the prediction stayed regression to the mean and the higher and lower values were inaccurate. The first phase had parameters like number of estimators, early stopping rounds and learning rate only. In the second phase, more parameters were added to make the model more accurate. During both phases, it was observed that predicted data had one time step ahead with respect to the test data. This phenomenon occurred due to the lag features, and it was corrected by shifting the predicted back to match the test data. The prediction model was able to trace the trends and patterns well. The predictions consistently exhibited a delay of approximately 15 minutes, which was later identified because of the lag features incorporated in the model. Therefore, the predicted values were shifted so that the predicted data and test data aligned, and the pattern was preserved. The short-term and mid-term patterns were well recognized by the model despite challenges in timing.

4.2.3 Long Short-Term Memory (LSTM)

In the LSTM model implementation, the dataset was first normalized using MinMaxScaler to scale the consumption values between 0 and 1. A sequence length of 96 was chosen, corresponding to one day's worth of data at 15-minute intervals. The data was split into training (80%) and testing (20%) sets to evaluate the model’s generalization performance. The training process involved preparing the data into sequences of 96-time steps as input (X) and the corresponding next time step as the target (y). The LSTM model was built using Keras' Sequential API, with one LSTM layer consisting of 64 units, followed by a Dense layer to output the predicted consumption value. The model was trained for 10 epochs using the Adam optimizer with a learning rate set to minimize the mean squared error (MSE) loss.

The model's performance was evaluated both visually and quantitatively. During training, the model demonstrated a tendency to capture the underlying trends in the data, but like the XGBoost model, it exhibited a regression-to-mean behaviour, especially during low-consumption periods. The predictions from the LSTM model showed a consistent lag, approximately one time step (15 minutes) behind the actual consumption data. This shift in predictions could be attributed to the model's ability to capture long-term trends but its challenge in precisely predicting the exact timing of consumption peaks. The model's ability to learn temporal dependencies over short periods (e.g., within a day) was evident, but its shortcoming was in predicting sudden spikes or drops in consumption.

The data was first scaled using MinMaxScaler of sklearn that transformed the data into the range between 0 to 1. It was done so that there is no bias toward features with higher values. The sequence

length was set to 96 that represents entire day of every 15 minutes data. The keras module was used to build the model with 64 neurons and dense value of 1. This was done since the most significant feature importance was one timestep lag. Mean Squared Error (MSE) was used as the loss function and Adam as the optimizer for the training. Initially, the number of epochs was set to 10 with 32 samples trained per batch. The data was split to 80 % to training and 20% for testing. The cost function had to be scaled back to the original scaling to assess its results. The comparison between the actual consumption values and the predicted values over sequential time steps is depicted in Figure 6. The visualization demonstrated how closely the model's prediction followed the real consumption patterns.

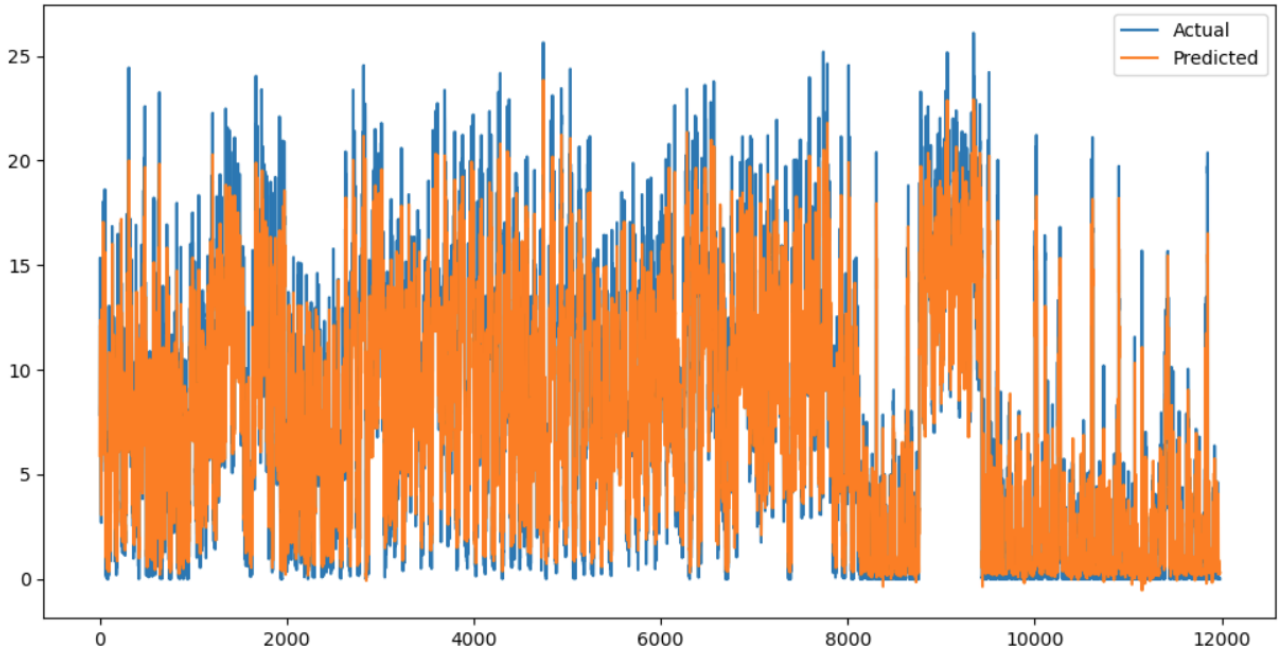


FIGURE 6. Comparison between actual and predicted data in LSTM

4.2.4 Random Forest

Random forest is an ensemble learning method commonly used for regression problems. The model creates decision trees based on which predictions are made. The preprocessing step included creation of lag features and splitting data to features and target. The target being the consumption value and features included the lag features. RandomForestAgressor was used to train the model. The data was split into 80% for training and 20 % for testing. The random forest was chosen to handle the nonlinear nature of data and to avoid overfitting. The hyperparameters included number of estimators and maximum depth that developed the basis of the predicting model. Furthermore, the model was tuned using GridSearchCV from sklearn library which automatically tested different combination of the provided

parameters. After the hyperparameters tuning the results were better assessed using the performance metrics. The comparison of the predicted values and the actual value is shown in Figure 7 where the predicted values in red colour has mostly overlapped the actual values.

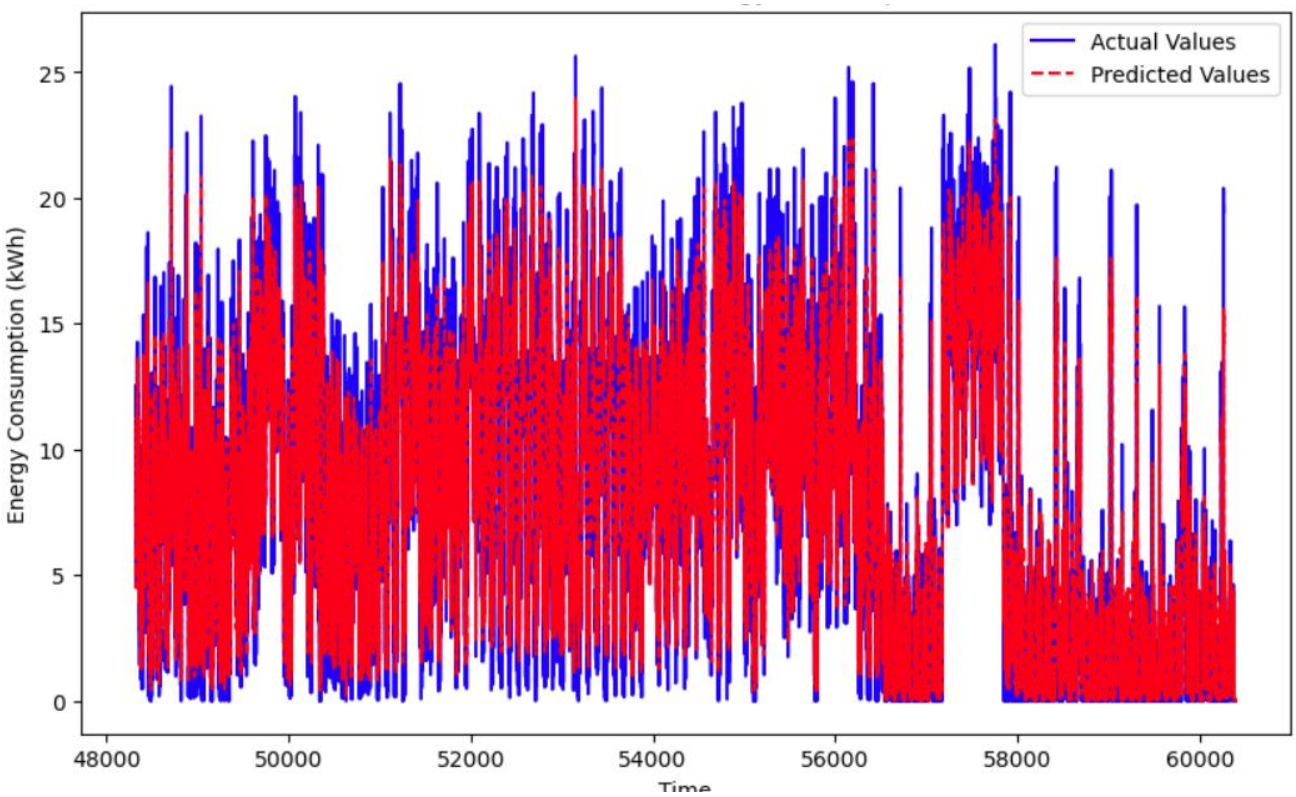


FIGURE 7. Prediction of Random Forest

The model's performance was evaluated using key metrics such as Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the coefficient of determination (R^2). The results of key metrics revealed outstanding results that has minimal prediction error and a higher R^2 value compared to other models experimented.

4.3 Performance Evaluation

The performance of the prediction model was evaluated using standard performance metrics that includes MSE, RMSE, MAE and coefficient of determination. (R^2) The models needed to be evaluated and compared so that this research can determine which model worked the best at predictions of electricity consumption. RMSE is an important metric as it is presented in the actual unit of data. It tells how far off the prediction is compared to the test data. The value of R^2 lies between 0 and 1 and

higher values indicate better performance. The comparison of the LSTM, XGBoost, and Random Forest models are presented in Table 1.

TABLE1.Performance comparison of forecasting models based on evaluation metrics

Model	**MSE**	**RMSE**	**MAE**	**R^2**
Base	31.87	5.64	4.60	0.07
XGBoost	6.17	2.48	1.76	0.83
LSTM	6.55	2.56	1.795	0.818
Random Forest	6.12	2.47	1.74	0.83

The performance evaluation shows that the Random Forest model significantly outperforms both the LSTM and XGBoost models. With the lowest RMSE (2.47) and MAE (1.74), as well as an impressive R^2 value of 0.83, it explains a substantial portion of the variance in the target variable, indicating a very good fit. In comparison, the LSTM model, with an RMSE of 2.56 and an MAE of 1.795, performed decently but lacked the predictive accuracy and fit of the Random Forest model. The XGBoost model, with the highest RMSE of 2.48, showed larger errors and, while still relatively accurate, fell behind the other two models in terms of performance. Overall, the Random Forest model is the most accurate and reliable for this regression task.

5 DISCUSSION

This chapter discusses about the results from the implementation of the prediction models developed for the electricity consumption forecasting. It interprets the significance of the findings considering the research objectives and discusses their broader relevance to energy management in renewable systems. The chapter also evaluates the model's practical applications, highlights its strengths and limitations, and explores how it could inform future developments in intelligent energy systems. The discussion answers questions related to the technical use of machine learning in real world in sustainable energy practices.

5.1 Practical Implications of Energy Management

The forecasting model implemented gives way for energy efficiency and management for virtual powerplants with data backed solutions. With the help of consumption prediction, the system can analyse the efficient of use of its energy sources. Although this research has focused on biogas plant, the methodology can be conveniently applied to other systems as well. A better decision can be made how to store the energy and how to make the best use of it. With the integration of spot prices of electric grid, the system can make the optimized use of electricity. Hence, reducing the energy purchase from the grid as much as possible. This leads to economic advantage and promotes sustainable environment. A predictable consumption facilitates the biogas plant to better allocate the resources and operate economically.

The adoption of these forecasting models at large scale support demand response and load balancing. This plays an important role in integration of multitude of renewable resources and balancing the load easily. This not only enhances operational sustainability but also aligns with national and EU-level energy policies aimed at decarbonization and increased energy efficiency.

In the longer term, the integration of multiple such intelligent forecasting systems across biogas plants and other decentralized renewable sources could contribute to the development of virtual power plants (VPPs). These VPPs aggregate the capacity of various distributed energy resources and operate them as a single entity in the electricity market. The forecasting model proposed in this study, while limited in scope to a single plant, demonstrates the foundational elements necessary for such future-oriented

systems. Accurate, timely, and interpretable forecasts are essential for coordinating production, consumption, and storage decisions across a network of facilities. Thus, the implications of this work extend beyond the immediate cost savings and energy efficiency benefits—it also offers a stepping stone toward a more flexible, intelligent, and resilient energy infrastructure.

5.2 Challenges and Limitations

This study, which focused on forecasting the electricity consumption of a biogas plant in Finland using machine learning, faced several challenges and limitations that must be acknowledged to contextualize its findings and guide future research. A major constraint was the quality and scope of the data available. The electricity consumption dataset contained inconsistencies, missing entries, and possible sensor malfunctions, which required preprocessing and imputation. While these steps are necessary for ensuring the model can function, they also risk removing nuanced behaviours that might carry predictive significance. Additionally, the weather data used in this research was based on the nearest available location rather than the precise coordinates of the biogas plant. This approximation introduced a level of inaccuracy, particularly in temperature readings, which are known to influence energy consumption. Furthermore, the dataset used for training the model represented the behaviour of a single biogas facility, limiting the generalizability of the results to other regions, climates, or operational contexts. Consumption patterns may vary significantly across different types of facilities or under different energy regulations, making it difficult to extend the findings beyond this specific case.

Another significant limitation was the modelling approach and its computational feasibility. The forecasting model was specifically designed to forecast short-term electricity consumption over a 24-hour period, meaning that long-term forecasting capabilities were outside the study's scope. Features such as historical electricity usage and weather data were prioritized, but the absence of external influencing variables—such as policy shifts, energy prices, or human interventions—meant that the model could not capture broader contextual dynamics.

The study focused on practical results and therefore simple to evaluate different models for preliminary cases were developed. The primary purpose was to identify peak and off-peak hours and the pattern of consumption itself rather than accuracy of the prediction. The experiment was conducted carefully so that overfitting does not happen in the training phase leading to inaccurate results for unknown data.

5.3 Ethical Considerations

Several ethical concerns were raised during the technical implementation of this research. From an environmental ethics perspective, the study supports the transition to cleaner energy by contributing to improved forecasting accuracy in biogas-based Virtual Power Plants (VPPs), thereby promoting operational efficiency and reduced carbon emissions.

In terms of data ethics, all datasets used in this research did not contain any personal or sensitive information. The data was handled following the ethical obligations outlined in the thesis contract, including the student's role as a data controller and adherence to data protection regulations and university guidelines. All data from the commissioning organization were used solely for the purposes of this thesis, and any confidential material was treated accordingly.

The data processing practices were followed responsibly during all phases of the research. The process of data collections, cleaning and preprocessing allowed to data integrity and better performance results. Moreover, model robustness was ensured with validation and performance evaluation using standard metrics such as MAE, RMSE, and R^2 metrics. These measures were crucial both for technical accuracy and ethical responsibility since forecasting of in energy has direct impact on environment and economics.

This thesis brought about personal and academic development. Challenges such as limited data, computational constraints, and model generalizability required structured problem-solving and critical thinking. Project planning, time management, and communication with supervisors further contributed to the overall learning process. This experience has provided a strong foundation for future research and professional work in data-driven energy optimization.

6 CONCLUSIONS

The research investigated the process of data analysis from its inception and further moving to the optimization of Virtual Power Plants (VPPs) powered by biogas. The study identified key challenges in electricity forecasting and balancing it with the spot prices. The challenges were addressed to bring improvement to existing system with evaluation of various machine learning technologies. The key problem statements including grid optimization, predictive models and business analytics were discussed and how the optimization can be enhanced further. The research advocates for intelligent forecasting solutions that could fit in similar systems and therefore achieve solutions to improve energy efficiency, reduce operational costs and ultimately support the transition into greener energy.

The research demonstrated the potential of machine learning for enhancing electricity consumption forecasting in a biogas-based Virtual Power Plant (VPP). The analysis provided insights into temporal consumption patterns and highlighted the practical implications of optimized energy management, supporting the effectiveness of a data-driven approach. Although challenges related to data limitations, generalizability, computational resources, and prediction timing were identified, the findings establish a foundation for developing intelligent forecasting solutions. The implementation of such approaches in similar systems may improve energy efficiency, reduce operational costs, and contribute to the broader transition toward sustainable energy within VPP frameworks. Future work should address the identified limitations by incorporating more accurate data sources, expanding datasets across multiple sites, integrating extra variables and leveraging more advanced infrastructure to enable real-time learning and enhance prediction accuracy.

REFERENCES

Cadenas, E., & Rivera, W. 2010. Wind speed forecasting in three regions of Mexico using a hybrid ARIMA–ANN method. *Renewable Energy*, 35(12), 2732–2738. Available at: https://doi.org/10.1016/j.renene.2010.04.022. Accessed: 2 April 2025

Gulzat, T., Lyazat, N., Siladi, V., Gulbakyt, S., & Maksatbek, S. 2020. Research on predictive model based on classification with parameters of optimization. *Neural Network World*, *30*(5), 295–308. Available at: https://doi.org/10.14311/nnw.2020.30.020. Accessed 2 April 2025.

Kahlen, M. T., Ketter, W. & van Dalen, J. 2018-11. Electric vehicle virtual power plant dilemma: Grid balancing versus customer mobility. *In Production and Operations Management*, 27(11), 2054–2070. Available at: https://doi.org/10.1111/poms.12876. Accessed 2 April 2025.

Kaiss, M., Wan, Y., Gebbran, D., Unsihuay Vila, C. and Dragičević, T., 2025. Review on Virtual Power Plants/Virtual Aggregators: Concepts, applications, prospects and operation strategies. *Renewable and Sustainable Energy Reviews*, 211, p.115242. Available at: https://doi.org/10.1016/j.rser.2024.115242. Accessed 2 April 2025.

Kozhakhmet, S., Lyazat, S., Siladi, G., Gulbakyt, K., & Maksatbek, K. 2020. Biogas production and its role in the energy system of Finland. *Renewable and Sustainable Energy Reviews*, 119, 109589. Available at: https://doi.org/10.1016/j.rser.2019.109589. Accessed 2 April 2025.

Lester, J. N., Cho, Y., & Lochmiller, C. R. 2020. Learning to Do Qualitative Data Analysis: A Starting Point. *Human Resource Development Review*, 19(1), 94-106. Available at: https://doi.org/10.1177/1534484320903890. Accessed 2 April 2025.

Naval, N. and Yusta, J.M., 2021. Virtual power plant models and electricity markets: A review. *Renewable and Sustainable Energy Reviews*, 149, p.111393. Available at: https://doi.org/10.1016/j.rser.2021.111393. Accessed 2 April 2025.

Price, I., Sanchez-Gonzalez, A., Alet, F. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 2025. Available at: https://doi.org/10.1038/s41586-024-08252-9. Accessed 2 April 2025.

Schiermeier, Q. 2016-07-14. Germany's renewable revolution awaits energy forecast. *In Nature (London)*, 535(7611), 212-213. Available at: https://doi.org/10.1038/535212a. Accessed 2 April 2025.

Singaravel, S., Suykens, J. & Geyer, P. 2018-10-01. Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction. *In Advanced Engineering Informatics*, 38, 81–90. Available at: https://doi.org/10.1016/j.aei.2018.06.004. Accessed 2 April 2025.

Xia, Z., Zhang, R., Ma, H. & Saha, T. K. 2024. Day-ahead electricity consumption prediction of individual household–capturing peak consumption pattern. *In IEEE Transactions on Smart Grid.* Vol. 15, no. 3. IEEE, 2971–2984. Available at: https://doi.org/10.1109/TSG.2023.3332281. Accessed 2 April 2025.