Decoding Energy Consumption: Machine Learning Predictions & model comparison at East Melbourne Waste Water Treatment Plant

Ryerson University

Decoding Energy Consumption: Machine Learning Predictions & Model Comparison at East Melbourne Waste Water Treatment Plant

Sazia Afreen

Student Number: 500995892

CIND860: Advanced Data Analytics Project

Supervisor: Dr. Tamer Abdou

Date of Submission: 15.04.2024

Table of Contents

Acknowledgement

I would like to express my sincere gratitude to Dr. Tamer Abdou for his invaluable guidance and supervision throughout this project. His expertise and support have been instrumental in shaping the outcome of this research.

I am also deeply thankful to all my professors, teaching assistants, colleagues, program director, and academic coordinator in my certificate program for imparting valuable skills and knowledge that have greatly contributed to the success of this project. Their mentorship and encouragement have been invaluable.

Lastly, I extend my heartfelt appreciation to all those who have supported and encouraged me along the way. Your assistance and encouragement have been instrumental in the completion of this project.

Decoding Energy Consumption: Machine Learning Predictions & Model Comparison in East Melbourne Waste Water Treatment Plant

Abstract

Wastewater treatment plants (WWTPs) are vital for maintaining environmental sustainability by adhering to stringent effluent quality standards. However, their energy-intensive operations pose challenges in terms of cost and carbon emissions. This study focuses on addressing these challenges, particularly at the East Melbourne Waste Water Treatment Plant, aiming to enhance environmental compliance. Through a comprehensive analysis of six years of data (2014−2019), we investigate the correlation between various organic, hydraulic, and climatic parameters of wastewater and daily energy consumption (EC) at the WWTP. The research aims to identify influential parameters for EC prediction, evaluate the predictive performance of machine learning models using historical data and environmental factors, and compare the effectiveness and efficiency of different machine learning algorithms.

To achieve these objectives, a holistic approach was adopted by integrating traditional methods like ridge regression and support vector regression (SVR), ensemble methods such as random forest (RF), gradient boosting (GB), and XGBoost, and deep learning techniques like artificial neural network (ANN), convolutional neural network (CNN), and long short-term memory network (LSTM). Feature selection employed SHAP values to assess each feature's relevance in predicting outcomes. Leveraging XGBoost's capabilities, feature importance was evaluated and identified influential features. Models underwent rigorous validation via time series cross-validation and were evaluated using key performance metrics like root mean square

error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and median absolute deviation (MAD).

The study identifies key features for predicting the long-term daily Electrical Conductivity (EC) at the Melbourne East WWTP, notably month, total nitrogen, COD, average inflow, and average temperature. Random Forest emerged as the most effective model, followed by Gradient Boosting, Ridge Regression, and LSTM. Residual analysis confirms unbiased predictions of RF model, while confidence interval analysis validates predictive validity, covering all unseen test dataset points. These findings highlight the robustness and reliability of the Random Forest model for EC prediction in wastewater treatment processes, offering valuable insights for optimizing energy consumption and enhancing sustainability in WWTP operations.

The study's computational analysis relied on Python 3.9 within the Google Colab platform, utilizing essential packages and libraries including Pandas, NumPy, Scikit-learn, Keras, TensorFlow, Scikeras, Matplotlib, and Seaborn for data manipulation, modeling, evaluation, and visualization.

The source code is available at https://github.com/saziaa/Decoding_Energy_Consumption

Introduction

Wastewater treatment plants (WWTPs) in Australia shoulder a substantial portion of the financial burden associated with electricity expenses, consistently ranking among the top three expenses in water utility operations. With an estimated annual energy consumption of 1000 gigawatt hours (GWh), these facilities account for approximately 0.4% of the nation's total electricity usage (Gregory et al., 2012; Australian Energy Statistics, 2016). However, the energy requirements for sewage treatment exhibit wide variability, ranging from 150 to 1400 kilowatt hours (kWh) per million liters (1 ML) of sewage treated (Annual report 2020-2021). Notably, the East Melbourne Waste Water Treatment Plant handles 40% of the total sewage treatment workload, with up to 90% of energy allocated to critical processes such as activated sludge aeration, pumping, and sludge treatment (Office of Environment & Heritage, n.d.).

Addressing the mounting challenges posed by escalating energy costs and environmental considerations necessitates the implementation of strategies aimed at bolstering efficiency, advocating onsite power generation, and adeptly managing energy demand.

Ensuring the proper treatment and sanitation of wastewater is imperative for safeguarding public health and the environment. This task involves significant energy consumption within water and wastewater systems, encompassing both onsite activities such as electricity usage for pumping and aeration, as well as offsite processes like the production and transportation of materials and chemicals for treatment (Wang et al., 2020).

Water agencies and WWTP operators are increasingly recognizing the importance of implementing energy-saving tools and methodologies, including benchmarking and energy audit procedures (Belloir et al., 2015; Panepinto et al., 2016; Foladori et al., 2015). Energy audits, characterized by a systematic approach to understanding energy consumption patterns in industrial plants, are crucial for optimizing energy management practices within WWTPs. However, the effectiveness of energy audits can be significantly enhanced by accurate predictions of energy consumption. Forecasting energy consumption in WWTPs allows operators to anticipate fluctuations in energy usage and pinpoint areas where energy-saving measures can be implemented proactively. This proactive approach not only enhances the feasibility of energy audits but also amplifies their impact, resulting in more sustainable and efficient WWTP operation.

Efficient management of energy consumption within WWTPs is paramount for achieving sustainable development goals, optimizing resource allocation, and preserving environmental integrity (Ahmad & Chen, 2018, Longo et al., 2016). Therefore, accurate forecasting of energy consumption is crucial for facilitating informed decision-making, enabling effective power management, ensuring operational vigilance, and reducing the carbon footprint associated with increased energy usage. These insights not only guide decisions regarding infrastructure planning and financial allocations but also promote community engagement and align WWTP operations with principles of responsible resource management.

However, forecasting energy consumption WWTPs is complex, with various factors, such as aeration processes in activated sludge, wastewater quality, weather conditions, and operational circumstances, contributing to the complexity (De Gussem et al., 2014).

This study serves as a replication of the original research titled 'Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach' conducted by Bagherzadeh et al. in 2021.

The original study focused on investigating the impact of various parameters, including wastewater, hydraulic, and climate-related factors, on daily energy consumption (EC) at the East Melbourne WWTP. Utilizing data collected from 2014 to 2019, the research employed several feature selection algorithms to identify the most relevant variables for training machine learning (ML) models. Subsequently, artificial neural networks (ANN), Gradient Boosting Machine (GBM), and Random Forest (RF) algorithms were utilized to predict EC, with GBM demonstrating superior predictive performance. Additionally, the study assessed prediction accuracy by establishing a 95% confidence interval, revealing a reasonable prediction error band. Key findings underscored the significant influence of total nitrogen, chemical oxygen demand (COD), and inflow-flow on energy consumption within the WWTP.

After reviewing the original study, the motivation for conducting this replication research arises from the importance of validating and extending existing findings in the field of wastewater treatment plant (WWTP) energy consumption prediction. Replication studies play a

vital role in corroborating the reliability and generalizability of research findings, contributing to the overall robustness of scientific knowledge.

Furthermore, while the original study by Bagherzadeh et al. provided valuable insights into the factors influencing energy consumption in WWTPs and the efficacy of machine learning models for prediction, there were certain aspects that warranted further investigation and validation. In particular, the replication study sought to explore the generalizability of the original findings to different modeling techniques and evaluation metrics, as well as to assess the robustness of the predictions under alternative validation methodologies.

To achieve this objective, several key modifications were made to the methodology employed in the original study. While maintaining consistency with the feature selection techniques utilized by Bagherzadeh et al., the replication study introduced novel machine learning algorithms, including Ridge Regression, Support Vector Regression (SVR), XGBoost, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM), alongside the previously used models such as ANN, GBM, and RF.

Additionally, a fundamental departure from the original approach was the adoption of time series cross-validation for model validation, aiming to better capture the temporal dynamics inherent in WWTP energy consumption data. This change in validation methodology represents

a notable departure from the original study and offers a more robust evaluation of model performance across different time periods.

In addition to validating the effectiveness of machine learning models for energy consumption prediction in WWTPs, this replication study aims to assess their efficiency and robustness under different evaluation metrics and validation methodologies. While the original study primarily focused on assessing predictive accuracy using metrics such as RMSE, MAE, MSLE, R^2, and J^2, our replication study extends this analysis to incorporate evaluations of both efficiency and robustness.

Efficiency considerations are crucial for real-world applications, as they directly impact the scalability and practical feasibility of deploying predictive models in operational settings. By comparing the computational efficiency of different machine learning algorithms, this replication study aims to provide insights into the trade-offs between predictive accuracy and computational cost, aiding stakeholders in making informed decisions when selecting models for deployment in WWTPs.

Furthermore, the replication study explores the robustness of predictive models by subjecting them to alternative validation methodologies, such as time series cross-validation. Robustness refers to the ability of a model to maintain consistent performance across different datasets and validation settings, ensuring reliable predictions in diverse operational scenarios.

Through a comprehensive comparison of models in terms of both effectiveness and efficiency, as well as an assessment of their robustness under different validation methodologies, this replication study endeavors to offer a holistic understanding of the capabilities and limitations of machine learning approaches for energy consumption prediction in WWTPs.

By undertaking these modifications, this replication study seeks to not only validate the findings of the original research but also extend the understanding of the factors influencing WWTP energy consumption and the effectiveness of machine learning models in this domain.

The paper adopts a systematic approach to explore the application of various ML models in predicting energy consumption at the East Melbourne WWTP. Section 2 provides a review of relevant literature. Section 3 describes the dataset used in the study and the project workflow. Section 4 outlines the research methodology, encompassing exploratory analysis, data preprocessing, feature selection, model selection, and evaluation metrics. Section 5 presents the experimental results, including a summary of the data, experimental setup, validation procedures, performance evaluation of individual models in terms of effectiveness, robustness, and efficiency, and comparison of models. The summary of key findings , comparison of the study with other studies and limitations of the study are discussed in Section 6. Finally, conclusions and future directions for research in this field are addressed in Section 7.

Literature review

Energy consumption prediction plays a pivotal role in the sustainable management of wastewater treatment plants (WWTPs), where energy-intensive processes are central to treatment operations. Accurate forecasting of energy usage not only facilitates efficient resource allocation but also enables optimization strategies to enhance environmental sustainability and operational cost-effectiveness. Traditional methods of energy consumption estimation often rely on simplistic models or historical averages, which may lack the precision required for effective decision-making in dynamic operational environments. In recent years, machine learning (ML) techniques have emerged as promising tools for energy consumption prediction in WWTPs, offering the potential to capture complex relationships within the data and improve forecasting accuracy. This literature review aims to provide a comprehensive overview of existing research in this domain, examining methodologies, findings, and opportunities for future advancements.

The study by Karadimoset al.2023, examines energy consumption (EC) in wastewater treatment plants (WWTPs), using machine learning (ML) to estimate EC based on project data from Greek WWTPs. Through statistical and correlation analyses, key predictive variables are identified, and a neural network (NN) model is constructed using selected variables. The optimized NN model achieves a mean squared error (MSE) of $8.99899 \times 10{-}5$, with inputs including treatment capacity, flow rate, influent load, and served population. The study highlights the potential generalizability of these models beyond Greece, offering valuable decision-making tools for stakeholders to improve resource allocation and energy-efficient designs.

Boncescu et al. 2021 aims to forecast the energy usage of a wastewater treatment plant in Romania by considering factors such as flow rate, BOD, TSS, COD concentrations, and overall energy consumption. Logistic regression is employed for the mathematical model, utilizing input data spanning two years (2015-2017) from a Romanian treatment facility, comprising a randomized dataset of 403 entries.

Żyłka et al.2020, assessed the application of a least squares linear regression model to predict energy consumption (EC) in a Polish dairy wastewater treatment plant (WWTP). This endeavor led to improved energy utilization. Additionally, they noted the impact of air temperature and biological load as notable factors influencing EC.

Yang et al. 2020, developed a regression model to predict the annual energy consumption (EC) of wastewater treatment plants (WWTPs) under varying influent conditions. They found that influent flow rate and chemical oxygen demand (COD) concentration were the main correlated features with EC in bioreactors. Their model demonstrated high accuracy in predicting annual EC.

Using Bayesian semiparametric quantile regression (QR), Yu et al.2019, utilized wastewater parameters such as biochemical oxygen demand (BOD), COD, pH, and total nitrogen to estimate WWTP EC. Strong correlations were observed between the dependent variable and COD (0.96) and BOD (0.86), respectively.

Zhang et al.2021, applied a random forest (RF) model to predict WWTP EC. Their study utilized a dataset of 2387 records from the China Urban Drainage Yearbook. The RF approach exhibited satisfactory predictive capabilities, achieving an R^2 value of 0.702.

Oulebsir et al.2020, proposed an approach using an artificial neural network (ANN) to develop an optimal EC model for WWTPs employing the activated sludge process. The ANN model showed good results, with a coefficient of determination ranging from 90% to 92% during training and 74% to 82% during testing.

Applying the most proficient model led to substantial energy gains across a significant portion of the dataset. Zhang et al.2012 utilized an ANN model to estimate EC of pumps within WWTPs, facilitating pump system operation scheduling aimed at minimizing EC. The ANN model demonstrated satisfactory performance, with a mean absolute error (MAE) of 0.78 and a mean absolute percentage error (MAPE) of 0.02, indicating accurate predictions.

Das et al. 2021, employed advanced machine learning models to predict WWTP energy consumption, identified GRU as the best-performing model, and recommended its use for future analysis.

In summary, the literature demonstrates a growing interest in utilizing machine learning techniques to predict energy consumption in wastewater treatment plants, with promising results across various methodologies and model architectures.

Dataset Description

The study utilizes a diverse dataset obtained from both the Melbourne water treatment plant (Melbourne water database, 2021) and nearby airport weather stations (Melbourne airport weather station, 2021) to advance energy consumption prediction techniques. Covering a period of five years from January 2014 to June 2019, the dataset includes nineteen variables outlined in Table 1. These variables encompass essential factors such as power usage, biological attributes, hydraulic elements, and weather conditions. Data on water quality and biological parameters are meticulously gathered using sensor-based measurements, while weather data is sourced from the conveniently located Melbourne airport weather station near the treatment plant. This integrated dataset provides researchers with a comprehensive understanding of the complex relationships affecting energy consumption at the wastewater treatment plant. Additionally, the dataset offers valuable time-domain information to enhance prediction accuracy in the study. More detailed insights into the dataset can be found in the extensive research conducted by Bagherzadeh et al. (2021).

Table: 1 Variables with units

Group	Parameters	Unit
Hydraulic	Average Inflow (Q_{in})	m^3/s
	Average Outflow (Q_{out})	m^3/s
	Energy Consumption (Ec)	MWh
Wastewater	Ammonia (NH_4-N)	mg/L

	Biological Oxygen Demand (BOD)	mg/L
	Chemical Oxygen Demand (COD)	mg/L
	Total Nitrogen (TN)	mg/L
Climate	Average Temperature (T_{avg})	℃
	Maximum temperature (T_{max})	℃
	Minimum temperature (T_{min})	℃
	Atmospheric pressure (AP)	(hPa)
	Average humidity (H)	%
	Total rainfall and / or snowmelt (Pr)	mm
	Average visibility (VIS)	Km
	Average wind speed (WS_{avg})	Km/h
	Maximum wind speed (WS_{max})	Km/h
Time	Year (year)	–
	Month (month)	–
	Day (day)	–

The source of the data is as follows: https://data.mendeley.com/datasets/pprkvz3vbd/1

Project Workflow

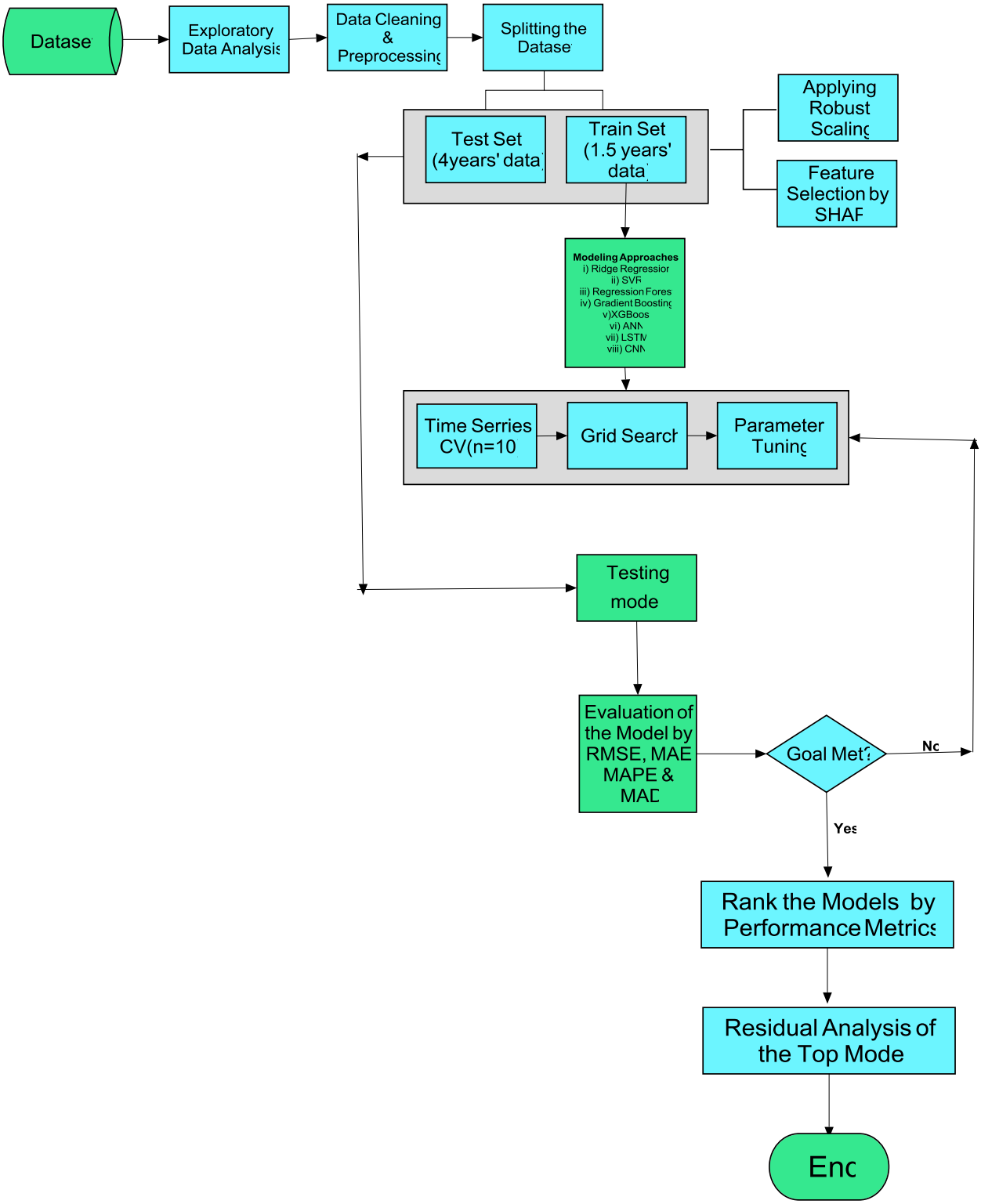

Fig1. Project Workflow

Research Methodology

This section outlines the methodology employed in the study to investigate the prediction of energy consumption in wastewater treatment plants (WWTPs) using machine learning techniques. The methodology encompasses data preprocessing, exploratory analysis, machine learning workflow, performance evaluation, and interpretation of results. Each step is detailed below to provide clarity on the research approach and analytical techniques utilized.

Exploratory Data Analysis (EDA):

Exploratory data analysis techniques were employed to gain comprehensive insights into the dataset. Summary statistics, data visualization, and correlation analysis were utilized to understand the relationships between variables and identify potential patterns or trends.

Upon examining the profile report, it becomes evident that the dataset features exhibit approximately normal distributions, characterized by moderate variability and skewness. Notably, there appears to be a negative association between the target variable, Energy_consumption, and certain features such as Avg_temp, Max_temp, Min_temp, Total_nitrogen, BOD, and Ammonia.

The correlation matrix further illuminates the interrelationships among variables. For instance:

- Avg_temp, Max_temp, and Min_temp display high correlations, indicating strong associations between these temperature-related features (correlation coefficients of 0.95 and 0.88, respectively).
- Similarly, a notable correlation exists among Avg_wind_speed, Max_wind_speed, and Max_gust_speed, underscoring their interconnectedness in measuring wind dynamics (correlation coefficients of 0.82 and 0.76, respectively).
- COD demonstrates moderate correlations with BOD (0.52) and Total_nitrogen (0.68). This observation is particularly significant as COD serves as a standard metric for assessing oxygen consumption during organic decomposition, while BOD quantifies the oxygen demand required for microbial degradation of organic matter. The moderate correlations suggest intertwined dynamics between these variables, where fluctuations in COD levels can influence both BOD concentrations and total nitrogen content. Specifically, increases in COD may contribute to elevated nitrogen levels due to the release of nitrogen compounds during organic decomposition.

In summary, the exploratory analysis reveals intricate relationships within the dataset, underscoring the interconnected nature of the variables and providing valuable insights into potential drivers of energy consumption in the studied context.

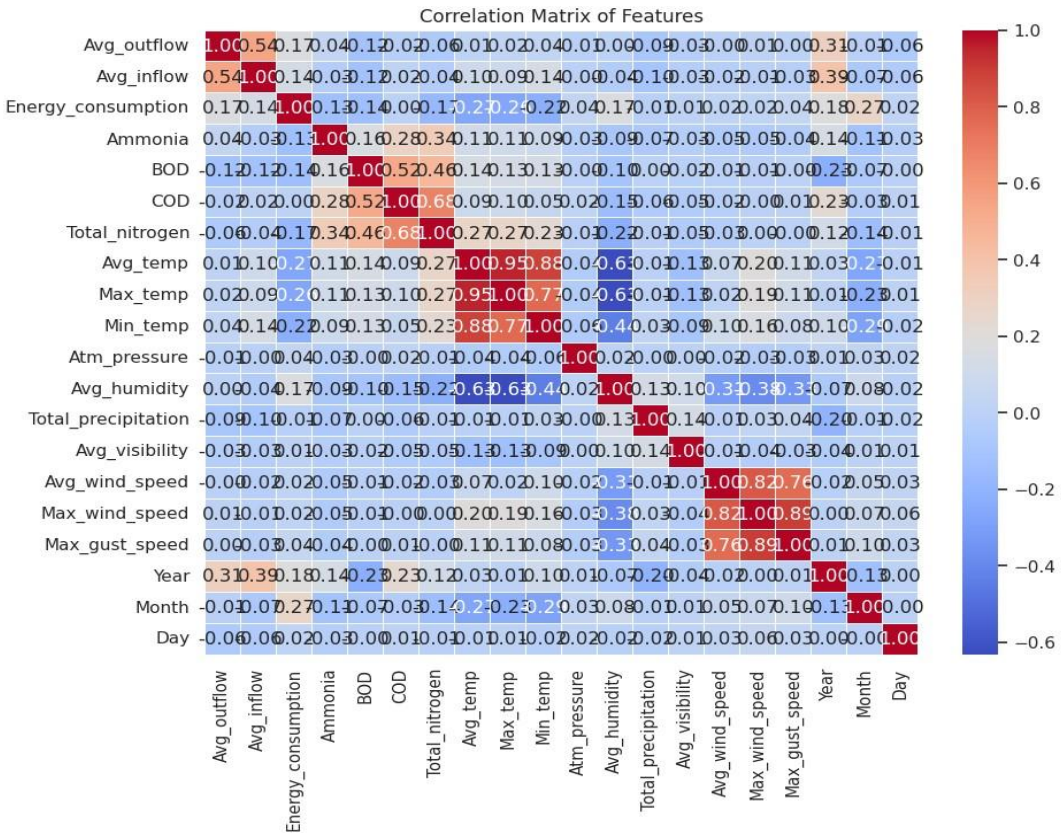

Fig 2. Correlation Matrix of the features

Data Preprocessing:

Prior to analysis, the raw data underwent thorough preprocessing to ensure its quality and suitability for further analysis. The preprocessing steps included handling missing and zero values, as well as addressing outliers.

i) Handling Zero Values:

Although no missing values were present in the dataset, zero values were identified in certain columns. These zero values were addressed by applying the K-nearest neighbors (KNN) method for imputation. Certain columns such as Total rainfall and/or snowmelt, Minimum

temperature, and Average visibility still contain zero value. Given the context of these variables, zero values were deemed logical and reasonable. Therefore, no imputation or removal of zero values was performed for these columns.

ii) Outlier Detection and Treatment:

Outliers can significantly affect model performance by skewing results and influencing predictions. Fig shows lots of outliers are present in almost all features. To identify and mitigate their impact, the Isolation Forest method was employed. This approach effectively isolates outliers by recursively partitioning the data into subsets until each observation becomes isolated. Following outlier detection, 5% of the total outliers from the dataset was removed to ensure a more robust and reliable analysis. This process will help to enhance the quality and accuracy of the models by minimizing the influence of extreme data points.

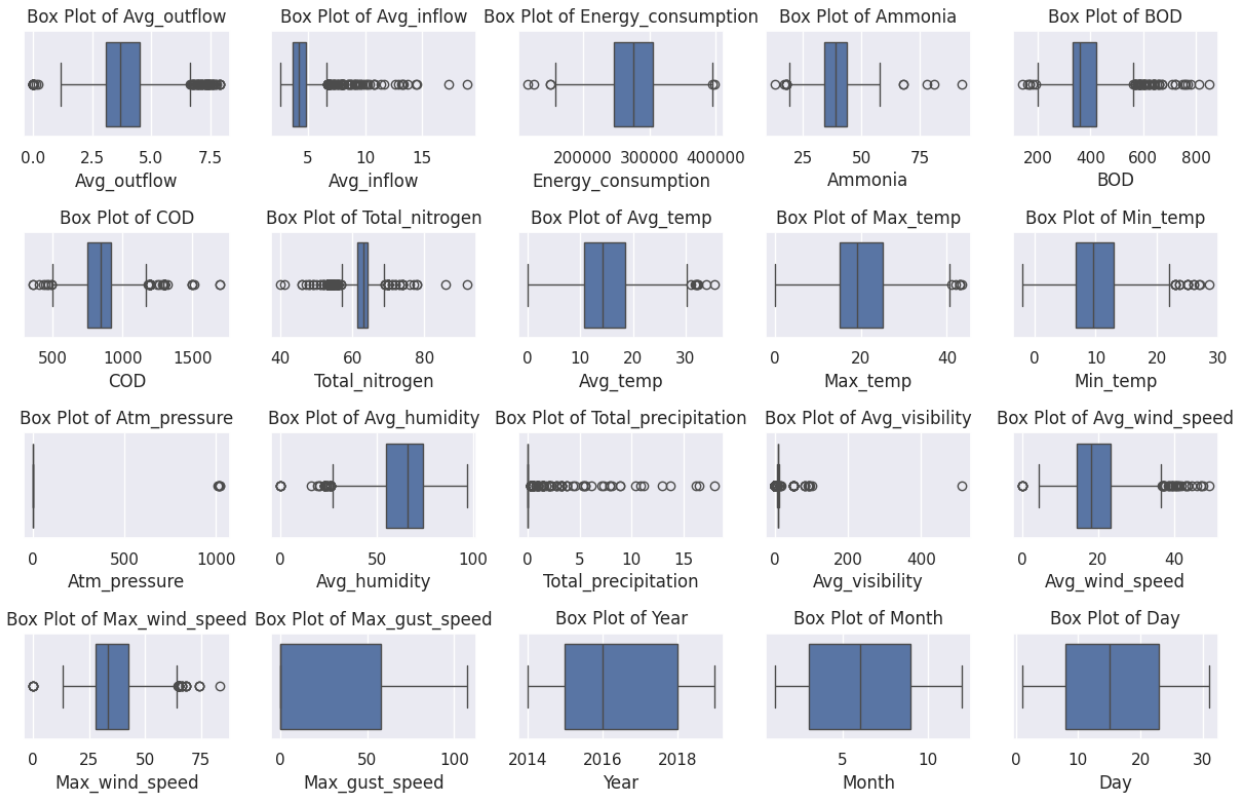

Fig 3. Outliers present in the features

iii) Data splitting:

The dataset is a time series dataset. It comprises records spanning 66 months. To keep the temporal pattern, 48 months' worth of data was allocated to the training set, while the remaining 18 months' data were reserved for the test set, maintaining approximately a 75-25 ratio.

iv) Feature Scaling:

Given the varying scales of features in the dataset, feature scaling was essential to ensure consistent model training. Robust scaling was selected due to the presence of outliers, as it is less susceptible to their influence compared to standard scaling techniques. With robust scaling, each feature's values are centered on the median and scaled by the interquartile range (IQR) instead of the standard deviation. This approach ensures that extreme values do not disproportionately affect the scaling process, resulting in more robust and reliable model performance. Fig shows the KDE plots of the features before and after scaling.

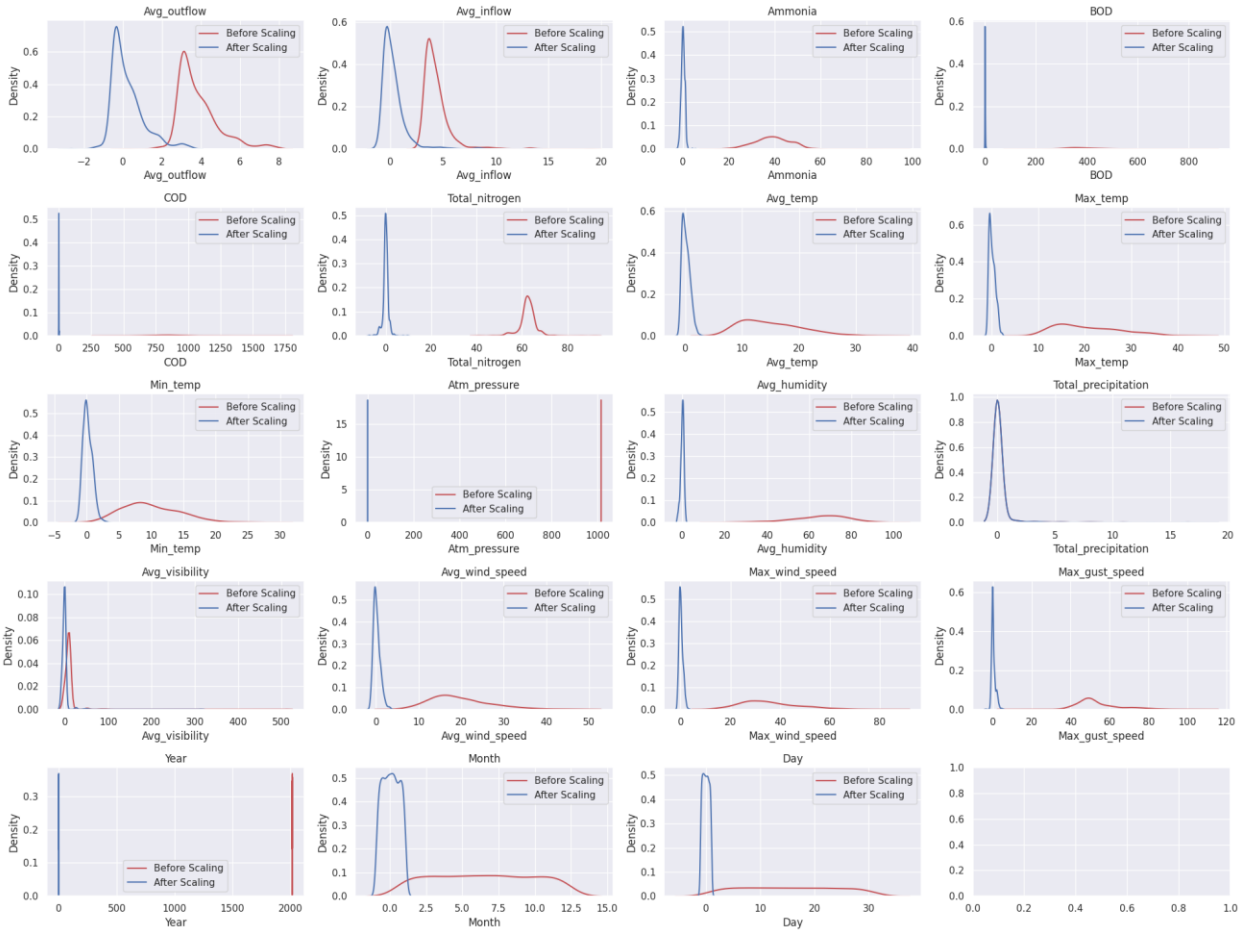

Fig 4. KDE plot of the features before and after Robust Scaling

Feature Selection:

Feature selection is a crucial step in the model-building process, aimed at identifying the most informative and relevant variables while eliminating redundant or non-informative ones (Luengo et al., 2020; Ranjan et al., 2021). In this study, the features were divided into four distinct groups: Time, Wastewater Parameters, Climate, and Hydraulic Parameters, each representing different aspects of the wastewater treatment process.

The time group encompasses temporal variables that capture the chronological aspects of the dataset. These variables include month, which reflects seasonal variations in energy consumption, as well as year, and day. Wastewater Parameters encompass variables related to the composition and characteristics of wastewater, such as Ammonia, Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD) and Total Nitrogen. Climate features consist of atmospheric variables like Average Temperature, Max Temperature, Min Temperature, Atmospheric Pressure, Average Humidity, Total Rainfall or Snowmelt, Average Visibility, Average Wind Speed, and Maximum Wind Speed. Hydraulic Parameters pertain to the flow rates of influent and effluent wastewater.

In this study, the feature selection process was refined by leveraging advanced techniques to identify the most influential predictors. Specifically, we employed SHAP (SHapley Additive exPlanations) values, a powerful tool for assessing the relevance of each feature in predicting outcomes. By analyzing the impact of individual features on model predictions, SHAP values provide valuable insights into feature importance. Using XGBoost the significance of each feature was assessed and determined the most influential ones. Initially, the top 13 features were considered based on their demonstrated influence. Among these, 'Month', 'Total_nitrogen', 'Avg_inflow', 'Avg_temp', 'Ammonia', 'Avg_humidity', 'BOD', 'COD', and 'Avg_wind_speed' emerged as key predictors. Notably, we omitted 'Max_temp' and 'Min_temp' due to their high correlation with 'Avg_temp'. Similarly, we prioritized 'Avg_inflow' over 'Avg_outflow' due to its higher importance. Additionally, we excluded the 'Year' variable, as its relationship with 'Month' rendered it redundant for our predictive modeling purposes. This refined feature selection strategy ensures that our models focus on the most impactful predictors, thereby enhancing their predictive performance and interpretability. During this process, we validated the XGBoost

model's performance using time series cross-validation, further ensuring the robustness of our feature selection approach. Fig shows the important features in their descending order.

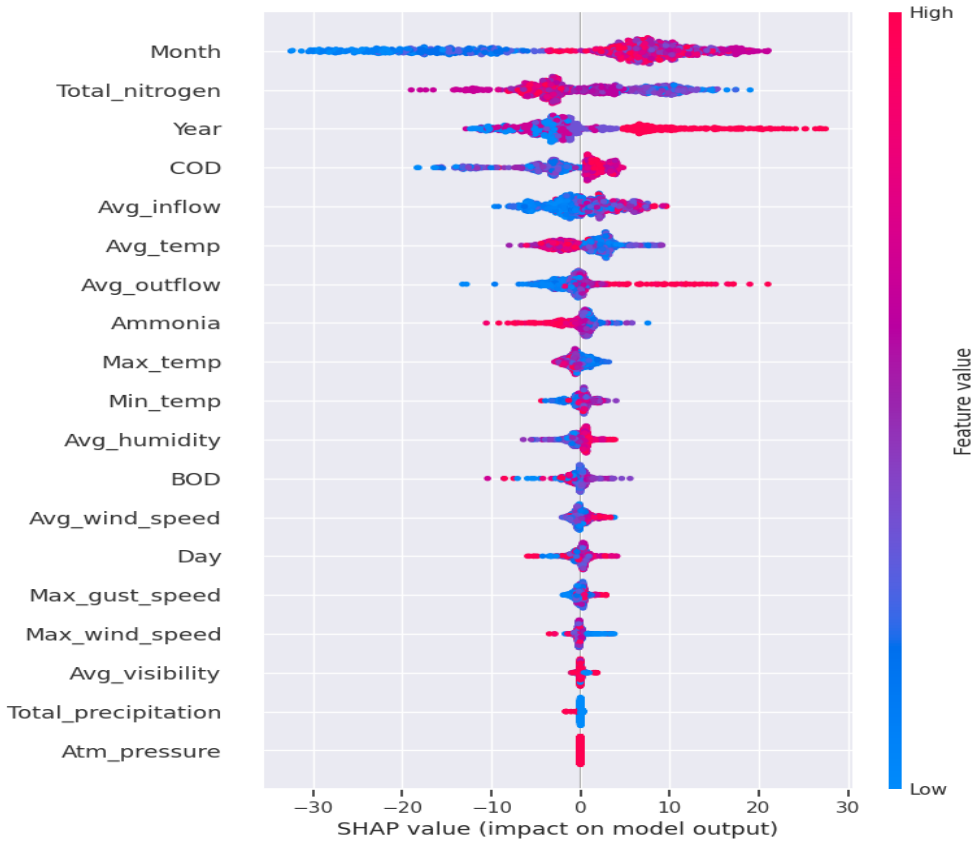

Fig 5. Important features in descending order

Model Selection:

In the model selection phase, a diverse set of machine learning algorithms was employed to construct predictive models for energy consumption.

The algorithms were chosen from three different groups: Traditional Method (Ridge Regression, Support Vector Regression), Ensemble Method (Gradient Boosting, XGBoost,

Random Forest) and Deep Learning Methods (Artificial Neural Networks, Long Short-Term Memory Networks, Convolutional Neural Networks).

Ridge Regression:

While investigating the applicability of multiple linear regression for the predictive modeling task, a series of hypothesis tests were conducted to evaluate the assumptions and significance of the predictor variables. The KS test indicated compatibility with the specified distribution, aligning with the assumptions of linear regression. However, the F-test revealed an extremely small p-value, providing compelling evidence against the null hypothesis and indicating the presence of at least one non-zero coefficient among the predictor variables.

Despite the overall significance of the model, further scrutiny uncovered that several features, such as Avg_outflow, Max_temp, Min_temp, and others, were statistically insignificant. This observation raised concerns about potential multicollinearity issues or the lack of importance in predicting the target variable. Consequently, the decision was made to opt for ridge regression over multiple linear regression.

Ridge regression offers notable advantages in addressing multicollinearity by incorporating a penalty term into the least squares objective function. This penalty term, typically represented as the $\ell 2$ penalty, helps stabilize the coefficient estimates and reduce their variance (Hoque et al.2022). Additionally, ridge regression effectively handles situations where

predictor variables are correlated or when certain features may lack statistical significance, as indicated by our hypothesis tests.

By leveraging the regularization properties of ridge regression, our aim is to enhance the robustness and generalization capability of the predictive model. Ultimately, this approach seeks to improve the model's performance in capturing the underlying relationships within the data, particularly in the presence of multicollinearity.

Support Vector Regression (SVR):

SVR, known for its adaptability and capacity to learn intricate patterns in data using kernel tricks, was chosen due to its robustness to outliers, which were retained in the dataset. Its fundamental principle involves mapping the training data into a higher-dimensional space and performing linear regression within that space. This technique enables SVR to effectively handle non-linear and non-separable datasets, as well as datasets with numerous features, while maintaining resilience against outliers.

However, leveraging SVR comes with computational demands, and its performance can be influenced by factors such as the choice of kernel function and regularization parameter (Yu et al., 2006; Hong et al., 2011). Despite these challenges, SVR remains a popular choice due to its ability to minimize structural risk, demonstrating reliable performance even with limited samples (Smola & Schölkopf, 2004). SVR is often considered less sensitive to outliers compared to some other regression techniques. This is because SVR focuses on finding a hyperplane that

fits the data within a certain margin (the ε-tube), and points outside this margin (outliers) have less impact on the model's performance. However, it's essential to note that SVR's robustness to outliers can vary depending on factors such as the choice of kernel function and the regularization parameter.

Ensemble Method (Gradient Boosting, XGBoost, Random Forest):

Ensemble methods are powerful techniques in machine learning that combine the predictions of multiple base models to produce a more accurate and robust final prediction (Zhou 2012). The underlying principle behind ensemble methods is that by aggregating the predictions of diverse models, they can compensate for individual model weaknesses and exploit their collective strengths. These models were selected based on their past success in similar domains and their ability to handle complex relationships in the data. Additionally, ensemble methods are known for their robustness and resilience to outliers.

Random Forest (RF):

Random Forest (RF) stands out as one of the most popular ensemble methods, known for its robustness and predictive power. This technique constructs multiple decision trees, with each tree trained on a random subset of the training data. During prediction, the outputs of individual decision trees are aggregated to generate the final prediction (Breiman, 2001; Lakshmanaprabu et al., 2019). The introduction of randomness in the training process helps alleviate overfitting and enhances the model's generalization capability.

Gradient Boosting (GB):

Gradient Boosting (GB) is a powerful ensemble method that constructs a robust predictive model by iteratively adding weak learners, typically decision trees, to the ensemble. Unlike Random Forest (RF), where trees are built independently, Gradient Boosting builds trees sequentially, with each tree aiming to correct the errors made by the previous ones. The learning process involves minimizing a loss function by iteratively fitting new trees to the residuals of the previous predictions (Natekin & Knoll, 2013). The final prediction is obtained by summing the predictions of all trees in the ensemble, weighted by a learning rate.

Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) represents an advanced iteration of Gradient Boosting Machines (GBM), integrating regularization techniques to mitigate overfitting and leveraging parallel processing to expedite training. Widely recognized for its efficiency, scalability, and superior performance on structured data, XGBoost has emerged as a favored choice in both competitive machine-learning arenas and real-world applications (Chen et al., 2020).

Distinguishing itself through the concept of "boosting trees," XGBoost orchestrates the training of each tree to rectify the errors of its predecessors via a gradient-based optimization

strategy. It offers a diverse range of regularization parameters, tree construction algorithms, and loss functions, providing users with granular control over the model's behavior and performance.

Artificial Neural Networks (ANN):

Artificial Neural Networks (ANNs) represent a versatile class of machine learning models renowned for their ability to discern intricate nonlinear relationships within data. Through the mechanism of backpropagation, errors are iteratively propagated backward through the network, allowing the model's parameters to be adjusted, thereby enhancing its performance over time.

ANN comprises interconnected layers of neurons, typically organized into three main layers: the input layer, hidden layers, and output layer. Within the architecture of an ANN, each neuron in a layer is fully connected to neurons in the previous and subsequent layers, facilitating the transmission of information across the network. The input layer receives raw input data, while the output layer generates the model's predictions or classifications.

The size of the input layer is typically determined by the number of independent input variables, while the output layer's size corresponds to the number of target features or output variables. Hidden layers, positioned between the input and output layers, play a pivotal role in capturing and representing complex relationships within the data.

Selecting the optimal size and number of hidden layers poses a challenge in ANN design and often necessitates experimentation to identify the most effective architecture. As model complexity increases, so does the size and intricacy of the hidden layers, underscoring the importance of meticulously tuning hyperparameters to achieve optimal performance (Raut and Dani, 2020; Tosun et al., 2016).

Convolutional Neural Networks (CNN):

Convolutional Neural Networks (CNNs) are commonly used in image processing tasks due to their ability to automatically learn hierarchical patterns and features from raw pixel data. However, CNNs can also be effectively applied to time series data as they can capture the temporal patterns and dependencies within localized segments of the series. In time series analysis, CNNs detect patterns and features irrespective of their temporal location.

The CNN model comprises three main layers: input, hidden, and output layers. The input layer typically receives a three-dimensional array, where dimensions represent height, width, and the number of channels. For a one-dimensional input of size N and no zero padding in the first layer, the feature output map is generated by convolving the input with a set of three-dimensional filters across all input channels. The resulting output is then passed through a non-linear activation function to produce feature maps (Barzegar et al, 2020).

In the hidden layer, which includes convolutional, pooling, and fully connected layers, the convolutional layer automatically extracts features across different regions of the input or intermediate feature maps using learnable filters. Neurons in this layer are connected only to a small region of the next layer, and shared weight matrices are employed for the convolutional operation, with weights updated during training (Hoseinzade & Haratizadeh, 2019). The pooling layer reduces the size of the input layer by converting values in the pooling window to a single value, typically through max-pooling (Zuo et al 2019). This layer aids in reducing computational costs and addressing overfitting (Hoseinzade & Haratizadeh , 2019).

In the hidden layer, the input feature map is convolved with a set of filters to create feature maps. The fully connected layer flattens the high-level features learned by the convolution layer and integrates them to obtain the final output. After passing through activation functions, the resulting feature values contribute to the final output. The output of the network, after multiple convolutional layers, is a matrix whose size depends on the filter size and number of filters used in the final layer (Borovykh et al.2017).

Long Short-Term Memory Networks (LSTM):

LSTM, a specialized form of RNN, tackles the issue of capturing long-term dependencies in sequential data. Its architecture incorporates memory cells and gating mechanisms, enabling the model to selectively retain or discard information from previous time steps, thereby facilitating the modeling of intricate temporal dynamics. LSTM has gained significant attention as a potent tool for time-series forecasting. Each memory cell within an LSTM is responsible for

storing and transmitting information across time steps. The essential components of an LSTM cell include the input gate, forget gate, output gate, and memory cell, which collaborate to control the information flow within the LSTM and determine which information should be remembered, forgotten, or outputted. The input gate regulates the storage of current input information in the memory cell, while the forget gate determines the amount of information from the previous cell state that should be discarded. The output gate governs the amount of information from the current cell state to be outputted. The candidate cell state represents a potential update to the current cell state, calculated based on the current input and previous hidden state. The final cell state is a combination of the previous cell state and the candidate cell state, weighted by the forget and input gates. The hidden state is derived by applying the output gate to the cell state after applying a hyperbolic tangent function. By leveraging memory cells and gating mechanisms, LSTMs can effectively capture and preserve relevant information over extended sequences, addressing challenges like the vanishing or exploding gradient problem often encountered in basic RNNs. This renders LSTMs particularly suitable for modeling complex temporal dependencies and capturing long-term patterns in sequential data, including energy consumption patterns in WWTPs (Harrou et al.2023).

Evaluation Metrics:

Accurately predicting energy consumption is essential for making informed decisions and managing resources effectively. To evaluate the performance of regression models, multiple assessment metrics are commonly used. In this case, we have selected four key metrics: Root

Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Median Absolute Deviation (MAD).

RMSE: RMSE measures the average magnitude of the errors between predicted and actual values. It penalizes larger errors more heavily than smaller ones due to the squaring operation. In the context of energy consumption prediction, RMSE provides a comprehensive understanding of the model's overall predictive accuracy. Since energy consumption values can vary significantly, RMSE gives insight into how well the model captures these variations.

MAE: MAE represents the average of the absolute errors between predicted and actual values. It is less sensitive to outliers compared to RMSE because it does not involve squaring the errors. MAE provides a straightforward measure of the model's average prediction error, making it easy to interpret. For energy consumption prediction, MAE offers a clear understanding of the typical magnitude of errors made by the model.

MAPE: MAPE calculates the average percentage difference between predicted and actual values, providing a relative measure of prediction accuracy. It is particularly useful when comparing the performance of models across different datasets or time periods. In the context of energy consumption prediction, MAPE allows stakeholders to assess the model's performance in terms of percentage accuracy, which can be valuable for decision-making and resource allocation.

MAD: MAD is a robust measure of variability that is less influenced by outliers compared to standard deviation. It is calculated as the median of the absolute deviations from the median of a dataset. MAD is particularly useful in situations where the data may contain outliers or exhibit non-normal distributions. It provides a more accurate representation of variability in such cases compared to standard deviation, which can be heavily influenced by outliers.

By considering these evaluation metrics collectively, a comprehensive understanding of the regression model's performance in predicting energy consumption can be gained. Each metric offers valuable insights into different aspects of prediction accuracy, enabling informed decisions regarding model deployment and optimization.

Experimental Results:

Summary of the data:

The table 2 shows the brief description of statistical properties of all the features after imputing the dataset for zero values.

Table 2: Summary of statistical properties of the features

Parameters	Max	Mean (50%)	Min	SD
Average Outflow (Q_{out})	7.920000	3.930608	0.000004	1.228778

Average Inflow (Q_{in})	18.968000	4.506338	2.589000	1.439583
Energy Consumption (Ec)	398.328000	275.159089	116.638000	44.640534
Ammonia (NH_4-N)	93.000000	39.222302	13.000000	7.761598
Biological Oxygen Demand (BOD)	850.000000	382.061708	140.000000	85.996012
Chemical Oxygen Demand (COD)	1700.000000	845.960434	360.000000	145.416540
Total Nitrogen (TN)	92.000000	62.740752	40.000000	3.571035
Average Temperature (T_{avg})	35.500000	15.036686	0.000000	5.398491
Maximum temperature	43.500000	20.530897	0.000000	7.096760

(T_{max})				
Minimum temperature (T_{min})	28.500000	10.037337	-2.000000	4.656887
Atmospheric pressure (AP)	1022.000000	3.675036	0.000000	61.010416
Average humidity (H)	97.000000	63.558611	0.000000	14.534205
Total rainfall and / or snowmelt (Pr)	18.030000	0.218090	0.000000	1.306923
Average visibility (VIS)	512.000000	9.101592	0.000000	16.322905
Average wind speed (WS_{avg})	49.100000	19.482706	0.000000	7.136868
Maximum wind speed (WS_{max})	83.500000	35.381259	0.000000	11.626362
Max_gust_speed (VG)	107.600000	29.184009	0.000000	31.366315

Year (year)	2019.000000	2016.258321	2014.000000	1.583491
Month (month)	12.000000	6.184515	1.000000	3.410588
Day (day)	31.000000	15.584660	1.000000	8.741422

Experimental Setup:

Initially, a baseline model was established to serve as a fundamental reference point for assessing the performance of more intricate models. This baseline model represents a simplistic approach, offering a benchmark against which the effectiveness of subsequent models is measured. While the baseline model provides a straightforward prediction strategy, it was anticipated that more advanced machine learning algorithms would surpass its performance by capturing intricate patterns and dependencies within the data.

During the construction of the Ridge regression model, an alpha value of 500 was chosen for regularization to control the extent of shrinkage applied to the coefficient estimates. This helps stabilize the model and reduce the impact of multicollinearity while ensuring that the model remains interpretable and generalizable.

IN SVR, a radial basis function (rbf) kernel was employed and performed grid search on hyperparameters 'C' and 'epsilon' with values [0.06, 0.07, 0.08] and [0.01, 0.1, 1], respectively. Time series cross-validation with n=10 folds was utilized for validation, ensuring the model's reliability across different time periods.

For the Random Forest model, we utilized hyperparameters obtained through grid search optimization. Specifically, we set the maximum depth of each tree to 10, with a minimum number of samples required to split an internal node set to 50 and the minimum number of samples required to be a leaf node set to 20. Additionally, the number of trees in the forest (n_estimators) was set to 1600 to ensure model stability and predictive accuracy. A random state of 42 was chosen for reproducibility purposes.

During the model selection phase, a grid search was conducted to identify the optimal hyperparameters for the Gradient Boosting model. The best parameters obtained from this search include a learning rate of 0.01, 500 estimators, a minimum sample split of 200, a minimum sample leaf of 100, a maximum depth of 15, and a subsample of 0.85, with a random state set to 42. These hyperparameters were chosen based on their ability to optimize the model's performance while avoiding overfitting.

A comprehensive grid search was conducted during model fitting to identify the optimal hyperparameters for the XGBoost model. The best parameters obtained from this search include a learning rate of 0.0001, 2000 estimators, a regularization alpha of 15, a maximum depth of 15, a minimum child weight of 4, a gamma value of 5, a subsample rate of 0.7, a column subsample rate of 0.7, and a random seed set to 42.

In the construction of the ANN model, a multi-layer architecture was employed including the following components:

- Input layer: The input layer received raw input data and had a size corresponding to the number of independent input variables in the dataset.

- Hidden layers: Multiple hidden layers were utilized, each comprising dense layers with rectified linear unit (ReLU) activation functions and dropout regularization to mitigate overfitting. The specific configuration included:

  First hidden layer: 50 neurons

  Second hidden layer: 50 neurons

  Third hidden layer: 50 neurons

  Fourth hidden layer: 20 neurons

- Output layer: The output layer generated the model's predictions and consisted of a single neuron, producing the final output.

This multi-layer architecture facilitated the effective capture and representation of complex relationships within the data. Additionally, L1 regularization with a regularization parameter of 0.1 was applied to each dense layer to further enhance model generalization and mitigate overfitting. By fine-tuning hyperparameters such as the number of epochs through grid search cross-validation, the performance of ANN was optimized.

The architecture used for CNN model construction consists of three Conv1D layers with 96 filters each, a kernel size of 2, and ReLU activation functions. Between the convolutional layers, AveragePooling1D layers with different pool sizes are added to reduce the dimensionality of the feature maps. The Flatten layer is used to convert the output of the convolutional layers into a one-dimensional array. Subsequently, two Dense layers with ReLU activation functions are added, with the first having 50 neurons and L1 kernel regularization, followed by a dropout layer with a rate of 0.3 to prevent overfitting. Finally, a dense layer with a single neuron is added

as the output layer. The model is compiled using the Adam optimizer and trained to minimize the mean absolute error loss function.

The architecture of the employed LSTM model consists of three LSTM layers, each with 150 units, Rectified Linear Unit (ReLU) activation functions, and kernel regularization with an L1 penalty of 0.6. Between the LSTM layers, Dropout layers with a dropout rate of 0.4 are added to prevent overfitting. Finally, a Dense layer with a single unit is added as the output layer. The model is compiled using the Adam optimizer with a specified learning rate and trained to minimize the Mean Absolute Error (MAE) loss function.

Table 3: The hyperparameters used by the models

Model	**Parameter**	**Value**
Ridge	Alpha(regularization parameter)	500
SVR	Kernel	RBF
	C	0.08
	Epsilon	0.01
RF	Number of estimators	1600
	Min sample leaf	20
	Min sample split	50
	Max depth	10

GB	Number of estimators	500
	Learning rate	0.01
	Min sample leaf	100
	Min sample split	200
	Max depth	15
XGB	Number of estimators	2000
	Learning rate	0.0001
	Reg_alpha	15
	Max depth	15
	Min_child_weight	4
	Gamma	5
	Sub_sample	0.7
	Col_sample_bytree	0.7
ANN	Input layer	50
	Hidden layer 1 (Relue)	50
	Hidden layer 2 (Relue)	50

	Hidden layer 3 (Relue)	20
	Output layer	1
	Loss	MAE
	Optimizer	Adam
CNN	Conv1D layer 1	96 filters
	Conv1D layer 2	96 filters
	Conv1D layer 3	96 filters
	Dense layer	50 neurons
	Dense layer(output)	1 neuron
	Kernel Regularizer	0.001
	Dropout	0.3
	Kernel size & pool size	2
	Loss	MAE
	Optimizer	Adam
LSTM	Input layer	(None,9,1)
	Hidden layer 1 (LSTM)	150

	Hidden layer 2 (LSTM)	150
	Hidden layer 3 (LSTM)	150
	Output layer	1
	Loss	MAE
	Optimizer	Adam
	Kernel Regularizer	0.6
	Dropout	0.4

Validation:

Validation of the models was performed using time series cross-validation (CV) to assess their predictive performance. Time series CV is particularly well-suited for sequential data analysis as it preserves the temporal order of the data during training and validation. The dataset was partitioned into 10 folds, with each fold representing a contiguous time period. Models were trained on historical data from earlier time periods and evaluated on subsequent time periods, ensuring that the evaluation process mimicked real-world scenarios where predictions are made on unseen future data. This approach helps assess the models' ability to generalize to new and unseen data while accounting for temporal dependencies. The performance metrics obtained from time series CV provide insights into the models' predictive accuracy and generalization capability across different time periods, enabling robust assessment of their effectiveness in capturing temporal patterns and dynamics.

Performance of Individual Models:

i) In terms of effectiveness:

The performance of each model was evaluated in terms of effectiveness, based on its ability to accurately predict the target variable. This assessment was conducted using a range of performance metrics, including RMSE, MAE, MAPE, and MAD, calculated on both the training and test datasets. Additionally, the consistency of performance across different evaluation metrics was considered to gauge the model's overall effectiveness in capturing the underlying patterns in the data.

Initially, the performance of the Naïve Baseline Model, which serves as a fundamental reference point for comparison, was evaluated using key metrics on the dataset.

The baseline model exhibited an RMSE of 48.255, indicating the average magnitude of errors between predicted and actual values. The MAE for the baseline model was 39.180, representing the average absolute difference between predicted and actual values. The MAPE was calculated at 14.231%, providing a relative measure of prediction accuracy. Additionally, the MAD for the baseline model was 34.499.

The Ridge Regression model was evaluated on both the training and test datasets. On the training set, the model achieved an RMSE of 38.714, MAE of 30.705, MAPE of 12.426%, and MAD of 25.001. However, on the test set, slightly higher error metrics were observed, with an RMSE of 45.581, MAE of 36.586, MAPE of 13.334%, and MAD of 32.709.

While the model performs reasonably well on both sets, there's room for improvement, particularly in reducing errors on the test set. Further optimization of parameters or exploration of alternative algorithms may enhance predictive performance

The Support Vector Regression (SVR) model was evaluated on both the training and test datasets. On the training set, SVR achieved an RMSE of 42.298, MAE of 33.565, MAPE of 13.496%, and MAD of 28.489. However, on the test set, slightly higher error metrics were observed, with an RMSE of 48.318, MAE of 39.282, MAPE of 14.227%, and MAD of 34.389.

The Random Forest model showed promising performance metrics on both the training and test datasets. On average, it achieved an RMSE of 31.06 and 43.77, MAE of 24.11 and 34.98, MAPE of 9.73% and 12.87%, and MAD of 19.51 and 30.03 on the training and test sets, respectively. While the model demonstrates effectiveness in capturing data patterns, there's potential for improvement, particularly in reducing test set errors.

The Gradient Boosting model exhibited respectable performance metrics on both training and test datasets. On average, it achieved an RMSE of 34.33 and 44.93, MAE of 27.03 and 36.17, MAPE of 10.93% and 13.28%, and MAD of 22.67 and 31.74 on the training and test sets, respectively. While showing effectiveness, there's potential for improvement, especially in reducing test set errors.

The XGBoost model achieved an average RMSE of 38.77 on the training set and 46.73 on the test set, indicating moderate prediction accuracy. Additionally, it obtained an average MAE of 30.66 on the training set and 37.66 on the test set, with an average MAPE of 12.38% on the training set and 13.67% on the test set. The model's average MAD was 25.79 on the training set and 32.87 on the test set. While demonstrating effectiveness, there's room for improvement, particularly in reducing test set errors.

The ANN model achieved an average RMSE of 39.69 on the training set and 45.64 on the test set, indicating moderate prediction accuracy. Additionally, it obtained an average MAE of

31.22 on the training set and 37.20 on the test set, with an average MAPE of 12.34% on the training set and 13.45% on the test set. The model's average MAD was 31.22 on the training set and 37.20 on the test set. While the ANN model demonstrates moderate performance with relatively low errors on the training set, there is potential for enhancement, especially in minimizing errors on the test set.

The CNN model demonstrates reasonable performance, as evidenced by its RMSE and MAE metrics on both the training and test sets. However, it's worth noting that the model's performance on the test set is comparable to the baseline model, suggesting that further refinement may be required to achieve significant improvements. Additionally, the relatively high MAPE and MAD values indicate that there may be room for enhancement in capturing the underlying patterns within the data. Further optimization of the model architecture or hyperparameters could potentially lead to better predictive performance.

The LSTM model demonstrates consistent performance across both the training and test sets, with average RMSE values of 43.19 and 45.69, and MAE values of 34.46 and 37.08, respectively. However, the model exhibits relatively higher MAPE and MAD values, indicating potential areas for improvement in capturing the relative and absolute deviations between predicted and actual values. Further refinement of the model architecture or hyperparameters may be necessary to enhance its predictive accuracy and robustness.

Table 4: The performance of different models

Model	Set	Evaluation Metrics			
		RMSE	MAE	MAPE	MAD
Naïve Baseline Model		48.255	39.180	14.231	34.499
Ridge Regression	Train	38.714	30.705	12.426	25.001
	Test	45.581	36.586	13.334	32.709
SVR	Train	42.298	33.565	13.496	28.489
	Test	48.318	39.282	14.227	34.389
RF	Train	31.061	24.110	9.726	19.510
	Test	43.765	34.977	12.871	30.032
GB	Train	34.334	27.026	10.933	22.666
	Test	44.932	36.167	13.284	31.740
XGB	Train	38.773	30.657	12.382	25.786
	Test	46.731	37.655	13.671	32.869
ANN	Train	39.686	31.220	12.338	31.220
	Test	45.636	37.204	13.454	37.204

CNN	Train	39.058	31.139	12.348	26.296
	Test	47.189	38.904	14.073	34.892
LSTM	Train	43.043	34.342	13.909	34.342
	Test	45.603	37.006	13.150	37.006

ii) In terms of robustness:

The stability and reliability of each model were assessed in terms of robustness, through cross-validation techniques, such as time series cross-validation. The dataset was divided into 10 folds and this approach helps determine the model's consistency in performance across different subsets of the data and its ability to generalize well to unseen data. Additionally, variations in performance metrics between training and test datasets were analyzed to gauge the model's resilience to overfitting and its capacity to maintain performance under varying conditions.

The tables below show the performance of the models in different folds of time series cross-validation.

Table 5: Performance Metrics for each fold for Ridge Regression

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	33.576	62.537	27.346	53.388	11.936	17.395	23.290	54.018

Fold 2	39.303	33.784	32.338	23.580	12.943	11.197	26.517	16.614
Fold 3	37.426	48.844	29.197	39.198	12.070	12.894	23.276	33.959
Fold 4	39.326	38.057	31.001	31.114	12.419	11.043	23.896	26.056
Fold 5	38.798	39.158	30.830	29.557	12.194	14.040	25.1438	24.715
Fold 6	38.279	41.296	30.139	31.891	12.151	14.797	24.281	26.734
Fold 7	38.553	56.918	30.247	46.932	12.375	15.083	24.404	40.508
Fold 8	40.210	46.201	31.514	38.029	12.632	13.119	26.015	35.289
Fold 9	40.885	43.713	32.221	36.084	12.818	12.031	26.647	35.663
Fold10	40.781	45.293	32.209	36.089	12.715	11.736	26.544	33.533
Average	38.714	45.580	30.704	36.586	25.001	32.709	12.425	13.334

Table 6: Performance Metrics for each fold for SVR

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	34.614	68.390	28.171	59.250	12.129	19.348	26.018	60.288
Fold 2	44.358	34.169	36.558	24.766	14.575	11.653	33.640	17.130

Fold 3	41.112	51.129	32.453	39.944	13.288	13.028	27.220	32.578
Fold 4	42.872	39.973	33.579	32.253	13.350	11.271	28.271	26.734
Fold 5	42.097	43.497	33.164	34.573	13.032	16.481	28.384	26.198
Fold 6	42.019	42.715	33.093	32.543	13.315	15.090	27.591	26.410
Fold 7	42.040	63.171	32.960	52.851	13.491	16.667	27.373	50.519
Fold 8	44.839	42.289	35.168	34.147	13.987	11.909	28.817	26.033
Fold 9	44.495	48.568	35.033	42.464	13.828	13.984	28.334	43.926
Fold10	44.533	49.270	35.471	40.030	13.961	12.838	29.242	34.067
Average	42.298	48.317	33.565	39.282	13.496	14.227	28.489	34.388

Table 7: Performance Metrics for each fold for RF

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	25.388	46.881	20.396	37.853	8.975	12.389	17.009	34.590
Fold 2	29.192	43.612	23.507	33.838	9.281	14.994	19.496	29.257
Fold 3	29.525	45.361	22.498	37.139	9.332	12.392	18.023	30.634

Fold 4	31.607	35.842	24.538	28.714	9.803	10.568	19.829	25.268
Fold 5	31.353	30.597	24.432	24.392	9.642	10.989	20.202	20.164
Fold 6	30.524	45.637	23.714	34.116	9.502	16.233	19.156	24.648
Fold 7	31.609	53.918	24.439	43.643	9.986	14.129	19.806	39.409
Fold 8	33.194	52.308	25.322	43.523	10.122	14.735	19.845	40.017
Fold 9	34.060	38.501	26.060	30.812	10.319	10.503	21.055	28.392
Fold10	34.157	44.988	26.190	35.738	10.291	11.778	20.679	27.940
Average	31.061	43.765	24.110	34.977	9.725	12.871	19.510	30.032

Table 8: Performance Metrics for each fold for GB

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	34.597	65.982	28.442	56.639	12.406	18.460	26.123	57.625
Fold 2	44.914	34.996	37.162	25.430	14.904	11.990	33.943	18.054
Fold 3	31.751	45.484	24.280	37.636	10.046	12.616	19.117	30.136
Fold 4	33.737	35.423	26.278	28.509	10.490	10.424	21.969	26.963

Fold 5	32.916	30.715	25.860	24.470	10.188	11.159	22.130	21.585
Fold 6	30.963	48.089	24.351	35.709	9.777	17.003	20.085	24.809
Fold 7	31.972	53.972	24.939	43.271	10.198	13.995	20.828	39.375
Fold 8	33.711	53.929	26.060	45.290	10.423	15.419	20.697	43.488
Fold 9	34.604	40.553	26.679	32.981	10.581	11.179	21.595	31.012
Fold10	34.173	40.177	26.208	31.734	10.313	10.595	20.176	24.348
Average	34.334	44.932	27.026	36.167	10.933	13.284	22.666	31.740

Table 9: Performance Metrics for each fold for XGB

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	31.395	62.698	25.629	53.120	11.186	17.272	22.545	53.081
Fold 2	40.031	33.203	32.814	23.448	13.144	11.078	29.467	16.154
Fold 3	37.625	49.264	29.544	38.209	12.171	12.478	25.205	31.603
Fold 4	39.193	38.395	30.587	31.024	12.221	10.940	25.822	26.282
Fold 5	38.550	41.804	30.321	33.134	11.979	15.787	25.478	26.201

Fold 6	38.560	42.535	30.258	32.262	12.175	15.052	25.020	24.234
Fold 7	38.762	62.399	30.265	52.109	12.367	16.449	25.391	49.004
Fold 8	41.264	42.644	32.198	34.565	12.878	12.013	26.240	27.768
Fold 9	41.095	46.290	32.257	39.980	12.811	13.217	26.174	41.548
Fold10	41.256	48.074	32.695	38.696	12.891	12.423	26.513	32.818
Average	38.773	46.731	30.657	37.655	12.382	13.671	25.786	32.869

Table 10: Performance Metrics for each fold for ANN

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	31.405	52.269	25.507	43.925	10.461	14.413	25.507	43.925
Fold 2	37.622	36.992	29.861	28.276	10.934	12.276	29.861	28.276
Fold 3	36.862	55.770	28.333	45.442	11.212	14.838	28.333	45.442
Fold 4	37.513	35.888	29.342	29.423	11.732	10.660	29.342	29.423
Fold 5	41.706	37.633	32.936	30.016	12.548	14.057	32.936	30.016
Fold 6	38.904	40.903	30.385	31.467	12.008	14.471	30.385	31.467

Fold 7	39.183	62.001	30.536	51.795	12.327	16.395	30.536	51.795
Fold 8	45.578	41.966	35.819	33.890	14.247	11.859	35.819	33.890
Fold 9	45.141	49.077	35.587	43.086	14.036	14.179	35.587	43.086
Fold10	42.942	43.862	33.891	34.716	13.871	11.384	33.891	34.716
Average	39.686	45.636	31.220	37.204	12.338	13.453	31.220	37.204

Table 11: Performance Metrics for each fold for CNN

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	40.206	70.552	33.781	63.512	13.837	21.102	28.602	63.567
Fold 2	39.518	35.614	32.196	28.564	12.292	12.221	27.756	23.152
Fold 3	34.122	46.363	26.964	38.200	11.165	12.925	23.454	37.214
Fold 4	37.047	41.926	29.414	32.335	11.525	11.779	26.215	24.856
Fold 5	38.746	34.432	30.767	28.709	11.643	12.741	24.964	27.761
Fold 6	36.944	40.421	29.074	32.054	11.626	14.607	24.544	26.230
Fold 7	37.346	53.726	29.476	44.113	12.103	14.668	25.209	38.337

Fold 8	41.446	43.310	32.325	33.930	13.419	12.219	25.316	29.330
Fold 9	41.266	45.517	32.016	36.972	12.776	12.323	26.387	36.353
Fold10	43.941	60.027	35.381	50.655	13.092	16.148	30.514	42.119
Average	39.058	47.189	31.139	38.904	12.348	14.073	26.296	34.892

Table 12: Performance Metrics for each fold for LSTM

Folds	RMSE		MAE		MAPE		MAD	
	Train	Test	Train	Test	Train	Test	Train	Test
Fold 1	34.694	59.691	28.704	50.419	12.652	17.093	28.704	50.419
Fold 2	44.918	45.606	37.137	36.955	14.856	13.034	37.137	36.955
Fold 3	41.916	45.493	33.316	36.848	13.919	13.006	33.316	36.848
Fold 4	43.557	42.708	34.374	34.284	13.926	12.390	34.374	34.284
Fold 5	43.027	45.431	33.954	36.788	13.082	12.990	33.954	36.788
Fold 6	42.790	46.015	33.817	37.346	13.441	13.140	33.817	37.346
Fold 7	42.726	44.890	33.584	36.285	13.763	12.859	33.584	36.285
Fold 8	45.574	43.503	35.819	35.014	14.257	12.550	35.819	35.014

Fold 9	45.931	40.946	36.549	32.681	14.932	12.191	36.549	32.681
Fold10	45.289	41.742	36.164	33.440	14.264	12.247	36.164	33.440
Average	43.042	45.602	34.342	37.006	13.909	13.150	34.342	37.006

iii) In terms of efficiency:

The efficiency of the models was assessed in terms of computational resources and time required for training and prediction. The time taken by each model for training and prediction is listed in the comparison section.

Comparison of Models:

Effectiveness Comparison:

Random Forest (RF) demonstrated remarkable consistency in its performance across all evaluated metrics, consistently securing the top rank. RF's ensemble learning approach, which aggregates predictions from multiple decision trees, enables it to effectively capture complex relationships and interactions within the data. Additionally, RF's inherent ability to handle nonlinearities, feature interactions, and high-dimensional datasets contributes to its superior predictive accuracy. The robustness of RF to overfitting and its resilience to noisy data further enhance its performance, making it a reliable choice for this energy consumption prediction task.

Following closely behind RF, Gradient Boosting (GB), Ridge Regression, and Long Short-Term Memory (LSTM) models exhibited commendable performance across the metrics, securing consistent rankings in the top tier. GB's sequential training of decision trees, where each

subsequent tree corrects the errors of the previous one, allows it to progressively improve predictive accuracy. Ridge Regression's regularization technique helps mitigate multicollinearity and overfitting, leading to stable and reliable predictions. LSTM, with its ability to capture long-term dependencies and temporal patterns in sequential data, proved effective in modeling the time-series nature of the dataset.

While these models may not have outperformed RF in every metric, their consistent rankings in the top tier underscore their efficacy and suitability for the prediction task at hand. Each model leverages unique strengths and techniques to achieve high predictive accuracy, demonstrating the importance of selecting the appropriate model based on the specific characteristics of the dataset and the prediction requirements.

SVR and CNN, while showing promise in certain scenarios, did not achieve competitive performance compared to other models in this evaluation. Several factors may have contributed to their relatively poor performance.

For SVR, its sensitivity to hyperparameters and the complexity of the dataset might have hindered its ability to generalize well to unseen data. Additionally, SVR's performance heavily relies on appropriate kernel selection and parameter tuning, which may not have been fully optimized in this study.

Similarly, CNN, although powerful in handling spatial information and capturing complex patterns, may not have been well-suited for the specific characteristics of the dataset. CNNs often require large amounts of data to effectively learn hierarchical representations, and insufficient data or suboptimal architecture design could have limited its performance.

Overall, the underperformance of SVR and CNN underscores the importance of model selection and tuning, as well as careful consideration of the dataset's characteristics and the model's suitability for the task at hand.

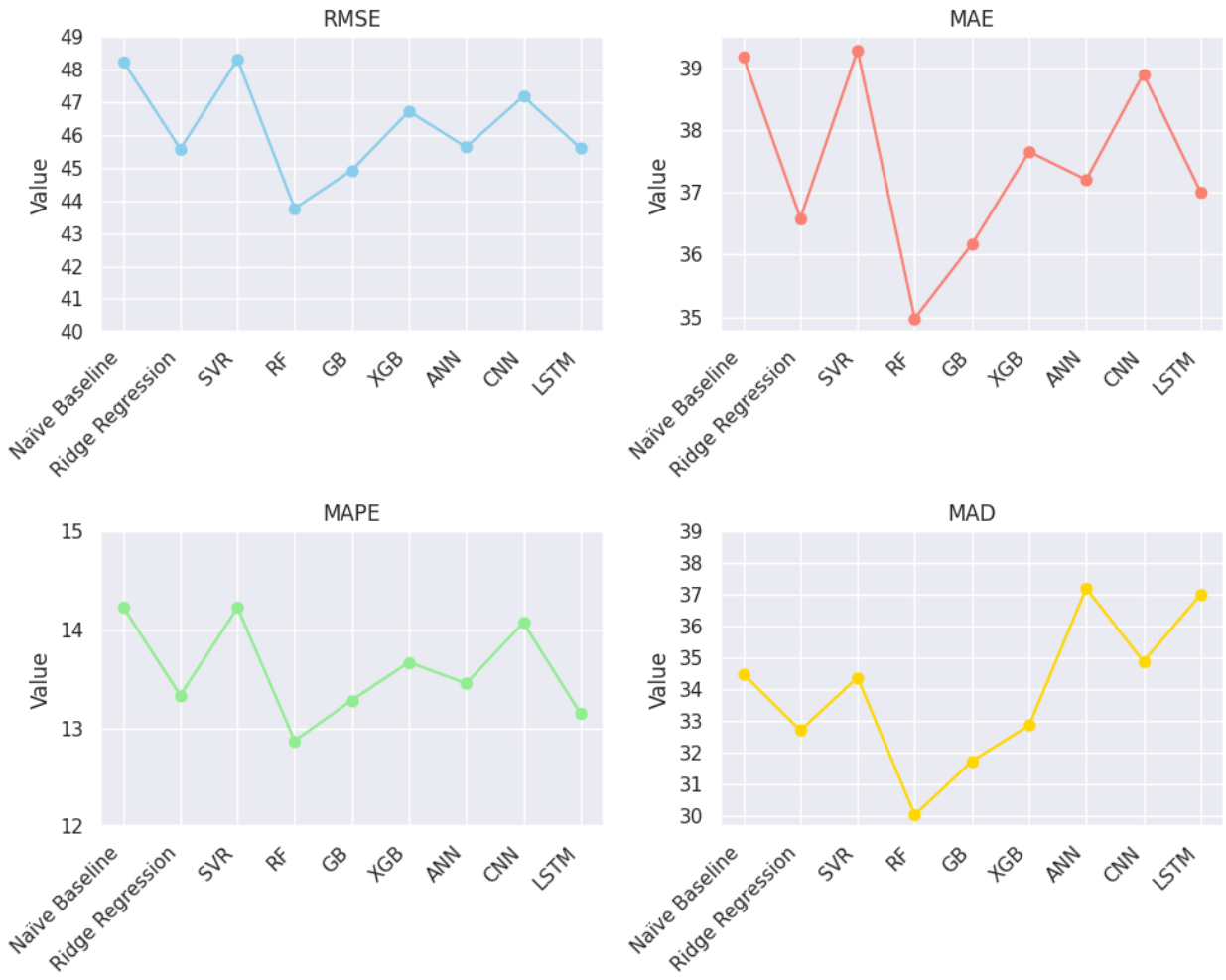

Fig 6. Comparison of the line graphs for each performance metric of different models

Table 13: Ranking of the models based on different performance metrics

Rank	RMSE	MAE	MAPE	MAD
1	RF	RF	RF	RF

2	GB	GB	LSTM	GB
3	Ridge Regression	Ridge Regression	GB	Ridge Regression
4	LSTM	LSTM	Ridge Regression	XGB
5	ANN	ANN	ANN	SVR
6	XGB	XGB	XGB	CNN
7	CNN	CNN	CNN	LSTM
8	SVR	SVR	SVR	ANN

Robustness Comparison:

Assessing the robustness of machine learning models is crucial in understanding their reliability and stability across different datasets or variations in data. A robust model should consistently deliver reliable predictions despite variations in the input data or sampling. In this section, we examine the robustness of each model by investigating its performance consistency across multiple folds or datasets. We utilize visualization techniques, statistical tests, and critical difference analysis to gain insights into the models' stability and reliability.

To assess the variability in model performance, box plots were constructed based on the root mean square error (RMSE) values of each model. Subsequently, a Friedman test was conducted to determine the significance of the performance comparison across models.

Additionally, critical difference (CD) values were calculated to elucidate the magnitude of performance differences between models. These analytical steps were instrumental in evaluating the robustness and significance of the observed differences in model performance.

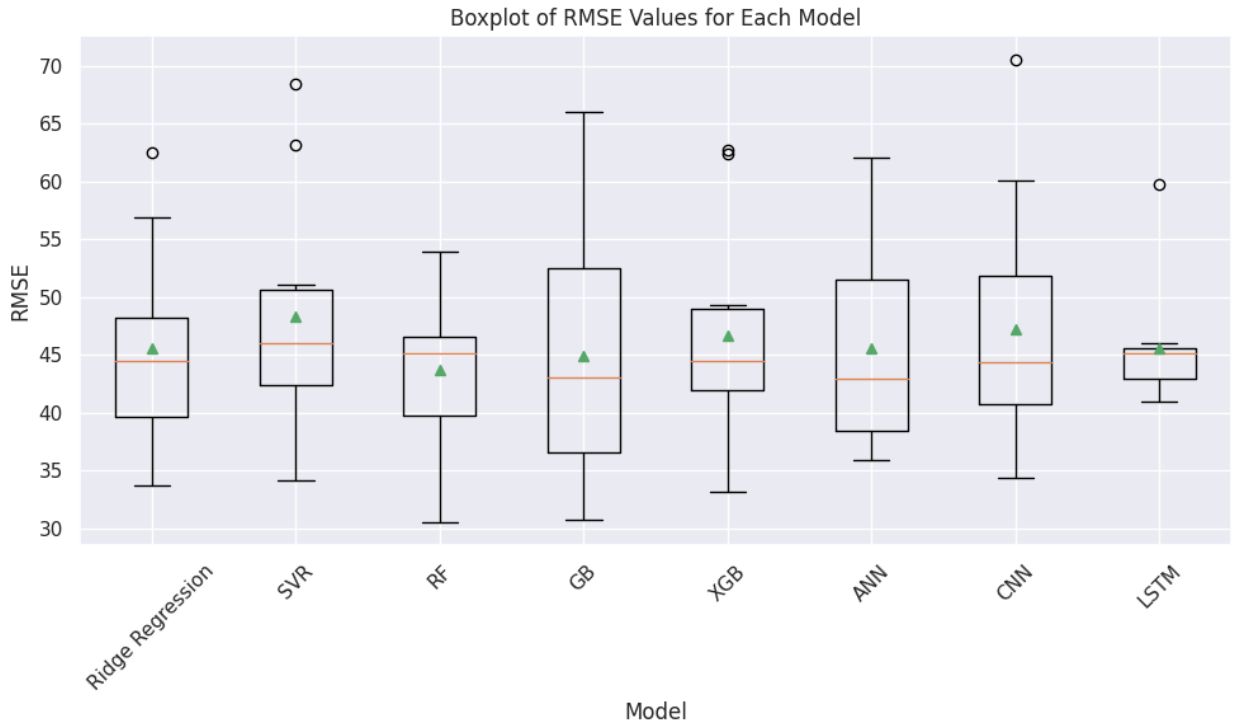

Fig 7. Box plot based on the RMSE values of each model

Based on the box plot analysis, the models with lower variability, such as LSTM, XGB, and RF, tend to exhibit less overfitting compared to models with higher variability. This suggests that these models are more robust and less prone to fluctuations in performance across different datasets or data variations.

The Friedman test results indicate a significant difference in the performance of the models (Friedman Statistic: 59.32, p-value: 8.19e-13), suggesting that at least one model significantly outperforms the others.

Additionally, the critical difference values provide insights into the pair wise comparisons between models. Models that are significantly different from each other indicate distinct performance characteristics. For example, Random Forest (RF) is significantly different from LSTM, ANN, XGB, CNN, and SVR, suggesting that RF performs differently and potentially better or worse than these models.

Interpreting these results, we can conclude that Random Forest demonstrates superior performance compared to many other models, including LSTM, ANN, XGB, CNN, and SVR. However, there are also significant differences observed between other models, such as GB, Ridge Regression, and SVR. These findings emphasize the importance of carefully selecting the appropriate model for the specific task at hand and considering the robustness of the model's performance across different evaluation metrics and datasets.

Efficiency Comparison:

Efficiency is a crucial aspect to consider when deploying machine learning models in real-world applications, as it directly impacts computational resources and operational costs. In this section, we evaluate the efficiency of each model by examining the time taken for training and prediction. The table below presents the time taken (in seconds) for both training and prediction tasks across different models:

Table 14: Efficiency of each model

Models	**Time taken for training and prediction (s)**
Naïve Baseline Model	0

Ridge Regression	1
SVR	9
RF	33
GB	7
XGB	42
ANN	240
CNN	180
LSTM	120

From the table, it is evident that the computational efficiency varies significantly across models. Models such as the Naïve Baseline Model, Ridge Regression, and GB exhibit relatively low computational overhead, with training and prediction times ranging from a few seconds to under a minute. On the other hand, more complex models like ANN, CNN, and LSTM require substantially more time for training and prediction tasks, with durations exceeding several minutes.

Discussion:

Summary of Key Findings:

The evaluation of the effectiveness, robustness, and efficiency of various machine learning models for predicting energy consumption in WWTPs yielded several noteworthy

insights. Firstly, in terms of effectiveness, Random Forest (RF) emerged as the top-performing model, consistently outperforming others such as Gradient Boosting (GB), Ridge Regression, and Long Short-Term Memory (LSTM). Conversely, Support Vector Regression (SVR) and Convolutional Neural Network (CNN) demonstrated relatively poorer performance across the metrics evaluated.

When considering the robustness of the models, LSTM exhibited the highest degree of stability, followed by XGBoost (XGB) and RF. Conversely, models such as GB, CNN, and Artificial Neural Network (ANN) exhibited greater variability in their performance across different evaluation scenarios.

Regarding efficiency, Ridge Regression and GB demonstrated low computational overhead, making them efficient choices for practical deployment. However, models such as ANN, CNN, and LSTM incurred significantly higher computational costs, potentially limiting their scalability in real-world applications.

Since RF emerged as the top-performing model in terms of effectiveness, a detailed analysis of the RF model's predictive accuracy was conducted through residual analysis.

The analysis revealed valuable insights into the model's predictive accuracy and potential areas for improvement.

The residuals, which represent the differences between the actual and predicted energy consumption values, were examined for patterns or systematic errors. We found that the residuals were mostly randomly distributed around the zero line, indicating that the model's

predictions were mostly unbiased. This suggests that the RF model effectively captured the underlying patterns in the data.

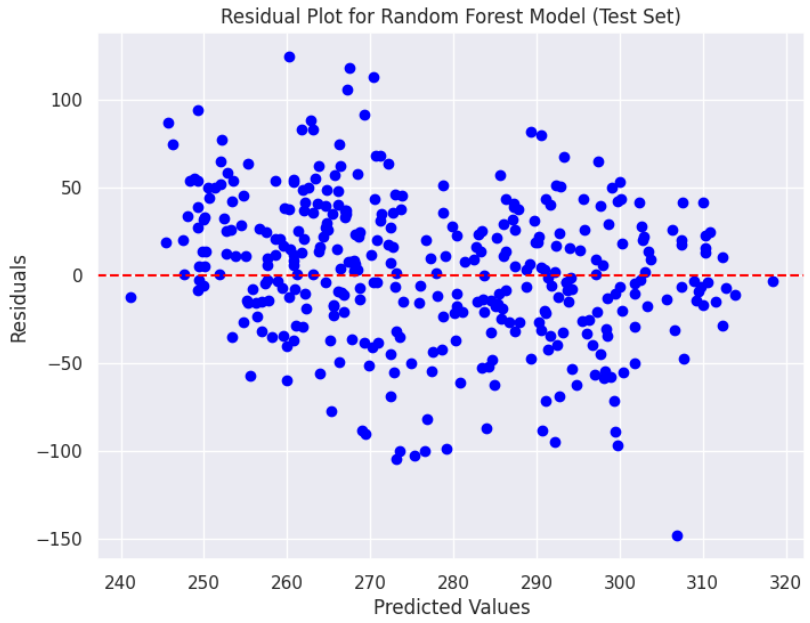

Fig 8. Residual plot of RF model

To visually assess the distribution of residuals, we plotted a histogram and a QQ plot. The histogram displayed the frequency distribution of residuals, while the QQ plot compared the quantiles of the residuals to those of a theoretical normal distribution. While the histogram showed a relatively symmetric distribution of residuals, the QQ plot revealed deviations from the straight line in the tails of the distribution. This departure from normality suggests the presence of outliers or heavy-tailed distributions in the residuals, which may warrant further investigation.

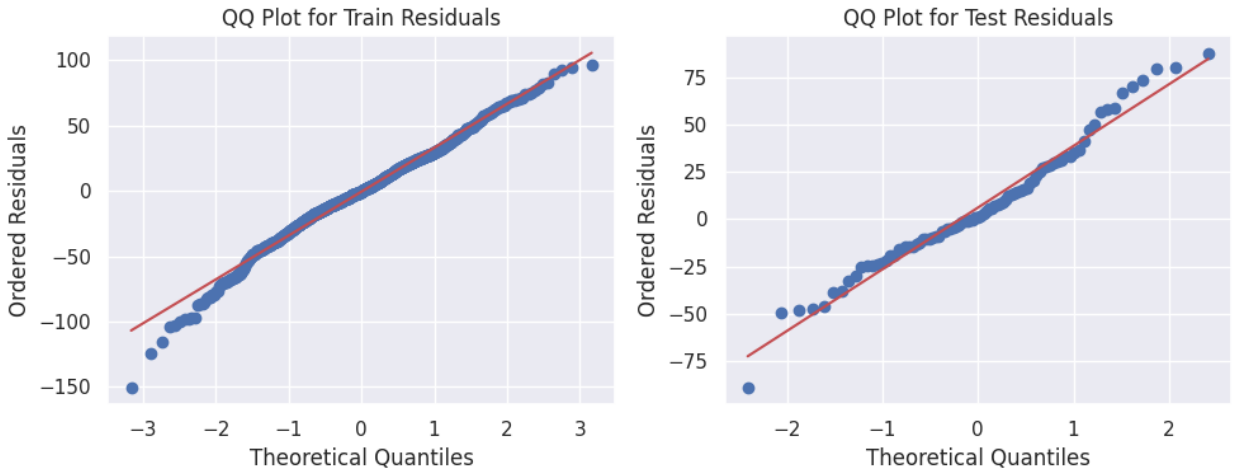

Fig 9. QQ plot for the residuals of RF model

Furthermore, we calculated a 95% confidence interval for the mean of the test residuals, providing insights into the variability of prediction errors. The confidence interval ranged from approximately -0.64 to 12.09, indicating that we are confident that the true mean of the test residuals falls within this range. Additionally, the mean value of 6.13 suggests that, on average, the residuals for the test set are higher than those for the training set. This observation highlights the variability in prediction errors when making predictions on unseen data.

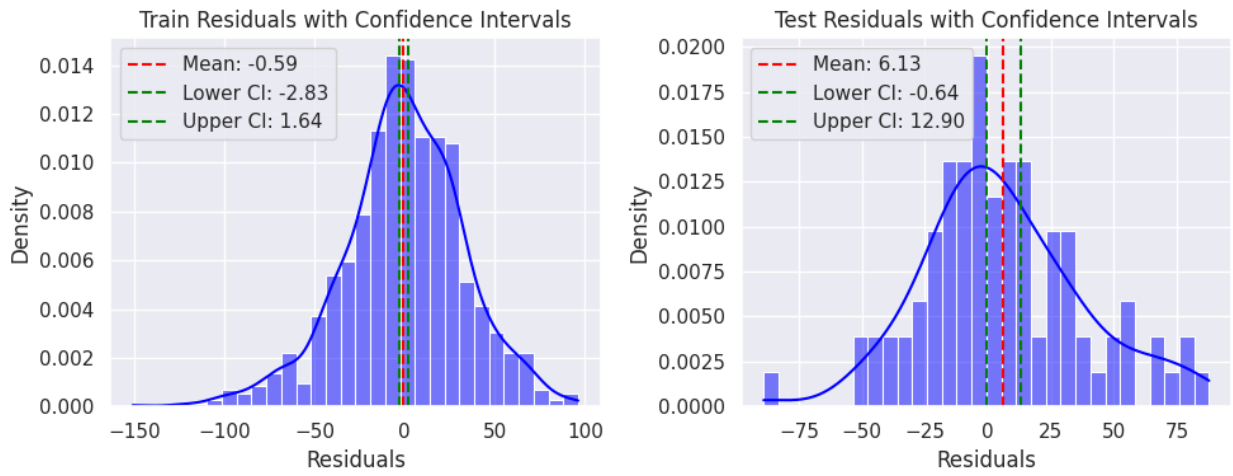

Fig 10. Residuals of RF with confidence intervals

Finally, we visualized the actual energy consumption values alongside the predicted values generated by the RF model, along with a 95% confidence interval. This visualization provides a clear comparison between the model's predictions and the ground truth, allowing us to assess the model's performance visually.

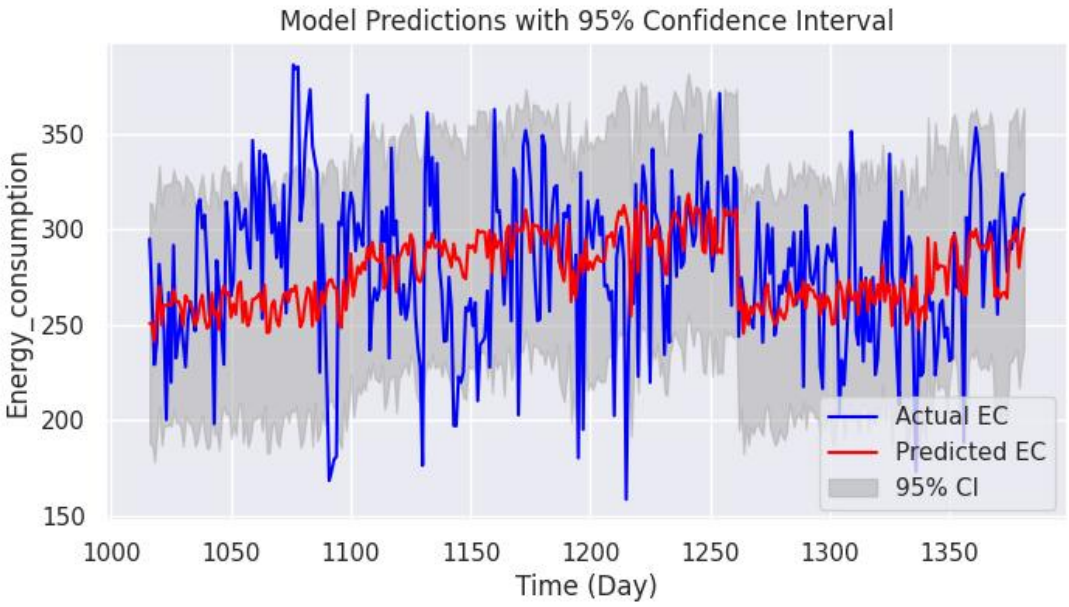

Fig 11. Model prediction with 95% confidence interval

Comparison with previous studies:

The performance of the models used in this study (Ridge Regression, SVR, RF, GB, XGB, ANN, CNN, and LSTM) is compared with findings from previous studies on WWTP energy consumption prediction. Table X presents a summary of the comparison, including the performance metrics used and the results obtained.

In a study conducted by Bagherzadeh et al. (2021), prediction results for GBM, RF, ANN, and RNN models were obtained based on selected features. Alali et al. (2023) investigated various machine learning techniques, including KNN, XGBoost, LightGBM, GPR, SVR, and Bagged Tree, using selected relevant features and lagged 1 energy consumption data. Harrou et al. (2023) significantly outperformed other machine learning approaches by employing data augmentation techniques, allowing for the capture of finer temporal details and variations in WWTP variables. The augmented data provided a richer representation of underlying patterns and dynamics, enhancing the deep-learning models' ability to learn from an expanded range of temporal dependencies.

Table 15: Comparison of the predictive performance of the models reported in the literature with this study.

Methods	**RMSE (MWh)**	**MAE (MWh)**	**MAPE(%)**	**R2**	**MAD (MWh)**	**Reference**
GBM	33.9	26.9	–	0.18	-	Bagherzadeh et al. 2021
RF	34.8	27.7	–	0.14	-	
ANN	39.8	32.1	–	0	-	
RNN	37.3	29.3	–	0.01	-	
KNN	37.33	28.23	10.65	–	-	Alali et al. 2023
XGBoost	37.14	28.5	10.81	–	-	

LightGBM	37.38	28.63	10.96	–	-	
GPRRQ	37.45	28.65	10.04	–	-	
GSVR	37.7	28.88	10.12	–	-	
BT	37.56	28.75	10.27	–	-	
RNN	7.498	5.924	2.182	0.964	-	Harrou et al.2023
GRU	6.513	4.698	1.738	0.973	-	
LSTM	5.821	3.685	1.36	0.978	-	
BiGRU	5.921	3.912	1.436	0.978	-	
BiLSTM	6.273	4.165	1.533	0.975	-	
BiLSTM	6.273	4.165	1.533	0.975	-	
Ridge Regression	45.581	36.586	13.334	-	32.709	This study
SVR	48.318	39.282	14.227	-	34.389	
RF	43.765	34.977	12.871	-	30.032	
GB	44.932	36.167	13.284	-	31.740	
XGBoost	46.731	37.655	13.671	-	32.869	
ANN	45.636	37.204	13.454	-	37.204	

CNN	47.189	38.904	14.073	-	34.892	
LSTM	45.603	37.006	13.150	-	37.006	

Limitations of the Study:

While this study aimed to provide comprehensive insights into energy consumption prediction in wastewater treatment plants (WWTPs) using machine learning models, several limitations should be considered when interpreting the results.

Limited Data: The study's reliance on a finite dataset suggests that increasing data volume or employing data augmentation techniques could enhance prediction accuracy. However, it's essential to acknowledge that data augmentation may introduce bias.

Model Architecture and Hyperparameter Tuning: The choice of model architecture and the process of hyperparameter tuning are crucial factors influencing predictive performance. Ensuring optimal model selection and parameter tuning is imperative for achieving accurate predictions.

Data Preprocessing: The methods employed for imputation and outlier removal during data preprocessing can significantly impact prediction outcomes. Careful consideration and validation of these preprocessing techniques are essential to minimize distortions in the dataset.

Time Series Stationarity: Given the dynamic nature of time series data, ensuring stationarity assumptions are met is vital for accurate predictions. Failure to account for temporal dependencies and changes in underlying patterns could affect the model's predictive capabilities.

Conclusion:

In conclusion, this study provides valuable insights into the application of machine learning models for energy consumption prediction in wastewater treatment plants (WWTPs). The study revealed that 'Month', 'Total_nitrogen', 'Avg_inflow', 'Avg_temp', 'Ammonia', 'Avg_humidity', 'BOD', 'COD', and 'Avg_wind_speed' are key predictors in predicting energy consumption in WWTPs. Through comprehensive evaluation, Random Forest is identified as the most effective model, demonstrating superior predictive accuracy compared to other models. The analysis highlighted the importance of considering model robustness and efficiency, with LSTM exhibiting the highest stability and Ridge Regression demonstrating low computational overhead.

Future Directions for Research:

Moving forward, several avenues for future research emerge from this study. Firstly, investigating ensemble methods that combine the strengths of multiple models could potentially enhance predictive performance further. Additionally, exploring advanced techniques for data augmentation and preprocessing, such as generative adversarial networks (GANs) and feature engineering, may offer new insights into improving model accuracy. Moreover, incorporating additional environmental variables into the modeling process could enhance the models' predictive capabilities in real-world scenarios. Furthermore, exploring the potential of

explainable AI techniques to improve model interpretability and stakeholder trust warrants further investigation.

Overall, by addressing these research directions, future studies can contribute to advancing the field of energy consumption prediction in WWTPs and ultimately support sustainable management practices in wastewater treatment facilities.

References

Ahmad T., Chen H. (2018).Utility companies' strategy for short-term energy demand forecasting using machine learning based models. *Sustain. Cities Soc. 39,* 401–417.

Alali Y., Harrou F., Sun Y. (2023). Unlocking the potential of wastewater treatment: machine learning based energy consumption prediction, *Water, 15* (13) , p. 2349

Bagherzadeh F., Nouri A.S., Mehrani M.J., Thennadil S. (2021) Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach. *Process Saf. Environ. Prot., 154,* 458–466.

Bagherzadeh F., Mehrani M.-J. , Basirifard M. , Roostaei J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance, *J. Water Process. Eng., 41* Article 102033

Belloir C., Stanford C., Soares A. (2015). Energy benchmarking in wastewater treatment plants: The importance of site operation and layout. *Environ Technol*, 36 (2), pp. 260-269

Breiman, L. (2001). Random forests. *Mach. Learn. , 45*, 5–32.

Barzegar, R., Aalami, M.T. & Adamowski, J. (2020) Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stoch Environ Res Risk Assess 34*, 415–433.

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*.

Boncescu C., Robescu L. D., Bondrea D. A., Măcinic M. E.(2020) Study of energy consumption in a wastewater treatment plant using logistic regression, *IOP Conf. Ser.: Earth Environ. Sci.* 664 012054 DOI 10.1088/1755-1315/664/1/012054

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2020). Extreme Gradient Boosting [R package xgboost version 1.2.0.1].

Das A., Kumawat P.K., Chaturvedi N.D. (2021). A Study to Target Energy Consumption in Wastewater Treatment Plant using Machine Learning Algorithms,*Computer Aided Chemical Engineering, Elsevier,Volume 50,* Pages 1511-1516, ISSN 1570-7946, ISBN 9780323885065, https://doi.org/10.1016/B978-0-323-88506-5.50233-3.

Foladori P., Vaccari M., Vitali F. (2015). Energy audit in small wastewater treatment plants: methodology, energy consumption indicators, and lessons learned, *Water Sci Technol, 72* (6), pp. 1007-1015

Gregory, A.G., Hall, M.R., & Cook, S. (2012). Energy use in the provision and consumption of urban water in Australia: an update.

Gussem K. D., Fenu A., Wambecq T. , Weemaes M. (2014).Energy saving on wastewater treatment plants through improved online control: Case study wastewater treatment plant Antwerp-South, *Water Sci. Technol., 69,* pp. 1074-1079

Hong W.C., Dong Y., Chen, L.Y., Wei S.Y. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Appl. Soft Comput. , 11*, 1881–1890.

Harrou F., Kadri F., Sun Y. (2020). Forecasting of photovoltaic solar power production using lstm approach, *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems 3*

Harrou F., Cheng T., Sun Y., Leiknes T., Ghaffour N. (2020). A data-driven soft sensor to forecast energy consumption in wastewater treatment plants: A case study, *IEEE Sens. J., 21* (4), pp. 4908-4917

Hoque J. M. Z., Ab Aziz N. A., Alelyani S., Mohana M., & Hosain, M. (2022). Improving Water Quality Index Prediction Using Regression Learning Models. *International journal of environmental research and public health*, *19*(20), 13702. https://doi.org/10.3390/ijerph192013702

Hoseinzade E., Haratizadeh S.(2019).CNNpred: CNN-based stock market prediction using a diverse set of variables,*Expert Systems with Applications*, *129*, Pages 273-285, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.03.029.

Harrou F., Dairi A., Dorbane A., Sun Y.(2023) Energy consumption prediction in water treatment plants using deep learning with data augmentation. *Results in Engineering*, *20*, 101428, ISSN 2590-1230. https://doi.org/10.1016/j.rineng.2023.101428.

Kazemi P., Bengoa C., Steyer J.-P. , Giralt J. (2021). Data-driven techniques for fault detection in anaerobic digestion process, *Process. Saf. Environ. Prot.*, *146* pp. 905-915

Karadimos P, Anthopoulos L.(2023). Machine Learning-Based Energy Consumption Estimation of Wastewater Treatment Plants in Greece. *Energies*. *16*(21):7408. https://doi.org/10.3390/en16217408

Longo S., d'Antoni B.M. , M. Bongards, A. Chaparro, A. Cronrath, F. Fatone, Lema J.M. , Mauricio M. I., Soares A. , Hospido A. (2016). Monitoring and diagnosis of energy consumption in wastewater treatment plants: A state of the art and proposals for improvement. *Appl. Energy*, *179*, 1251-1268

Luengo J., García-Gil D., Ramírez-Gallego S., García S., Herrera F.(2020). Big Data Preprocessing *Springer*

Lakshmanaprabu K. , Shankar K. , Ilayaraja M. , Nasir A.W. , Vijayakumar V. , Chilamkurti N. (2019). Random forest for big data classification in the internet of things using optimal features, *Int. J. Mach. Learn. Cybern., 10* , pp. 2609-2618

Melbourne Water. Annual report 2020-2021 Retrieved from https://www.melbournewater.com.au/about/what-we-do/annual-report accessed 31 March 2024.

Melbourne Airport Weather Station (2021) Retrieved from https://en.tutiempo.net

Melbourne Water Database (2021) Retrieved from https://data-melbournewater.opendata.arcgis.com

Natekin A., Knoll A. (2013). Gradient boosting machines, a tutorial, *Front. Neurorobot.* , p. 7

Office of the Chief Economist. (2016). Australian Energy Statistics. Retrieved from https://www.energy.gov.au/publications/australian-energy-update-2016, accessed 31 March 2024.

Office of Environment & Heritage. (n.d.). Energy efficiency opportunities in wastewater treatment facilities. Retrieved from https://www.environment.nsw.gov.au/resources/business/wastewater-treatment-facilities-energy-efficiency-opportunities-190114.pdf

Oulebsir R., Lefkir A., Safri A., Bermad A. (2020). Optimization of the energy consumption in activated sludge process using deep learning selective modeling, *Biomass and Bioenergy, Volume 132,* 105420, ISSN 0961-9534, https://doi.org/10.1016/j.biombioe.2019.105420.

Picos-Benítez A.R. , Martínez-Vargas B.L. ,Duron-Torres S.M. , Brillas E. , Peralta-Hernández J.M. (2020). The use of artificial intelligence models in the prediction of optimum operational conditions for the treatment of dye wastewaters with similar structural characteristics, *Process. Saf. Environ. Prot., 143* , pp. 36-44

Panepinto D., Fiore S., Zappone M., Genon G. , Meucci L. (2016). Evaluation of the energy efficiency of a large wastewater treatment plant in Italy.*Appl Energy, 161* , pp. 404-411.

Raut P., Dani A. (2020). Correlation between number of hidden layers and accuracy of artificial neural network, Springer, Singapore

Ranjan K.G., Prusty B.R., Jena D. (2021). Review of preprocessing methods for univariate volatile time-series in power system applications, *Electr. Power Syst. Res., 191* , Article 106885

Shaikh A.K., Nazir A., Khalique N., Shah A.S., Adhikari N. (2023). A new approach to seasonal energy consumption forecasting using temporal convolutional networks, *Results Eng., 19*, Article 101296

Smola, A.J., Schölkopf, B. (2004).A tutorial on support vector regression. *Stat. Comput. 14*, 199–222.

Tosun E. , Aydin K. , Bilgili M. (2016). Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures, *Alexandria Eng. J., 55* , pp. 3081-3089.

Wang S., Zou L., Li H., Zheng K., Wang Y., Zheng G., Li J.(2020). Full-scale membrane bioreactor process WWTPs in East Taihu basin: Wastewater characteristics, energy consumption and sustainability. *Science of the Total Environment*, *723*, https://doi.org/10.1016/j.scitotenv.2020.137983.

Yu P.S., Chen S.T., Chang, I.F.(2006). Support vector regression for real-time flood stage forecasting. *J. Hydrol. , 328*, 704–716.

Yu Y., Zou Z., Wang S. (2019) Statistical regression modeling for energy consumption in wastewater treatment, *Journal of Environmental Sciences,Volume 75*, Pages 201-208, ISSN 1001-0742, https://doi.org/10.1016/j.jes.2018.03.023.

Yang X., Wei J., Ye G., Zhao Y., Li Z., Qiu G., Li F., Wei C.(2020) The correlations among wastewater internal energy, energy consumption and energy recovery/production potentials in wastewater

treatment plant: An assessment of the energy balance, *Science of The Total Environment,Volume 714*,136655, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.136655.

Zhao L. , Dai T. , Qiao Z. , Sun P. , Hao v, Yang Y. (2020). Application of artificial intelligence to wastewater treatment: a bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse, *Process. Saf. Environ. Prot.*, *133*, pp. 169-182

Zhang, S., Wang, H., & Keller, A. A. (2021). Novel machine learning-based energy consumption model of wastewater treatment plants. *ACS ES&T Water, 1*(12), 2531-2540. https://doi.org/10.1021/acsestwater.1c00283

Zhou, Z. (2012). Ensemble Methods: Foundations and Algorithms; CRC Press: Boca Raton, FL, USA, pp. 15–55.

Zuo R, Xiong Y, Wang J, Carranza EJM (2019) Deep learning and its application in geochemical mapping. *Earth Sci Rev 192*:1–14. https://doi-org.ezproxy.lib.torontomu.ca/10.1016/j.earscirev.2019.02.023

Żyłka R., Dąbrowski W., Malinowski P., Karolinczak B. (2020). Modeling of Electric Energy Consumption during Dairy Wastewater Treatment Plant Operation. *Energies*.; *13*(15):3769. https://doi.org/10.3390/en13153769

Zhang Z., Zeng Y., Kusiak A.(2012).Minimizing pump energy in a wastewater processing plant, *Energy,Volume 47*, Issue 1,Pages 505-514, ISSN 0360-5442,https://doi.org/10.1016/j.energy.2012.08.048.