Engineering Applications of Artificial Intelligence 162 (2025) 112338

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research Paper

Toward sustainable wastewater treatment: Transformer ensembles and multitask learning for energy consumption and quality management

Hager Saleh [a,b,g,*], Sherif Mostafa [c], Shaker El-Sappagh [d,e], Abdulaziz AlMohimeed [f], Michael McCann [g], Saeed Hamood Alsamhi [b,h,i], Niall O'Brolchain [b], John G. Breslin [b], Marwa E. Saleh [j]

[a] Faculty of Computers and Artificial Intelligence, Hurghada University, Hurghada, Egypt
[b] Insight Research Ireland Centre for Data Analytics, School of Engineering, University of Galway, University Road, Galway, H91 TK33, Ireland
[c] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
[d] Faculty of Computer Science and Engineering, Galala University, Suez, 435611, Egypt
[e] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha, 13518, Egypt
[f] College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 13318, Saudi Arabia
[g] Department of Computing, Atlantic Technological University, Letterkenny, Donegal, Ireland
[h] Department of Computer Science and Engineering, College of Informatics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea
[i] Department of Electrical Engineering, Ibb University, Ib, Yemen
[j] Computer Science Department, Faculty of Computers and Information, Minia University, Minia, Egypt

ARTICLE INFO

Keywords:
Wastewater treatment plant
Energy consumption
Transformer model
Multitask learning
Voting ensemble model
Wastewater quality indicators
Quality management
Multitask Bidirectional Gated Recurrent Unit

ABSTRACT

Wastewater treatment plants (WWTPs) are among the most energy-intensive components of urban infrastructure and bear strict regulatory responsibilities for wastewater quality. These dual challenges, minimizing energy consumption and maintaining environmental compliance, are deeply interrelated and must be managed simultaneously to achieve sustainable plant operation. This study proposes a framework that comprises two customized components. The first component employs a voting ensemble model based on transformer architecture to predict energy consumption. It processes heterogeneous feature domains — including hydraulic, wastewater, and climatic variables — through parallel attention-driven streams. The outputs from these streams are then aggregated using a weighted voting mechanism to produce the final prediction. Second, a multitask Bidirectional Gated Recurrent Unit (Bi-GRU) forecasts wastewater quality indicators concurrently (ammonia, Biochemical Oxygen Demand (BOD), and Chemical Oxygen Demand (COD)), capturing shared temporal dependencies and reducing model complexity. A hybrid preprocessing strategy is applied, incorporating domain-aware outlier detection (z-score and Interquartile Range (IQR)), K-Nearest Neighbors (KNN) Imputation, and feature selection using Extreme Gradient Boosting (XGBoost).

Experimental results showed that. The voting ensemble model achieved the best results for energy consumption prediction with 31.61 of Root Mean Squared Error (RMSE). The multitask Bi-GRU achieved the best results for wastewater quality indicators with RMSE at 6.1689, 48.0323, and 88.2214 for ammonia, BOD, and COD, respectively. This work is among the first to integrate transformer ensembles and multitask learning in a unified WWTP forecasting system. Simultaneously addressing energy efficiency and water quality assurance, this offers a practical, scalable, and intelligent decision-support tool for sustainable wastewater management.

1. Introduction

Wastewater treatment plant facilities (WWTPs), a major source of operating expenses and greenhouse gas emissions in the industry, are under pressure to improve efficiency while reducing energy consumption as urbanization increases and environmental restrictions tighten. About 3% of the world's power consumption is attributable to WWTPs, predicted to increase as the population and industrial demand expand (Cardoso et al., 2021). Since energy is a significant cost contributor (Molinos-Senante et al., 2018), increases in energy

* Corresponding author at: Insight Research Ireland Centre for Data Analytics, School of Engineering, University of Galway, University Road, Galway, H91 TK33, Ireland.
E-mail address: hager.saleh@insight-centre.org (H. Saleh).

https://doi.org/10.1016/j.engappai.2025.112338
Received 16 March 2025; Received in revised form 11 August 2025; Accepted 12 September 2025
Available online 24 September 2025
0952-1976/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

consumption may result in a substantial increase in carbon dioxide emissions in the water sector when fossil fuels are utilized in power plants and a significant increase in operating costs (Hao et al., 2015). Because of the significant energy footprint, WWTPs are an important target for optimization initiatives meant to promote climate-resilient infrastructure and sustainable urban development. However, lowering energy use in WWTPs is fraught with challenges. Numerous variables, such as hydraulic load, pollution levels, and ambient environmental conditions, all of which change over time, affect the energy consumption in these facilities.

Energy management has always depended on linear models and rule-based methodologies, which frequently fall short of capturing the intricate relationships between these factors (Zhou et al., 2024). Consequently, the development of more sophisticated modeling methods that can interpret and forecast energy usage across varying operating scenarios is urgently required. The traditional analysis methods and nonlinear and interconnected linkages are difficult for traditional energy management techniques, which frequently rely on static or rule-based models, to capture, resulting in inefficiencies and lost energy-saving potential.

Promising solutions to overcome challenges in traditional analysis methods are provided by recent developments in deep learning (DL) and machine learning (ML). DL models like Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Transformers can efficiently analyze multivariate time-series data by utilizing algorithms that are very proficient in pattern recognition. The intelligent techniques allow them to capture both short-term and long-term dependencies in WWTP operations (Niu et al., 2023; Ramli and Hamid, 2018; Ahmed et al., 2023; Bagherzadeh et al., 2021a). WWTP operators can modify their operating methods to reduce energy consumption while preserving treatment efficiency when they have access to quick and precise energy consumption data. For instance, when influent loading is minimal or influent quality is generally good, they may optimize aeration levels, which lowers energy use (influent is untreated or partially treated wastewater). The volume of research examining DL in wastewater treatment has increased recently (Alvi et al., 2023; Baarimah et al., 2024). However, the limitations of ML and DL approaches include their inability to fully capture intricate, nonlinear interactions and dependencies among multivariate data, which sometimes necessitates considerable feature engineering or data augmentation. Furthermore, disappearing gradients and performance degradation over extended time sequences are problems that can plague models like RNNs, LSTMs, and GRUs, which reduces their ability to estimate energy usage properly in dynamic and multivariate settings like WWTPs. However, WWTPs must balance energy efficiency with treatment effectiveness. Energy use (especially for aeration, pumping, and heating) is tightly coupled with achieving effluent quality standards. Therefore, simultaneous prediction of energy consumption is essential for coordinated, intelligent control, enabling cost savings without compromising regulatory compliance.

This study proposes a unified predictive framework that supports the sustainable and cost-effective operation of wastewater treatment plants. By jointly forecasting energy consumption and critical effluent quality indicators, the framework enables data-driven decision making that balances energy efficiency with compliance with environmental regulations. Such dual forecasting is crucial to optimize aeration, pumping schedules, and treatment cycles in real time, promoting both operational efficiency and environmental sustainability. We propose an integrated prediction framework that combines a multi-task deep learning model with a voting ensemble of transformer models. By combining forecasts from many data categories — hydraulic, wastewater, and climate — the transformer-based voting ensemble increases the precision and robustness of energy consumption forecasting. Simultaneously, the multi-task model leverages the shared common features across domains to forecast ammonia, COD, and BOD concurrently. This paper not only enhances the prediction performance but also empowers resource management in real time and contributes to sustainability and energy efficiency.

1.1. Motivations and contributions

WWTPs are under increasing pressure to optimize energy consumption due to rising energy costs, increasing urbanization, and tightening of environmental regulations. The increasing energy needs of urbanization and strict environmental restrictions highlight the pressing need for creative energy management solutions in WWTPs. Most previous research in the domain of wastewater treatment and energy prediction has taken a single-task, single-model approach, often relying on conventional machine learning models (e.g., GBM, RF, SVR) or basic deep learning architectures (e.g., LSTM, GRU) applied in isolation. These approaches typically target either energy consumption or a single wastewater quality indicator and often lack integration across multiple feature domains (hydraulic, climatic, and wastewater). These traditional energy management systems often fail to capture complex and non-linear interactions between dynamic parameters such as hydraulic load, wastewater composition, and climate, even with advances in DL. These restrictions result in inefficiencies, lost optimization, and increased greenhouse gas emissions.

This study bridges that gap by introducing a unified deep learning framework that addresses both energy consumption forecasting and wastewater quality prediction in parallel, using an architecture that is, to our knowledge, novel in the literature. Our first contribution is a transformer-based voting ensemble model that independently processes features from hydraulic, climate, and wastewater categories using separate transformer blocks. It then fuses their outputs through a weighted voting mechanism. Unlike traditional monolithic or early-fusion approaches, this architecture enables specialized learning in each domain while maintaining robustness through ensemble integration. Our second contribution is a multi-task Bi-GRU model that predicts ammonia, BOD, and COD concurrently using shared temporal representations. This enables knowledge transfer between related prediction tasks, improving data efficiency and model generalization. Additionally, we introduce a dual-strategy preprocessing pipeline that handles outliers and missing values using statistically appropriate methods tailored to the distributional characteristics of each feature. We also apply Bayesian optimization for hyperparameter tuning, ensuring high model performance and efficiency. Overall, this study is the first to jointly leverage transformer-based ensemble modeling and multitask recurrent networks for comprehensive energy and quality management in WWTPs, thus offering a generalizable, scalable, and high-performing framework for sustainable infrastructure operations. The contributions of this paper are summarized as follows:

- *Transformer-based ensemble model for domain-aware energy prediction:* We propose a voting ensemble deep learning model that uniquely applies multiple Transformers models for energy consumption forecasting in WWTPs. The proposed voting ensemble model improves performance and resilience by integrating outputs from diverse transformer models based on different data categories (i.e., hydraulic, wastewater, and climatic) in contrast to single-model techniques.
- *Multi-task Bi-GRU model for joint quality forecasting:* We propose a novel multi-task deep learning model for concurrently predicting numerous vital wastewater quality indicators (i.e., ammonia, BOD, and COD). Bi-GRU has improved temporal and sequential data learning for leveraging shared features across domains. We use optimal feature selection by using the XGBoost regressor to guarantee the model is computationally efficient while maintaining significant prediction accuracy performance.
- *Hybrid preprocessing and feature selection pipeline:* We apply a tailored and efficient data preprocessing strategy — z-score for normally distributed features, IQR + KNN for skewed features — and XGBoost-based feature selection. In particular, models trained on selected features outperformed full-feature models, confirming the value of compact, high-quality input.

- *Empirical validation on real-world WWTP data:* Our models achieve competitive results: an RMSE of 31.61 for energy prediction and 6.1689 for ammonia, 48.0323 for BOD, and 88.2214 for COD, respectively. These results highlight the effectiveness of the proposed integrated framework for sustainable WWTP operation.

In summary, while prior studies have focused on either energy consumption or effluent quality separately, WWTPs require integrated monitoring and forecasting of both aspects to meet both their economic and sustainability goals. Operational decisions (e.g., aeration control) directly influence both energy use and treatment effectiveness. Hence, this paper introduces a unified deep learning framework that supports holistic decision support by simultaneously forecasting energy needs and water quality indicators. The goal is not just improved accuracy but a practical, dual-purpose tool to guide sustainable WWTP operation. The proposed approach is innovative in that it explicitly models the operational complexity of wastewater treatment plants by combining transformer-based domain-specific learners and a multi-task temporal forecaster. Our model aligns with real-world WWTPs' goals by jointly forecasting interdependent targets while adapting to the structure of the input domains (hydraulic, climatic, and wastewater).

1.2. Paper structure

The remainder of this paper is arranged as follows. Section 2 provides overview of previous studies related to this work. The proposed framework describes the proposed model of energy consumption and wastewater quality factors introduced in Section 3. The experimental results are discussed in Section 4. Section 5 discusses the summary of results. Finally, conclusions are shown in Section 6.

2. Related work

Recent research has shown notable progress in using a variety of ML and DL approaches to optimize energy use in WWTPs. Bagherzadeh et al. (2021b) used feature selection and models such as artificial neural networks (ANN), Gradient Boosting Machine (GBM), and Random Forest (RF) to examine the impact of climatic, hydrological, and wastewater characteristics on energy usage. Models trained and evaluated using public and open datasets (thanks to the Victoria Government's open data policy) were collected from the Melbourne water facility for six years from 2014 to 2019. The results showed that GBM recorded the best performance; their investigation revealed that GBM was the best-performing model. Using the same dataset, Alali et al. (2023) made comparisons between 24 ML models that were optimized by Bayesian optimization for energy consumption prediction. Ensemble models using RF and XGboost were applied to select the most relevant features. In addition, they proposed lagged measurements as inputs to enhance the ML models' ability to improve the model's performance. Harrou et al. (2023) applied different DL models, RNN, LSTM, GRU, BiLSTM, and BiGRU, with feature selection (FS) and data augmentation techniques for predicting the WWTP's energy consumption. They applied a cubic spline as a data augmentation method to increase the dataset size and enhance the model's performance. They made comparisons between models before and after applying data augmentation. The results showed that BiGRU recorded the best performance with data augmentation.

Additional studies by Zhang et al. (2021) used RF to build energy consumption models using urban drainage data from the China Statistical Yearbook, emphasizing the importance of discharge standards in energy predictions. Ramli and Hamid (2018) presented data-based modeling for a wastewater treatment facility employing ML techniques. The authors examined methods such as ANN, KNN, Support Vector Regression (SVR), and LR to forecast energy use. Data used for energy consumption was gathered from Tenaga Nasional Berhad (TNB) electrical bills in Malaysia between March 2011 and February 2015. According to the study, ANNs outperformed the other machine learning techniques regarding prediction accuracy, having the lowest root mean square error. Torregrossa et al. (2016) have highlighted the benefits of data-driven approaches in achieving energy efficiency by emphasizing the effective use of ANNs, support vector regression (SVR), and linear regression (LR) to forecast energy usage. Bagherzadeh et al. (2021b) proposed studies of the effect of wastewater, hydraulic, and climate-based parameters on energy consumption (EC) and sustainable energy-saving WWTPs using feature selections (FS), ML, and ANN, and Gradient Boosting Machine (GBM), and Random Forest (RF). Furthermore, many creative approaches and procedures have been developed to improve energy consumption modeling in WWTPs. While Oliveira et al. (2021) produced remarkable results with a CNN model for energy predictions, Das et al. (2021) examined several DL models and found GRUs to be superior in predicting energy usage based on real-world data. Yusuf et al. (2019) demonstrated the efficacy of LSTM in enhancing prediction accuracy by combining ARIMA and LSTM models to forecast electric load. By combining RF and neural networks to improve forecast accuracy, Torregrossa et al. (2018) presented a machine learning cost modeling technique that successfully evaluated energy costs across 317 WWTPs in Europe.

The fuzzy clustering approach by Qiao and Zhou (2018) enhanced fuzzy neural networks' effluent quality and energy consumption modeling capabilities. Last, Oulebsir et al. (2020) promoted energy efficiency and sustainability in wastewater treatment processes by optimizing energy usage at the Boumerdes-WWTP using a data-driven approach. However, our proposed framework bridges a significant gap by offering a comprehensive strategy that integrates various datasets and cutting-edge modeling techniques, opening the door for more effective and sustainable wastewater treatment processes than previous studies, which mainly concentrated on individual modeling techniques or limited data sources.

Mekaoussi et al. (2023) proposed a hybrid ML model based on an extreme learning machine (ELM) optimized by the Bat algorithm (ELM-Bat) for predicting BOD in WWTP. The results showed that the hybrid ELM-Bat recorded the best performance compared with those of the multilayer perceptron(MLP), the RFR, and Gaussian process regression (GPR). In Baki et al. (2019), the authors used a dataset collected from 2025 to 2026 from the laboratory of Antalya Hurmathat to predict BOD. The dataset includes different features: pH, chemical oxygen demand, suspended sediment, total nitrogen, total phosphorus, electrical conductivity, and input discharge. In Alsulaili and Refaie (2021), the authors applied an ANN model with a minimal set of influent variables (temperature, conductivity, COD) to predict BOD. With R^2 values in the 0.61–0.75 range, these models show promise for application as soft sensors in WWTP real-time control systems, speeding up decision-making and optimization. In Saleh and Kayi (2021), the authors applied an ANN with different features: chloride ions, nitrate ions, phosphate ions, sulfate ions, ammonia, and BOD, which are included in the dataset collected from WWTP North Gas Company/Kirkuk, to predict COD. Abba and Elkiran (2017) applied an ANN and MLP with different features: pH, TSS, total nitrogen, total phosphates, conductivity, and SS, which are included in the dataset collected from WWTP models, to predict the COD in a wastewater treatment plant in Nicosia, North Cyprus. In Heddam et al. (2016), the authors proposed a generalized regression neural network (GRNN) with five variables: COD, pH, temperature, suspended solids, and electrical conductivity, which were collected from WWTPs in the east of Algeria, to predict BOD. In Qambar and Al Khalidy (2022), the authors applied different ML models: DT, RF, adaptive boosting, gradient boosting (GB), and extreme gradient boosting algorithms, using two WWTPs in the South Kingdom of Bahrain to predict BOD. The GB model obtained the best results.

To handle the previous challenges, we propose a comprehensive and purposefully integrated DL framework by combining transformer-based voting ensembles for energy prediction and multi-task temporal models

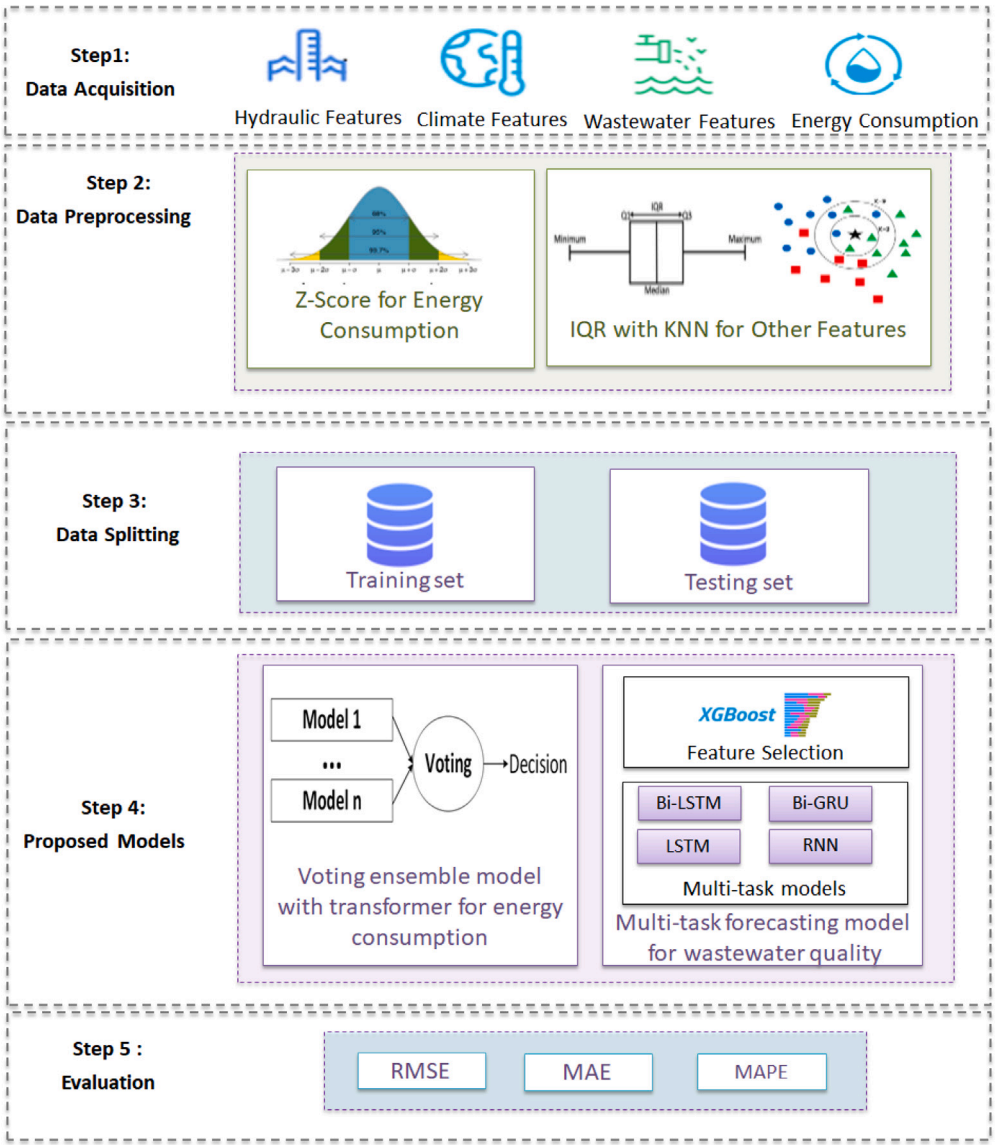

Fig. 1. Proposed architecture for energy consumption and wastewater quality.

for effluent quality forecasting. Our approach mirrors the operational realities of WWTPs and leverages shared patterns in climate, hydraulic, and wastewater data. This integration is not incidental, but central to achieving accurate, robust, and interpretable forecasting that can be realistically adopted in practice.

3. Methodology

Fig. 1 presents the main components of the proposed methodology, which comprises five key stages: (1) **Data acquisition**, incorporating hydraulic, climate, wastewater features, and energy consumption data; (2) **Data preprocessing**, where outliers in energy consumption are handled using the z-score method, while outliers in the other features are treated with the IQR method combined with KNN imputation; (3) **Data splitting**, dividing the dataset into training and testing subsets; (4) **Proposed models**, including a voting ensemble Transformer-based model for forecasting energy consumption, and a multi-task Transformer model for predicting wastewater quality indicators such as BOD, COD, and ammonia; and (5) **Evaluation**, performed using RMSE, MAE, and MAPE as performance metrics. Each of these steps is described in detail in the following subsections.

3.1. Data acquisition

Multivariate data from the Melbourne water treatment plant and airport weather station[1] is used to enhance methods for predicting energy consumption. The dataset covers five years, from January 2014 to June 2019, and is comprised of 1382 entries that includes nineteen variables, as outlined in Table 1. The variables cover essential elements, including energy consumption, biological, hydraulic factors, and climatic. Information regarding water quality and biological traits was acquired through sensor readings, while weather data was obtained from the Melbourne airport weather station. The dataset encompasses time-related information, which is utilized to enhance prediction accuracy. Further information about the used dataset is given in Bagherzadeh et al. (2021b).

3.2. Data preprocessing

To guarantee the reliability of the data, a meticulous data-cleaning procedure is conducted through the following steps:

[1] https://data.mendeley.com/datasets/pprkvz3vbd/1.

Table 1
Features and corresponding name of dataset.

Categories	Parameters (Abbreviation)	Unit
–	Energy consumption	MWh
Hydraulic	Average inflow	m^3/s
	Average outflow	m^3/s
Wastewater	Ammonia (NH_4–N)	
	Biological Oxygen Demand (BOD)	mg/L
	Chemical Oxygen Demand (COD)	mg/L
	Total Nitrogen (TN)	mg/L
Climate	Average temperature	°C
	Maximum temperature	°C
	Minimum temperature	°C
	Atmospheric pressure	hPa
	Average humidity	
	Total rainfall (Pr)	
	Average visibility	km
	Average wind speed	km/h
	Maximum wind speed	km/h
Time	Year	–
	Month	–
	Day	–

- Outliers in energy consumption — data points exhibiting unusually high or low usage — were removed using the z-score method. This choice was grounded in empirical distribution analysis, which indicated that the energy consumption followed a near-normal distribution. The z-score method, being sensitive to deviations in Gaussian distributions, was thus appropriate for this context (Chikodili et al., 2020).
  We applied the z-score method using a standard threshold of ±1.96. Data points with z-scores greater than +1.96 or less than −1.96 were considered statistical outliers and were excluded from further analysis. This threshold corresponds to the conventional 95% confidence interval for normally distributed data. The decision was supported by visual inspection, which confirmed that energy consumption followed a near-Gaussian distribution.
  Approximately 5.43% of the data points were removed through this step. The z-score indicates how many standard deviations a data point is from the mean of the dataset, allowing us to systematically detect and eliminate outliers.
  To compute a z-score for a specific data point, Eq. (1) is used.

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

  where z is the z-score, x is the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset.
- In contrast, the distributions of other features (hydraulic, wastewater, and climatic variables) were found to be skewed or non-Gaussian, as confirmed by descriptive statistics and visual inspections. Consequently, we employed the Interquartile Range (IQR) method for detecting outliers in these variables, given its robustness to distributional assumptions (Vinutha et al., 2018), a statistical technique that measures data dispersion. Detected outliers were subsequently imputed using the K-Nearest Neighbors (KNN) algorithm (Peterson, 2009). This approach was particularly suitable as the majority of features exhibited skewed or non-Gaussian distributions, for which the IQR method is known to be more robust and effective compared to other outlier detection techniques.
  To determine the IQR for a provided dataset, Eq. (2) is applied after arranging the data in ascending order.

$$IQR = Q3 - Q1, \tag{2}$$

  where Q1 represents the median of the lower half of the dataset, while Q3 represents the median of the upper half of the dataset. An outlier is defined as any value that falls outside the range:

$$[Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR]$$

Following outlier detection, once outliers are detected across all features. The features are treated as missing values and handled using the KNN imputation procedure, which involves calculating distances and imputing missing values based on the nearest neighbors as follows: The Euclidean distance between two data points a and b in an n dimensional space is calculated by using Eq. (3).

$$Distance\,(a, b) = \sqrt{\sum_{i=1}^{n} (a - b)} \tag{3}$$

Then, the nearest neighbors K is used to determine the data point with missing values based on the calculated distances. The value of K = 3 was empirically chosen after conducting a grid search across values from 1 to 10. The optimal K was determined based on imputation accuracy metrics (RMSE, MAE) on a validation subset, and K = 3 achieved a balance between local sensitivity and computational stability, minimizing potential overfitting.
For imputing the missing value m (in feature f_m) of a data point, the imputed value is calculated as the weighted average of the corresponding feature values from the K nearest neighbors, utilizing Eq. (4).

$$\hat{m} = \frac{\sum_{i=1}^{K} w_i \times m_i}{\sum_{i=1}^{K} w_i}, \tag{4}$$

where w represents the weighted average, which can be computed through different weighting methods including inverse distance weighting as shown in Eq. (5).

$$w_i = \frac{1}{\text{Distance}(a, b_i)} \tag{5}$$

3.3. Data splitting

In this step, the dataset has been split into two subsets: a training set comprising 80% of the data, and a testing set comprising the remaining 20%. The training set has been used to train the proposed models, while the testing set (unseen data) has been employed to evaluate their performance.

3.4. Proposed models

3.4.1. Voting ensemble model for energy consumption

A novel voting ensemble model has been developed as an integrated predictive tool to enhance performance and data analysis capabilities significantly. The integrated model combines three distinct feature categories (i.e., hydraulic, wastewater, and climate) allowing for a comprehensive approach to improving predictive accuracy, as depicted in Fig. 2. To effectively capture the nuances of each feature set, the model utilizes parallel transformer architectures, each dedicated to one of the three feature categories. This parallelization takes advantage of the unique strengths and specialized capabilities of each model type, ensuring that the specific characteristics of each feature set are fully leveraged.

After generating predictions from the individual models, a weighted average voting method is used to combine them. In this approach, each model's prediction is assigned a weight that reflects its relative importance or reliability. These weights were determined based on the models' individual performance on the validation dataset. Specifically, we evaluated each model using performance metrics such as RMSE and MAE, and assigned higher weights to models that demonstrated better predictive accuracy. Therefore, these weights were manually selected through iterative experimentation to balance performance across different input sources, rather than being learned during training.

The final output is then obtained by multiplying each model's prediction by its corresponding weight and summing the results. This ensemble approach enhances the robustness and accuracy of the final predictions by reducing the influence of less reliable models. A step-by-step outline of this method is provided in Algorithm 1.

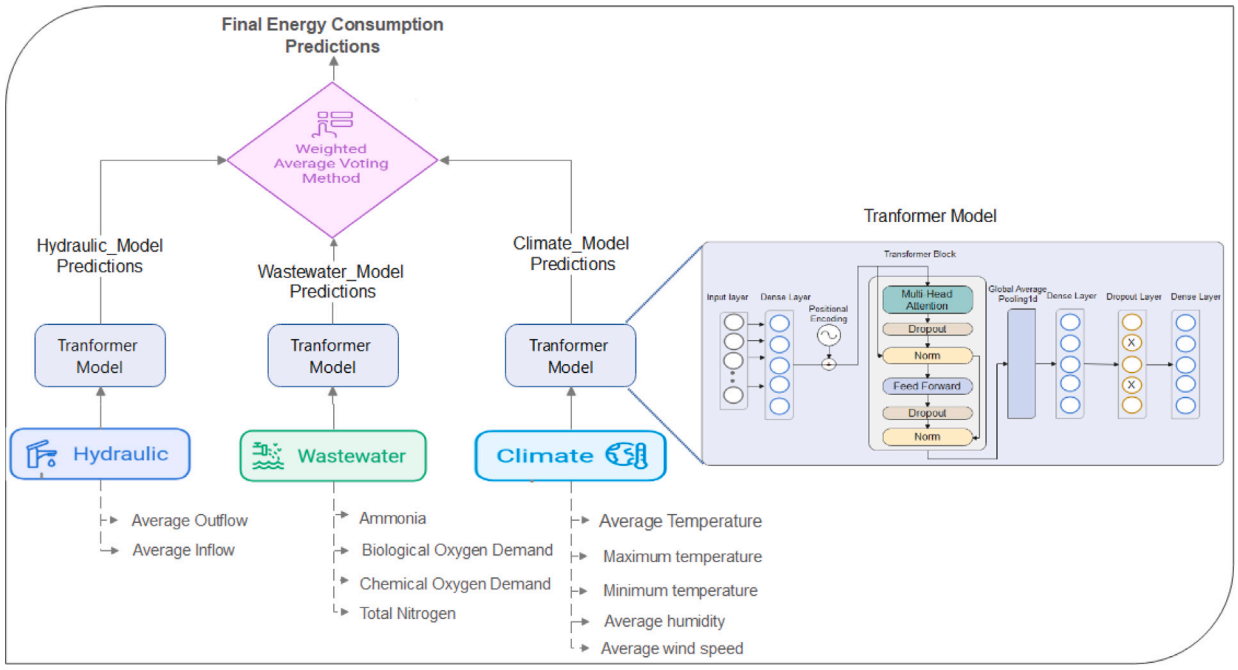

Fig. 2. Voting ensemble model with transformer model for energy consumption.

Algorithm 1 Weighted Average Voting for Multi-Source Prediction

Require: Hydraulic_features, Wastewater_features, Climate_features
Ensure: Final_predictions
1: $P_{\text{hyd}} \leftarrow$ Hydraulic_Model(Hydraulic_features)
2: $P_{\text{ww}} \leftarrow$ Wastewater_Model(Wastewater_features)
3: $P_{\text{cl}} \leftarrow$ Climate_Model(Climate_features)
4: $w_{\text{hyd}} \leftarrow 0.2$
5: $w_{\text{ww}} \leftarrow 0.5$
6: $w_{\text{cl}} \leftarrow 0.4$
7: Final_predictions $\leftarrow w_{\text{hyd}} \cdot P_{\text{hyd}} + w_{\text{ww}} \cdot P_{\text{ww}} + w_{\text{cl}} \cdot P_{\text{cl}}$
8: **return** Final_predictions

The transformer model employed in the voting ensemble model, as depicted in Fig. 2, begins with an input layer that receives a tensor X of shape (1, number of features) at each time step. This input is then passed through a dense layer, which projects the features into an embedding space, as defined in Eq. (6).

$$E = W_E X + b_E, \tag{6}$$

where E represents the embedding output, X is the input, W_E is the weight matrix, and b_E is the bias vector. Then, Positional Encoding (PE) is added to the embedding E in (7).

$$E_{PE} = E + PE \tag{7}$$

where E_{PE} represents what is then processed through the transformer block, which incorporates a Multi-Head Attention mechanism and a Feed-Forward Neural Network (FFNN) with normalization and dropout layers, as illustrated in Fig. 2. The advantage of the Multi-Head Attention mechanism is its ability to capture complex dependencies between different parts of the input by concentrating on various "head" or aspects of the data (Lv et al., 2022). Therefore, E_{PE} goes through a Multi-Head Attention mechanism as follows:

$$attn\_output = MultiHeadAttention(Q, K, V), \tag{8}$$

where $Q, K, and V$ are the Query, Key, and Value matrices derived from E_{PE} since this is a self-attention mechanism (Yang et al., 2024). The Multi-Head Attention output for each head is computed as follows:

$$Attention(Q, K, V) = softmax(d\left(\frac{QK^T}{\sqrt{d_k}}\right))V, \tag{9}$$

where d_k is the dimensionality of the keys used for scaling. The final Multi-Head Attention output is obtained by concatenating the outputs of all heads and projecting them using a weight matrix, as in the following equation.

$$attn\_output = Concat\left(head_1, \ldots, head_h\right) W_O, \tag{10}$$

where W_O is the output projection matrix. After that, dropout is applied to the attention output, and a residual connection adds the original input to the attention output, using Eqs. (11) and (12). This step helps stabilize the training by maintaining gradient flow.

$$attn\_output = Dropout\left(attn\_output, p = 0.1\right), \tag{11}$$

where p is the dropout probability. The output from the attention block, out1, is given by:

$$out1 = LayerNorm\left(E_{PE} + attn\_output\right) \tag{12}$$

It is processed through a Feed-Forward Network consisting of two Dense layers, as described in Eq. (13).

$$ffn\_output = ReLU\left(W_1 * out1 + b_1\right) W_2 + b_2, \tag{13}$$

where W_1 and b_1 are the weights and biases for the first dense layer, W_2 and b_2 are the weights and biases for the second dense layer, and ReLU is the activation function applied to the first dense layer. Subsequently, dropout is applied to the FFNN output, and a residual connection adds $out1$ to the FFNN output, as shown in Eqs. (14) and (15).

$$ffn\_output = Dropout\left(ffn\_output, p = 0.1\right), \tag{14}$$

where p is the dropout probability. Next, the output from the transformer block is given by:

$$trans\_out = LayerNorm\left(out1 + ffn\_output\right) \tag{15}$$

It passes through the Global Average Pooling1D layer, which reduces the data's dimensions by averaging each feature across all time steps, using Eq. (16):

$$max\_out = max(trans\_out) \tag{16}$$

Subsequently, a fully connected layer is applied, consisting of a dense layer with a ReLU activation function (giving y), followed by a dropout layer to mitigate overfitting (giving y_{drop}). Finally, a dense layer produces a single output value ($\tilde{y}$), making it suitable for regression tasks. Since this is a regression model, the output layer does not use an activation function, as demonstrated by the following equations:

$$y = ReLU(W_{fc1} * max\_out + b_{fc1}), \tag{17}$$

where W_{fc1} is a weight matrix of a fully connected layer and b_{fc1} is the bias vector.

$$y_{drop} = Dropout(y, p = 0.1), \tag{18}$$

where p is the dropout probability (0.1 in this case).

$$\tilde{y} = W_{out} * y_{drop} + b_{out}, \tag{19}$$

where W_{out} and b_{out} are a weight matrix and the bias vector of the output layer, respectively.

3.4.2. Multi-task model for wastewater quality

The second proposed model is designed to analyze the influence of multiple features on wastewater characteristics. It is structured as a multitask forecasting model, where the same set of input features is shared across multiple output targets, enabling the model to make predictions on various wastewater levels simultaneously. The model takes energy consumption, hydraulic data, climate variables, and time-based features (such as the day and month) as inputs. The shared features are then used to predict three different wastewater characteristics: ammonia, BOD, and COD. To identify which input features are most influential, the model employs the XGBRegressor technique, which is applied across all input features.

The architecture of the proposed multitask model is illustrated in Fig. 3. The architecture is comprised of a shared input layer and three distinct Bi-GRU sub-networks, one for each target output (ammonia, BOD, and COD). Each of these sub-networks produces a separate regression output, effectively forecasting the values for each wastewater characteristic. The first layer in the model is a shared input layer that expects a tensor with dimensions (1, number of features) for a single time step. Let X denote this input tensor. The next component of the model consists of three independent sub-networks, one for each target. Each sub-network includes a Bi-GRU layer, which processes the input data in both forward and backward directions.

To prevent overfitting and improve generalization, a dropout layer with a rate of 0.2 is applied after the Bi-GRU layer. This is followed by a dense layer with ReLU activation, which introduces non-linearity and allows the model to capture more complex patterns in the data. Another dropout layer is added to further regularize the network. Finally, the model ends with a dense output layer for each sub-network, which produces a single regression value for each target output (ammonia, BOD, or COD). These outputs are generated using a linear activation function to ensure that the predicted values are continuous and can represent the range of possible wastewater characteristics, as shown in the following equations.

$$h_t = \mathrm{GRU}(X_t, h_{t-1}) \tag{20}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(X_t, \overleftarrow{h_{t-1}}) \tag{21}$$

$$h_t^{\mathrm{BiGRU}} = \left[h_t, \overleftarrow{h_t}\right] \tag{22}$$

$$h_{drop} = Dropout\left(h_t^{\mathrm{BiGRU}}, 0.2\right) \tag{23}$$

$$d = ReLU\left(h_{drop} * W_1 + b_1\right), \tag{24}$$

where W_1 is a weight matrix of a dense layer and b_1 is the bias vector.

$$d_{drop} = Dropout\left(d, 0.2\right) \tag{25}$$

$$y_i = d_{drop} * W_{out} + b_{out}, \tag{26}$$

where W_{out} and b_{out} are the weight matrix and the bias vector of the output layer. This equation produces a single scalar output y_i for each target. The model's design allows it to learn shared patterns across multiple outputs while also accounting for the unique characteristics of each target. By leveraging the Bi-GRU layers and regularization techniques such as dropout, the model aims to make accurate and robust predictions on wastewater characteristics, even in the presence of complex temporal dependencies and potentially noisy data.

In summary, the choice to model each feature category (hydraulic, climatic, wastewater) with a separate transformer stream stems from operational insights in WWTPs. Each domain affects energy and effluent behavior differently, e.g., rainfall impacts inflow dynamics, temperature affects biological processing, and influent COD levels determine aeration load. This structure allows the model to specialize in the temporal and contextual patterns of each domain before fusing insights through ensemble voting.

3.5. Baseline deep learning models

We compared the proposed models with several DL models: Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bidirectional (Bi-LSTM), Bi-GRU (Bi-GRU).

- **RNN:** RNN is a form of neural network that interprets data in succession, keeping a hidden state updated throughout each time step, permitting it to record temporal dependencies (Sherstinsky, 2020). The key procedure integrates the current input with the prior hidden state, resulting in the current hidden state (Ming et al., 2017). Hidden state h_t updated each time t as:

  $$h_t = f(w_h h_{t-1} + w_x x_t + b_h) \tag{27}$$

  where w_h is the weight matrix for the hidden state, w_x is the weight matrix for the input, x_t is the input at time t, b_h is the bias term, and f is the activation function used. The output can be calculated as:

  $$y_t = w_y h_t + b_y \tag{28}$$

  where w_y is the weight matrix, and b_y is the bias vector for the output. Throughout training, gradients are generated using the chain rule, considering the dependencies within time steps.
- **GRU:** GRU is a variant of RNN that uses gating techniques to boost efficiency on sequence tasks, specifically for tackling the vanishing gradient problem (Shewalkar et al., 2019). GRUs are intended to record long-term dependencies in sequential data with greater efficiency. z_t is the update gate, r_t is the reset gate, and $\odot$ represents the element-wise multiplication. $h_{(t-1)}$ is the hidden state at previous time stamp $t-1$, x_t input at time t, w_z, w_r, w_h are the weight matrices for the update, reset, and hidden state, respectively, b_z, b_r, b_h bias vectors, and σ is the sigmoid activation function.

  $$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{29}$$

  $$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{30}$$

  $$\bar{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{31}$$

  $$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \bar{h}_t \tag{32}$$
- **LSTM:** LSTM is a powerful variant of RNN architecture that employs a cell state, and three gates, i.e., the input gate, governs the cell's input, the forget gate decides what information to give up, and the output gate regulates the cell's output (Yu et al.,

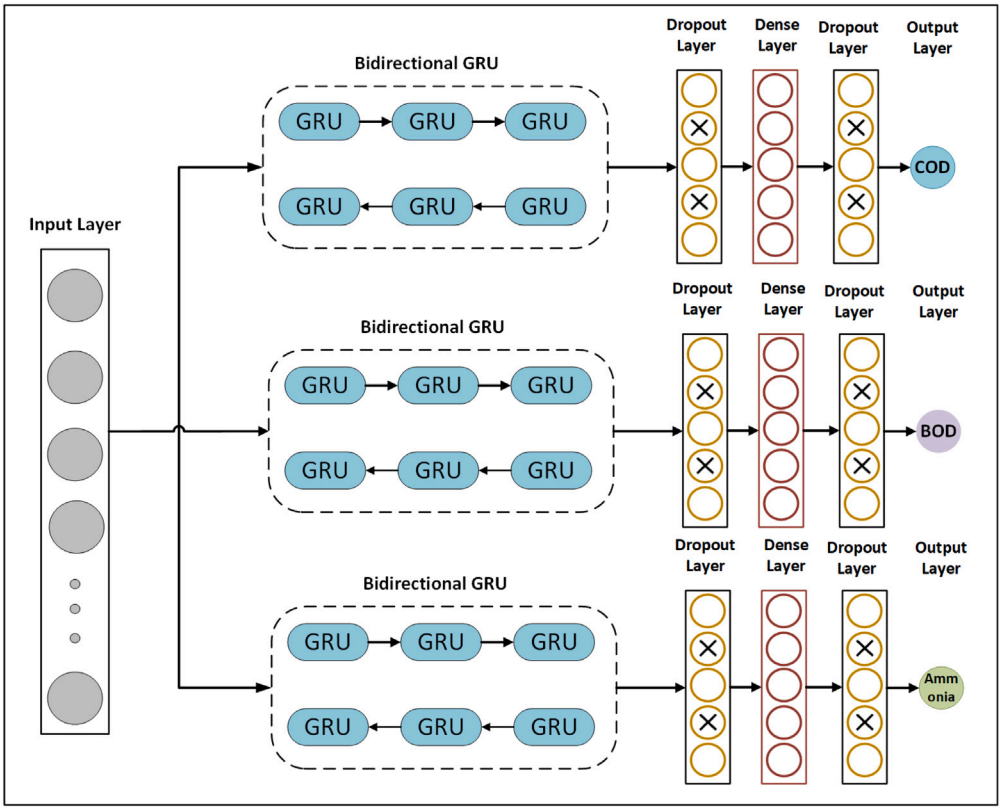

Fig. 3. Proposed multitask forecasting model for wastewater quality.

2019). The gating mechanisms enable LSTMs to retain and forget information selectively, making them an appealing option in various machine learning and deep learning applications (Belletti et al., 2019). where σ is sigmoid activation function, C_i is cell state at time t, $h_{(t-1)}$ is the hidden state at previous time stamp t−1, x_t is the input at time t, w_f, w_i, w_c, w_o are the weight matrices for the forget, input, candidate cell state and output respectively, $\odot$ represents the element-wise multiplication., and b_f, b_i, b_c, b_o are bias vectors.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{33}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{34}$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{35}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{36}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{37}$$

$$h_t = o_t \odot \tanh(C_t) \tag{38}$$

- **Bi-LSTM:** Bi-LSTM is a type of LSTM that analyzes input sequences in both forward and reverse directions (Siami-Namini et al., 2019). Bi-LSTM comprises two independent LSTM layers: forward LSTM, which handles the sequence from beginning to end. Backward LSTM evaluates the sequence from end to beginning (Siami-Namini et al., 2019). The outputs of both LSTM layers concatenate using summation, resulting in an expanded version of the input sequence.

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{39}$$

$$\overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t-1}}) \tag{40}$$

$$h_t^{\text{BiLSTM}} = \left[h_t, \overleftarrow{h_t}\right] \tag{41}$$

where (h_t) is the hidden state of the backward LSTM, this dual approach allows them to capture richer contextual information, making them highly effective for various applications in sequential data tasks.

- **Bi-GRU:** Bi-GRU leverages GRUs to process sequences in both directions. It integrates the hidden states of the forward and backward GRU, permitting it to access context from both ends of the sequence, improving the performance on various sequence-based tasks (Abdelgwad et al., 2022). The outputs of both GRU layers can be concatenated, summed, or merged to offer a more complete representation of the input sequence.

3.6. Performance evaluating

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used to evaluate models.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\text{obs}} - y_i^{\text{pred}}\right)^2} \tag{42}$$

$$\mathbf{MAE} = \frac{1}{\mathbf{n}}\sum_{\mathbf{i}=1}^{\mathbf{n}}\left|\mathbf{y_i^{obs}} - \mathbf{y_i^{pred}}\right| \tag{43}$$

The Mean Absolute Percentage Error (MAPE) is a common metric for evaluating forecasting accuracy. The formula for MAPE is:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{44}$$

4. Results and discussion

The results include two main sections that present the results of models for predicting energy consumption and the results of predicting wastewater targets (ammonia, COD, and BOD) using multitask models.

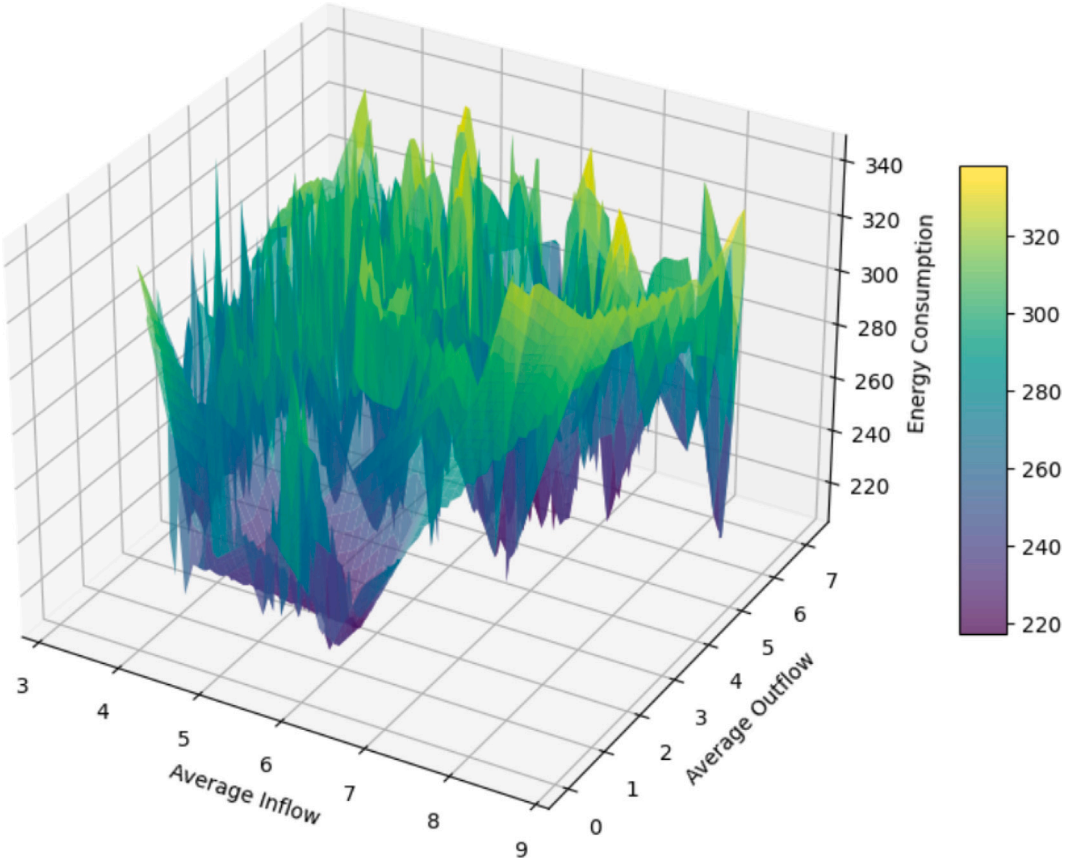

Fig. 4. 3D surface of average inflow, average outflow, and energy consumption.

4.1. Experimental setup

A NVIDIA RTX 3090 GPU was utilized to run our experiments. The programming language implemented for this experimentation is Python 3, which is used for implementations based on different libraries. Seaborn and Matplotlib are used to visualize results, and TensorFlow and Keras are used for the DL models. We removed atmospheric pressure from the climate category because it has the same value of 0. Different data-cleaning processes are implemented: Outliers in energy consumption are removed using the z-score method. In contrast, outliers across all features are identified using IQR and replaced by applying KNN.

4.1.1. Dataset splitting

The dataset is split into a 75% training set and a 25% testing set. A training set was used to train models, and a testing set was used to evaluate models.

4.1.2. Hyperparameter values

Hyperparameter values for transformer models, Each transformer model within the voting ensemble (hydraulic, wastewater, and climate branches) follows a consistent architecture composed of four Multi-Head Attention. The input features from each domain are embedded into a vector of size 64. The feed-forward network following each attention layer uses a 128-dimensional hidden layer and a ReLU activation, with a dropout rate of 0.1 applied after both the attention and FFN blocks. The models were trained using the Adam optimizer with a learning rate of 0.001. The batch size was set to 16, and training was conducted for 50 epochs, with early stopping triggered if the validation RMSE did not improve for 10 consecutive epochs. The same hyperparameter configuration was used for all three domain-specific transformers to ensure consistency and fair ensemble integration.

Hyperparameter values for DL models: LSTM, GRU, Bi-LSTM, Bi-GRU, RNN, The models were trained using the Adam optimizer with a learning rate of 0.001, 50 epochs, with early stopping triggered if the validation RMSE did not improve for 10 consecutive epochs. Models include 400 hidden units with Relu as activation functions, with a dropout layer of 0.3.

4.2. Data analysis

This section presents the effect of each category on energy consumption and wastewater.

4.2.1. Data analysis of energy consumption

The correlation between three variables, i.e., average inflow, average outflow, and the energy consumption is depicted in a three-dimensional surface graph, an in Fig. 4. The x-axis shows average inflow, the y-axis shows average outflow, and the z-axis shows energy consumption. Variations in energy consumption are displayed on the surface graph according to average inflow and outflow levels. The complex and sweeping surface of the graph, which has several peaks and valleys, indicates that energy consumption was significantly impacted by the interplay between average input and average outflow. The surface's colors correspond to energy consumption; warmer hues (red and yellow) indicate higher consumption, while bluer and purpler hues indicate lower consumption. Pumping and treatment operations usually use more energy at higher average inflow rates. Pumps must work harder as more water reaches the plant, which boosts energy usage. Additionally, sustaining pressure and flow rates often requires more energy when outflow is higher. Therefore, optimizing energy efficiency and operational costs requires effectively controlling these features. From the graph, it is clear that there is a direct relationship

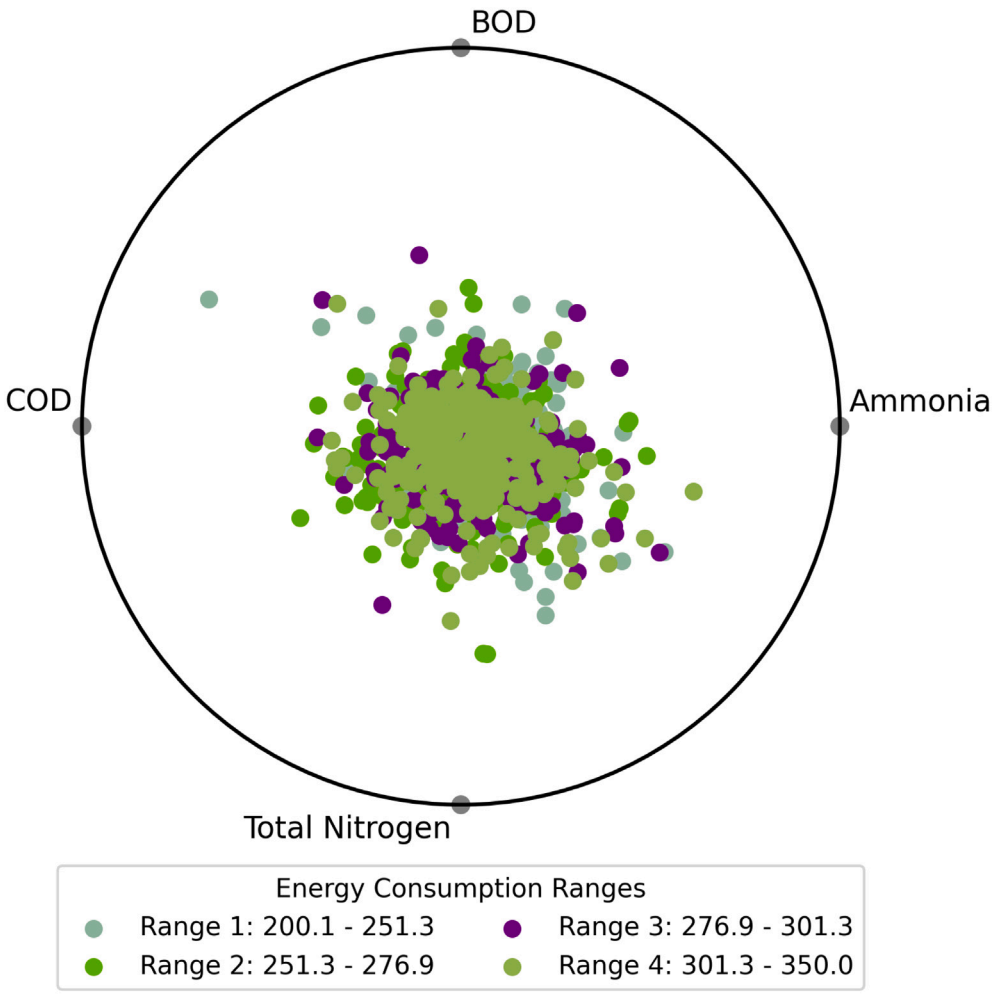

Fig. 5. RadViz visualization of wastewater factors impacting energy consumption.

between the higher the average inflow and average outflow, the higher the energy consumption.

The RadViz graph in Fig. 5 is an important visualization technique, that allows for the simultaneous analysis of multiple interconnected factors, resulting in a comprehensive understanding of the system and guiding decision-making to improve the environmental and economic sustainability of the wastewater treatment process. Here the graph depicts the interactions between four parameters in a wastewater treatment system: BOD, COD, ammonia, and total nitrogen. The three vertices of the top triangular graph indicate the extremes of BOD, COD, and ammonia, respectively. In contrast, the data points within the graph reflect a variety of these three elements in terms of total nitrogen content. Higher energy consumption is indicated colors (yellow, orange, and green), and lower energy consumption is indicated by purple colors. The data points in the graph are labeled with colors corresponding to the respective ranges of energy use. The amounts of organic matter in the water are indicated by high BOD levels, while the use of more energy-intensive biological treatment techniques. Since maintaining microbial activity is one of the most energy-intensive activities in biological treatment systems, increased BOD greater aeration. COD tracks all oxidizable chemicals in water, and more significant levels frequently demand chemical treatments such as oxidation, which can be energy-intensive. Like BOD, greater COD levels necessitate more aeration and mixing, increasing energy usage. Ammonia raises the oxygen demand in biological treatment procedures, requiring extra energy for aeration and nitrification. Temperature can also influence the energy required for nitrification, with colder temperatures requiring more energy to maintain treatment efficiency. Nitrogen removal frequently entails both nitrification and denitrification procedures, which can be energy-intensive depending on the need for aeration and mixing. Therefore, high levels of BOD, COD, ammonia, and total nitrogen increase energy consumption in a wastewater treatment plant.

Using data points grouped according to the similarities in the feature values, RadViz in Fig. 6 shows how various climate-related variables correspond with different amounts of energy usage. When more energy consumption points (dark green) are drawn towards Maximum Temperature or Average Humidity, it may indicate that the aspects of the climate influence energy consumption. A link between a feature and greater energy consumption may be implied if a certain energy consumption range has a propensity to cluster around specific anchors (for example, Range 4 close to Average Temperature). Similar feature profiles are probably shared by points that are closer to one another inside the circle. locations near the “Average Temperature” and “Average Humidity” anchors, for instance, may have greater values for these characteristics than other locations. On the other hand, points that are more centrally located are drawn equally by several features, indicating that the feature values are balanced.

4.2.2. Data analysis of wastewater quality

This section presents the relationship between all factors and wastewater factors (ammonia, COD, BOD). Firstly, we determine the most critical features that affect wastewater. Decreasing the input data dimensions helps us to simplify the model and obtain better accuracy. XGBRegressor, as a feature selection method, is used to select the

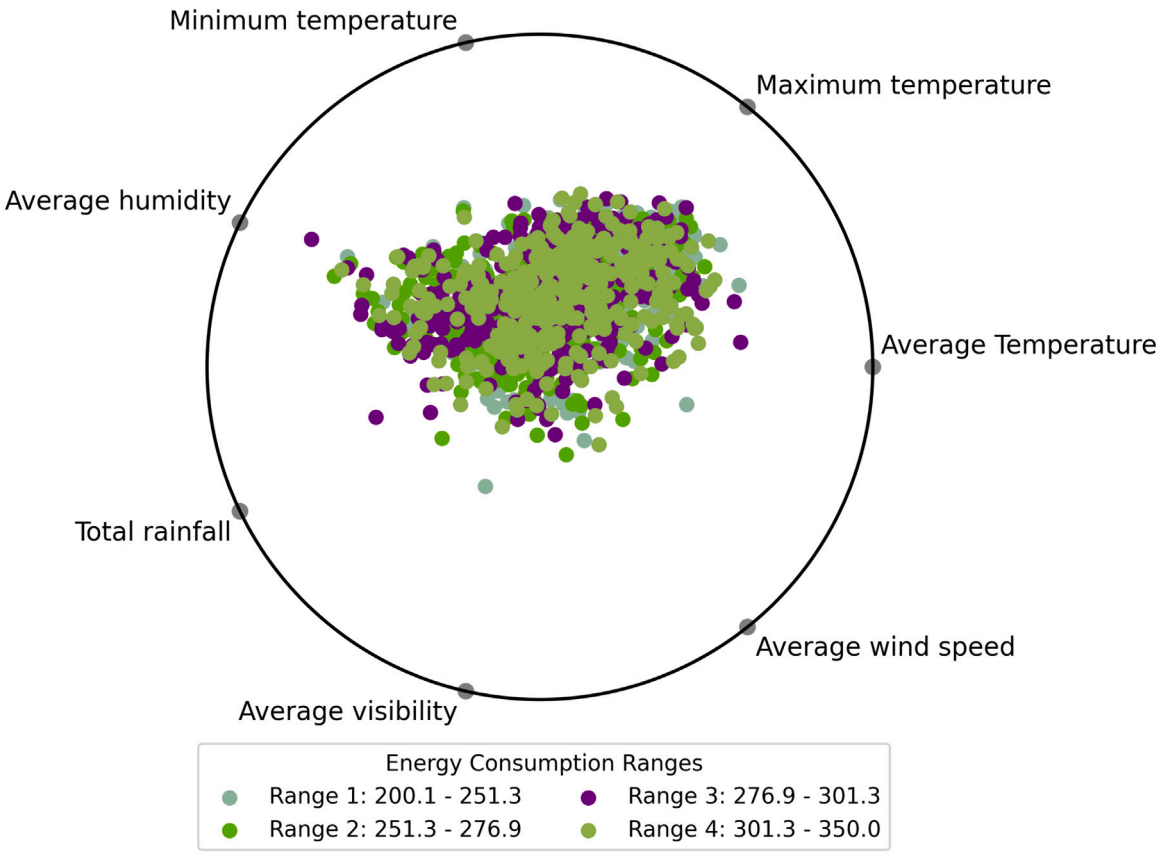

Fig. 6. RadViz visualization of climate factors impacting energy consumption.

crucial features from the dataset. Fig. 7 shows a feature importance analysis of factors affecting wastewater factors (ammonia, BOD, and COD). Average Outflow is the most influential feature, significantly affecting ammonia, COD, and BOD with F scores of 722, 775, and 823, respectively. Total Rainfall has the lowest impact on ammonia, COD, and BOD, with F scores of 23, 22, and 21, respectively. In our study, Average Outflow, Average Inflow, Energy Consumption, Average Temperature, Maximum Temperature, Average Humidity, Average Visibility, Average Wind Speed, and Day are used for applying multitask models to predict ammonia, COD, BOD in parallel.

This bubble diagram in Fig. 8 illustrates the relationship and interaction between the average temperature represented on the x-axis and the ammonia levels on the y-axis. The diagram consists of a set of bubbles whose size is proportional to the amount of rainfall, while their color gradient indicates the ammonia levels. Dark purple bubbles indicate high levels of ammonia, while lighter, greener colors indicate lower levels. It is also noticeable from the diagram that there is an inverse relationship between average temperatures and ammonia levels. This may be because high temperatures enhance the conversion of ammonia to nitrate, while low temperatures slow this process, which increases the amount of ammonia. The RadViz plot in Fig. 9 analyzes and interprets the changes in BOD levels in the wastewater treatment plants between full features. The graph shows the direct and influential effect of temperature, inflow, and outflow of water and humidity on BOD rates due to their impact on the biological and chemical processes in the wastewater and hence on the energy consumption rates (Srivastava et al., 2020). The least influential factors were the visibility rate and the rainfall rate again.

The RadViz plot in Fig. 10 illustrates the relationship between full features with COD. The purpose of the plot is to help analyze the relationship between COD requirements and the other variables. It is clear from the plot that this relationship exists, and it is a direct relationship between COD and both low temperature and wind and, thus, energy consumption. This result can be attributed to the fact that at low temperatures, chemical and biological reactions are reduced, which in turn leads to the decomposition of organic waste and thus increases the levels of COD in the water (Zheng et al., 2013). Conversely, it was found that there is an inverse relationship between high temperatures, which in turn accelerate the oxidation process of organic compounds, leading to a decrease in COD levels (Ma et al., 2021).

4.3. Results of predicting energy consumption

This section presents the results of LSTM, GRU, Bi-LSTM, Bi-GRU, RNN, and transformer models to predict energy consumption using different evaluation methods: RMSE, MAE, and MAPE with optimized and non-optimized data.

4.3.1. Comparison of models' performance on non-optimized data

Table 2 shows the results of models (LSTM, GRU, Bi-LSTM, Bi-GRU, RNN, and transformer) for each category hydraulic, climate, wastewater using different evaluation methods RMSE, MAE, and MAPE. It shows that transformer models achieve the best performance with the lowest error for each category. Models in the hydraulic and wastewater categories record the best results. In the hydraulic category, Bi-LSTM has the weakest performance and largest errors (RMSE = 36.6316 and MAPE = 10.7585). Regarding RMSE, LSTM and GRU report roughly the same values, at 36.4135 and 33.9426, respectively. The transformer operates the best with the fewest mistakes (RMSE = 35.6379 and MAPE = 10.6556). In the climate category, Bi-GRU performs the worst with the largest errors (RMS = 37.9915 and MAPE = 10.7109). Similar findings are recorded by Bi-LSTM and GRU with RMSE values of 37.2306 and 37.2357, respectively.

The transformer has the lowest mistakes and the best performance (RMSE = 35.9522 and MAPE = 10.7109). In the wastewater category,

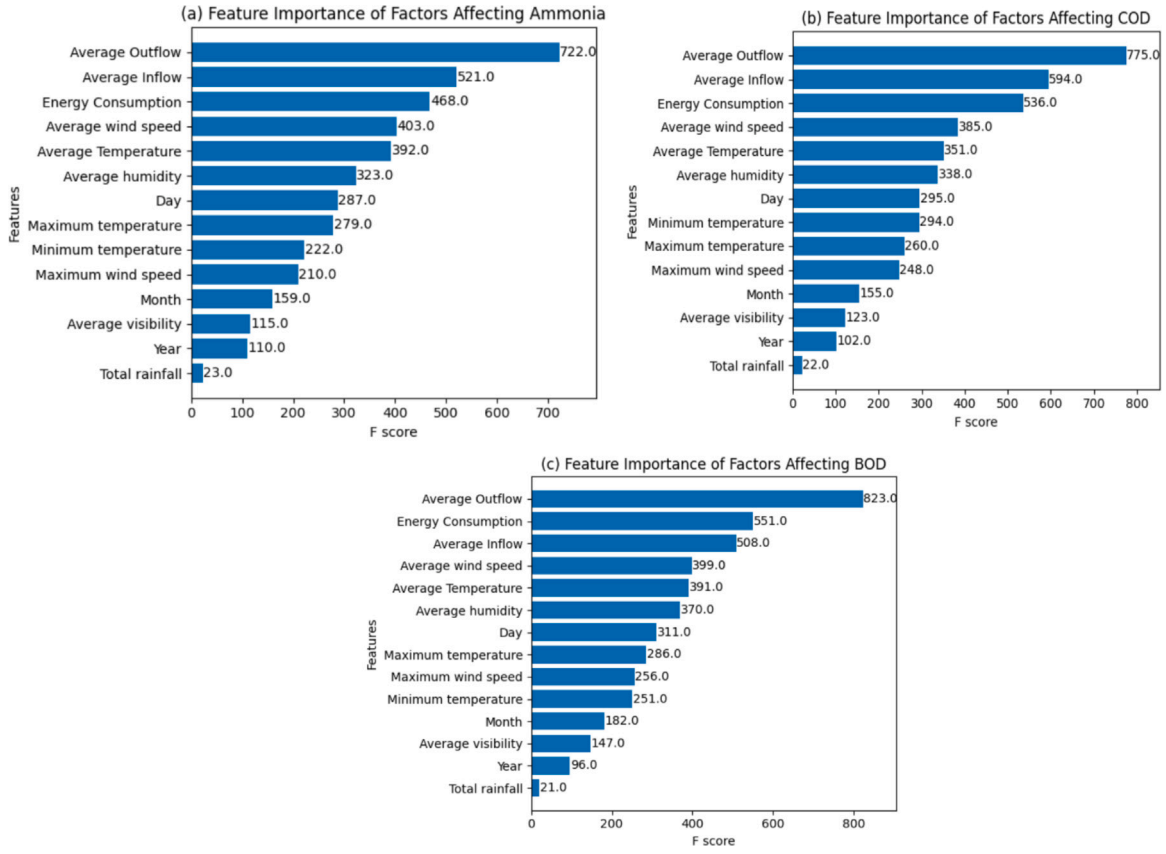

Fig. 7. Feature importance.

LSTM performs the worst with the largest mistakes (RMSE = 37.3055 and MAPE = 10.8828). The same results are recorded by Bi-LSTM and Bi-GRU, with RMSE values of 36.4002 and 36.4840, respectively. The transformer has the lowest mistakes and the best performance (RMSE = 35.1189 and MAPE = 10.0532). The proposed model voting model based on the voting from each transformer model across categories achieves the best performance with a value of 34.3219 for RMSE, which is a 1% improvement compared to transformer models. The results of the paired t-tests indicate statistically significant differences between the Voting model and other transformer models. Specifically, the p-values for the comparisons were as follows:p = 0.000431, 0.003296, and 0.003300 for Hydraulic-Transformer, Wastewater-Transformer, respectively. Since all p-values are below the standard significance threshold of 0.05, these results demonstrate that the Voting model's performance differs significantly from each of the compared models.

4.3.2. Comparison of models' performance on optimized data

Table 3 shows the results of models (LSTM, GRU, Bi-LSTM, Bi-GRU, RNN, and transformer) for each category, hydraulic, climate, and wastewater, using different evaluation methods, RMSE, MAE, and MAPE.

Transformer models achieve the best performance with the lowest error for each category. Models in the hydraulic and wastewater categories record the best results. According to the hydraulic category, Bi-LSTM performs the worst with the highest errors (RMSE = 35.1701 and MAPE = 10.148). LSTM and GRU record approximately the same results, with 33.9705 and 33.9426 RMSE, respectively. The transformer performs the best with the lowest errors (RMSE = 33.1359 and MAPE = 9.7757). According to the climate category, GRU performs the worst with the highest errors (RMS = 35.8630 and MAPE = 10.3181). Bi-LSTM and RNN record approximately the same results, with 34.0697 and 34.0278 RMSE, respectively. The transformer performs the best with the lowest errors (RMSE = 33.6919 and MAPE = 9.7918). According to the climate category, GRU performs the worst with the highest errors (RMS = 34.8288 and MAPE = 10.2256). Bi-LSTM and Bi-GRU record approximately the same results, with 33.7006 and 33.8133

Table 2
Comparison of models' performance in prediction energy consumption on non-optimized data.

Categories	Models	RMSE	MAE	MAPE
Hydraulic	LSTM	36.4135	29.8117	10.7300
	GRU	36.3303	29.7578	10.7367
	Bi-LSTM	36.6316	30.0186	10.7585
	Bi-GRU	35.9076	29.3364	10.7101
	RNN	36.3316	29.1489	10.8706
	Transformer	35.6379	28.5425	10.6556
Climate	LSTM	37.5717	30.8289	10.9247
	GRU	37.2357	30.5595	10.8717
	Bi-LSTM	37.2306	30.5557	10.8709
	Bi-GRU	37.9915	31.1724	10.9987
	RNN	36.9790	29.5944	11.1755
	Transformer	35.9522	29.3904	10.7109
wastewater	LSTM	37.3055	30.6166	10.8828
	GRU	37.1953	30.5259	10.8653
	Bi-LSTM	36.4002	29.8209	10.7443
	Bi-GRU	36.4840	29.8923	10.7531
	RNN	37.1210	30.4632	10.8532
	Transformer	35.1189	28.1657	10.3136
–	Voting model	34.3219	27.5504	10.0532

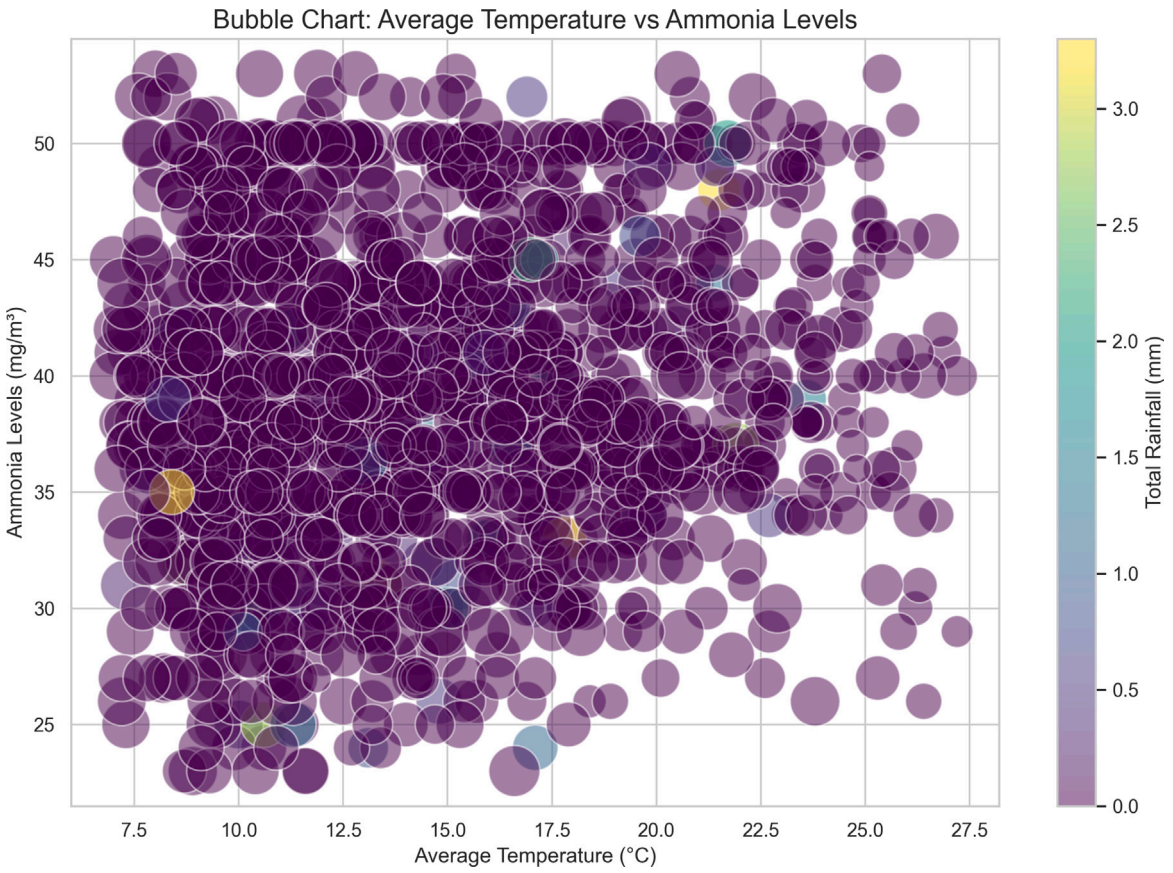

Fig. 8. Bubble chart for ammonia.

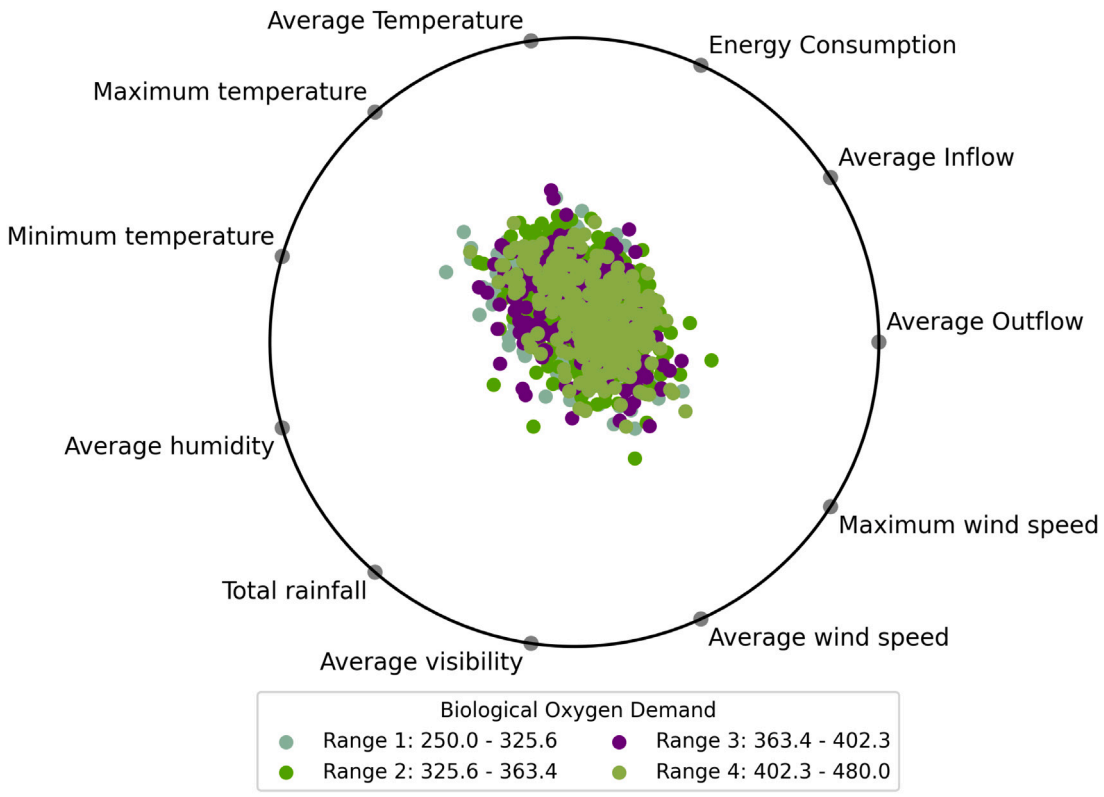

Fig. 9. RadViz visualization of factors impacting BOD.

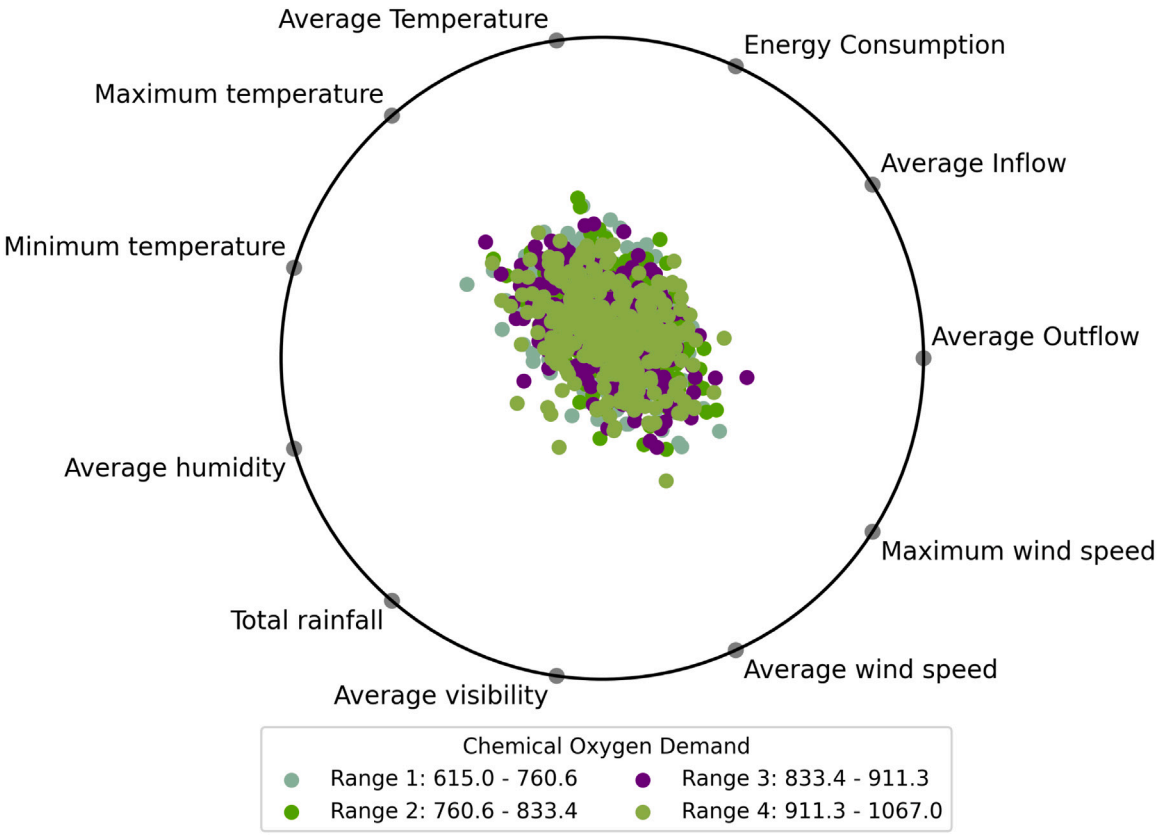

Fig. 10. RadViz visualization of climate factors impacting COD.

RMSE, respectively. The transformer performs the best with the lowest errors (RMSE = 33.1569 and MAPE = 9.9301). The proposed voting model, based on the voting from each transformer model across categories, achieves the best performance, with RMSE of 31.6163, 3% improvement compared to transformer models.

The results of the paired t-tests indicate statistically significant differences between the Voting model and other transformer models. Specifically, the p-values for the comparisons were as follows: p = 0.000373, 0.000102, and 0.002588 for Hydraulic-Transformer, Wastewater-Transformer, respectively. Since all p-values are below the standard significance threshold of 0.05, these results demonstrate that the Voting model's performance differs significantly from each of the compared models.

The lowest RMSE value reported by Voting model in Table 3, 31.6163 kWh/m^3, represents a significant improvement compared to transformer model and LSTM, GRU, Bi-LSTM, Bi-GRU and RNN models, which ranged from 33 to 37 kWh/m^3. While the absolute value of RMSE depends on plant-specific energy usage patterns, values in the low 30 s are considered accurate enough for supporting operational energy planning in WWTPs. Notably, this level of accuracy allows plant operators to anticipate high-demand periods, optimize aeration or pumping schedules, and identify anomalies in energy usage that may signal equipment inefficiencies. Therefore, the model's predictive power is not only statistically significant but also practically useful in a real-world decision-support context.

Fig. 11 displays fluctuations in energy consumption predictions across transformer models of features for wastewater, climate, and hydraulic, and the proposed model (voting) with some degree of variation between them. The real data line appears to have more frequent and extreme peaks, while the other models generally track it but with differing degrees of accuracy and smoothness. The Voting method (yellow) represents a combined model to improve accuracy over individual model predictions.

Table 3
Comparison of models' performance in prediction energy consumption on optimized data.

Categories	Models	RMSE	MAE	MAPE
Hydraulic	LSTM	33.9426	27.6674	9.9319
	GRU	33.9705	28.0494	9.9834
	Bi-LSTM	35.1701	28.3291	10.148
	Bi-GRU	33.7598	27.8797	9.9579
	RNN	33.6251	27.3770	9.8309
	Transformer	33.1359	27.0887	9.7757
Climate	LSTM	34.4386	28.4531	10.0624
	GRU	35.8630	29.5826	10.3181
	Bi-LSTM	34.0697	28.1344	9.9987
	Bi-GRU	34.6578	28.6268	10.0978
	RNN	34.0278	28.0977	10.1573
	Transformer	33.6919	27.3345	9.7918
Wastewater	LSTM	34.7499	28.6990	10.1131
	GRU	34.8288	28.4492	10.2256
	Bi-LSTM	33.7006	27.8302	9.9511
	Bi-GRU	33.8133	27.9231	9.9640
	RNN	33.565	26.2028	9.4705
	Transformer	33.1569	27.3378	9.9301
–	Voting model	31.6163	24.3304	8.9026

4.4. Results of single-task models to predict wastewater quality

This section presents the results of baseline models (LSTM, GRU, Bi-LSTM, Bi-GRU, and RNN) for predicting BOD, COD, and ammonia as individual values. They applied both the full feature set and selected features.

Table 4 shows that models with selected features achieved the best performance compared to models with full features. Comparing Bi-GRU to other models, it records the best performance with the fewest errors. Bi-GRUs analyze the data both forward and backward,

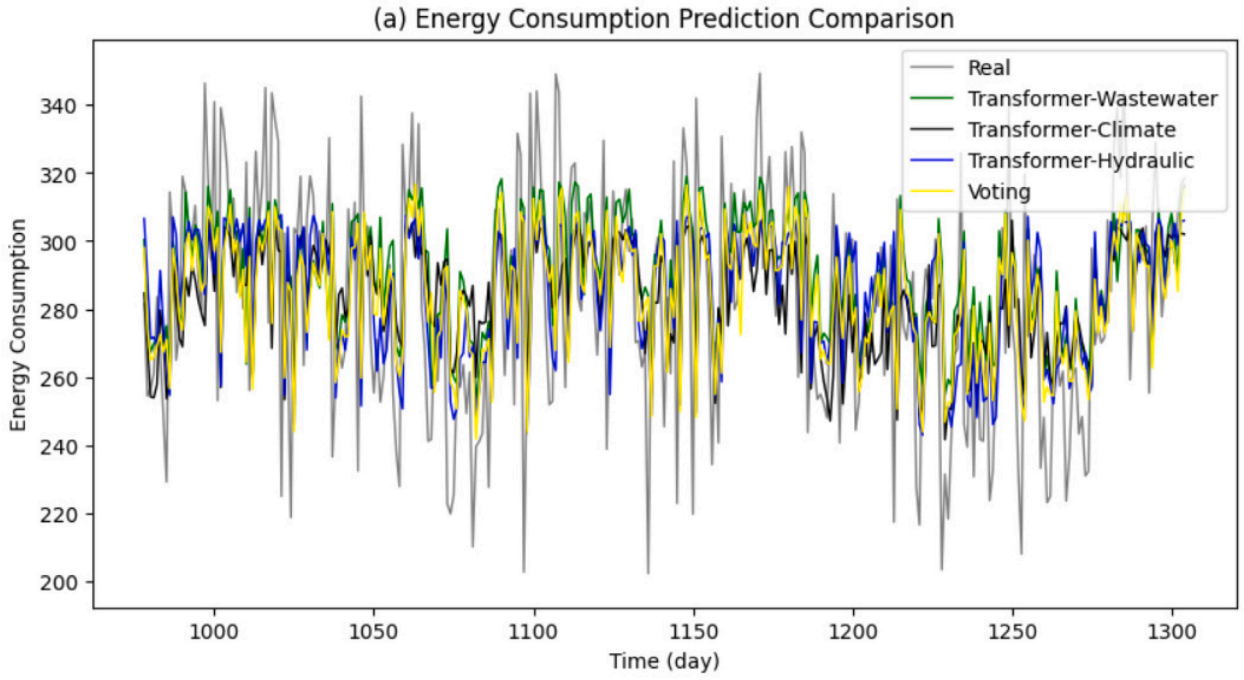

Fig. 11. Comparison between actual data and predicted data of energy consumption.

efficiently capturing context from the sequence's previous and subsequent phases. Furthermore, this simpler GRU structure helps Bi-GRUs learn relationships in sequences more effectively and lowers computing effort.

In full features, the LSTM performs the worst and has the highest errors, with RMSEs of 8.8120, 51.0455, 108.8735, for ammonia, BOD, and COD respectively. The best results are obtained by Bi-GRU, which has the lowest RMSE for BOD and COD, respectively, at 7.0700, 47.5811, 93.6207. In selected features, BI-GRU records the best performance with the lowest error across all targets: ammonia (RMSE = 6.2299 and MAPE = 14.2190), BOD (RMSE = 47.9506 and MAPE = 11.4435) and COD (RMSE = 89.0222 and MAPE = 9.2546). LSTM performs the worst results, with the highest error of 6.7138 RMSE for ammonia and 55.5588 RMSE for BOD.

4.5. Results of multitask models to predict wastewater quality

This section presents the results of multitask models (LSTM-Multitask, GRU-Multitask, BI-LSTM-Multitask, BI-GRU-Multitask, and RNN-Multitask) to predict BOD, COD, and ammonia in parallel using different evaluation methods: RMSE, MAE, and MAPE. They applied both the full feature set and selected features.

Table 5 shows the results of the multitask models to forecast wastewater values with optimized data and full features and selected features. We can see that models with selected features achieved the best performance compared to models with full features. Bi-GRU-Multitask records the best performance with the lowest errors compared to other models. The model with ammonia records the lowest RMSE and the highest MAPE. Models with COD record the highest RMSE and the lowest MAPE.

In full features, LSTM-Multitask records the worst performance with the most significant errors, with RMSE at 8.7950, 51.4978, and 104.2075 for ammonia, BOD, and COD, respectively. Bi-GRU-Multitask achieves the best performance with the lowest RMSE at 6.6578, 46.9277, and 90.6838 for ammonia, BOD, and COD, respectively.

In selected features, BI-GRU-Multitask records the best performance with the lowest error across all targets: ammonia (RMSE = 6.1689 and MAPE = 13.0589), BOD (RMSE = 48.0323 and MAPE = 12.2143) and COD (RMSE = 88.2214 and MAPE = 9.0292). LSTM-Multitask performs the worst results, with the highest error of 6.4416 RMSE for ammonia and 97.4463 RMSE for COD. Fig. 12 contains three line plots comparing the performance of different multitask models in predicting three parameters over time: ammonia, BOD, and COD with the x-axis indicating Time (day) and the y-axis indicating the parameter value.

Table 4
Comparison of single-task models' performance in prediction wastewater.

Features	Models	Targets	RMSE	MAE	MAPE
Full features	LSTM	Ammonia	8.8120	7.3847	20.9270
		BOD	51.0455	41.8896	12.8770
		COD	108.8735	86.5003	10.3927
	GRU	Ammonia	7.3856	5.8361	15.3636
		BOD	54.0472	44.9566	13.9884
		COD	105.6655	85.9383	9.7571
	Bi-LSTM	Ammonia	7.0303	6.3280	16.4437
		BOD	52.1656	44.5405	13.3564
		COD	105.6414	85.8574	9.7697
	Bi-GRU	Ammonia	7.0700	5.6184	16.2050
		BOD	47.5811	39.5620	12.4540
		COD	93.6207	78.1180	9.2229
	RNN	Ammonia	7.9533	6.4196	19.9019
		BOD	53.3189	44.6520	12.3839
		COD	99.5006	79.5850	9.4540
Selected features	LSTM	Ammonia	6.7138	5.4788	14.5395
		BOD	55.5588	45.2308	14.8189
		COD	102.2044	84.8195	9.6165
	GRU	Ammonia	6.3275	4.9330	15.9104
		BOD	50.0969	40.9679	12.8555
		COD	107.9807	86.0623	9.7553
	Bi-LSTM	Ammonia	6.5463	5.2277	15.8665
		BOD	53.6910	45.3937	14.1784
		COD	94.8595	75.8942	9.5583
	Bi-GRU	Ammonia	6.2299	5.0258	14.2190
		BOD	47.9506	40.0708	11.4435
		COD	89.0222	73.5907	9.2546
	RNN	Ammonia	6.3509	5.2929	15.3396
		BOD	48.8841	39.3360	13.0508
		COD	98.5006	77.5850	9.8540

Table 5
Comparison of multitask models' performance in prediction wastewater.

Features	Models	Targets	RMSE	MAE	MAPE
Full features	LSTM-Multitask	Ammonia	8.7950	7.1638	20.7605
		BOD	51.4978	42.2318	13.3483
		COD	104.2075	84.6781	9.6434
	GRU-Multitask	Ammonia	6.8893	5.5797	14.9091
		BOD	52.6701	43.0118	13.7473
		COD	104.6424	83.2028	9.6252
	Bi-LSTM-Multitask	Ammonia	6.9120	5.7069	15.3801
		BOD	51.4063	41.8812	13.1628
		COD	103.3811	83.9029	9.5614
	Bi-GRU-Multitask	Ammonia	6.6578	5.5123	15.0829
		BOD	46.9277	37.4812	12.0856
		COD	90.6838	74.7268	8.7368
	RNN-Multitask	Ammonia	7.8755	6.3774	18.2660
		BOD	51.4036	40.4983	11.9897
		COD	97.9017	78.0267	9.2862
Selected features	LSTM-Multitask	Ammonia	6.4416	5.1371	16.1310
		BOD	54.3157	44.4416	14.5893
		COD	97.4463	76.9276	9.7217
	GRU-Multitask	Ammonia	6.1996	4.8373	15.8807
		BOD	49.4814	39.6500	12.7779
		COD	104.2822	82.8013	9.7936
	Bi-LSTM-Multitask	Ammonia	6.3161	5.0355	15.7557
		BOD	51.5169	42.1989	13.8947
		COD	92.9662	73.8588	9.4497
	Bi-GRU-Multitask	Ammonia	6.1689	4.8518	13.0589
		BOD	48.0323	39.5788	12.2143
		COD	88.2214	70.5225	9.0292
	RNN-Multitask	Ammonia	6.2517	4.9733	14.8637
		BOD	47.3206	38.9054	12.8414
		COD	96.1963	75.4696	9.5437

Note that the COD predictions consistently show the highest RMSE but also the lowest MAPE across all models. This discrepancy is primarily due to the scale and distribution of COD values in the dataset, which tend to be substantially higher than those of ammonia and BOD. RMSE captures absolute error in the same units as the target variable, and therefore, naturally increases with the magnitude of the values being predicted. In contrast, MAPE measures relative error and becomes smaller when predictions deviate by a modest percentage—even if the absolute deviation is large. This highlights a known issue in regression evaluation where RMSE and MAPE may diverge in interpretability depending on the scale of the target variable. As a result, the reported metrics should be interpreted jointly rather than in isolation to assess model effectiveness across different targets.

In summary, the Bi-GRU-Multi-task model performed best across quality indicators, supporting the hypothesis that temporal and statistical dependencies exist among Ammonia, BOD, and COD. This reflects biological and chemical linkages in treatment processes, such as how ammonia oxidation affects COD removal under varying aeration and sludge retention times.

4.6. Comparison with literature studies

Table 6 compares previous studies and our work based on the number of features, types of models, and pre-preprocessing techniques. In our work, we proposed a novel ensemble voting model combining predictions from multiple transformer models to enhance the accuracy and robustness of energy consumption prediction. We applied special pre-processing techniques to optimize the dataset.

The authors (Bagherzadeh et al., 2021b) applied GBM, RF, ANN, RNN, and KNN at RMER 33.9, 34.8, 39.8, 37.3, and 37.33, respectively. The models used included nine selected features: Months, Total Nitrogen, Ammonia, BOD, Maximum Temperature, Average Humidity, Total Rainfall, and Average Inflow. The authors in Alali et al. (2023) applied XGBoost, LightGBM, GPRRQ, and GSVR with features selected by XGboost and lag column of energy consumption that recorded 37.33, 37.14, 37.38, 37.45, and 37.7. Harrou et al. (2023) applied different DL models RNN, LSTM, GRU, BiLSTM, and BiGRU with selected features by XGBoost algorithms, including the month, daily inflow rate (Q), average humidity, TN, BOD, and ammonia, and recorded RMSE at 38.655, 38.683, 37.759, 39.064, and 39.114, respectively.

We can see that the transformer of each category records the best results compared to other studies, and the voting ensemble model gets the best results compared to the transformer with 31.6163 RMSE. The optimized data with transformer models improve the performance of models for energy consumption, and the voting ensemble model using full features improved the performance of models.

5. Study limitations and future work

This study presents a significant advancement for both the practical and technical aspects of WWTP management. From a business and operational perspective, the proposed framework addresses urgent challenges in energy optimization and wastewater quality monitoring—two critical domains for sustainable urban infrastructure. By providing accurate, real-time predictions for energy consumption and key water quality indicators (ammonia, BOD, and COD), the study enables WWTP operators to make informed decisions that enhance operational efficiency and reduce environmental impact. From a deep learning standpoint, the study introduces a novel ensemble model based on transformer architectures, which aggregates predictions across heterogeneous feature categories (hydraulic, climatic, and wastewater). Additionally, a multi-task Bi-GRU model is proposed for the simultaneous prediction of multiple wastewater quality parameters, enabling better generalization and reduced computational burden.

The study also employs a dual-method data preprocessing strategy — using z-score and IQR/KNN — to improve data quality based on the statistical characteristics of different variables. Importantly, the hyperparameter tuning process is guided by Bayesian optimization, which offers an efficient and principled approach to enhancing model performance without exhaustive grid searching. Despite these strengths, the study has several limitations that warrant future investigation.

First, while the proposed models demonstrate high accuracy, their interpretability remains limited. Deep learning architectures, particularly transformer ensembles and recurrent networks, are often viewed as black boxes. For WWTP operators to trust and adopt such systems in practice, it is essential to provide explanations for model outputs. Future work will therefore focus on enhancing model interpretability using explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations), attention visualization, and local surrogate models like LIME. These tools will help uncover feature importance and provide transparent insights into how predictions are generated.

Second, the current study is trained and evaluated on data from a single wastewater treatment facility in Melbourne, which may limit its generalizability. Operational dynamics and environmental conditions vary significantly across regions and plant configurations. To improve robustness and external validity, future research should incorporate data from multiple WWTPs and investigate transfer learning strategies to adapt the model with minimal retraining. This would ensure that the framework can be deployed more broadly across varying contexts.

Third, while the model incorporates a comprehensive set of hydraulic, climate, and wastewater quality variables, it does not include additional operational or real-time control parameters such as chemical dosing, aeration levels, or sensor health indicators. These factors can provide valuable contextual information, particularly under abnormal or stress conditions (e.g., stormwater surges or system faults). Future studies should explore the integration of such features to further improve the accuracy and responsiveness of the model.

Fourth, the current framework optimizes for prediction accuracy but does not explicitly incorporate economic factors such as cost savings,

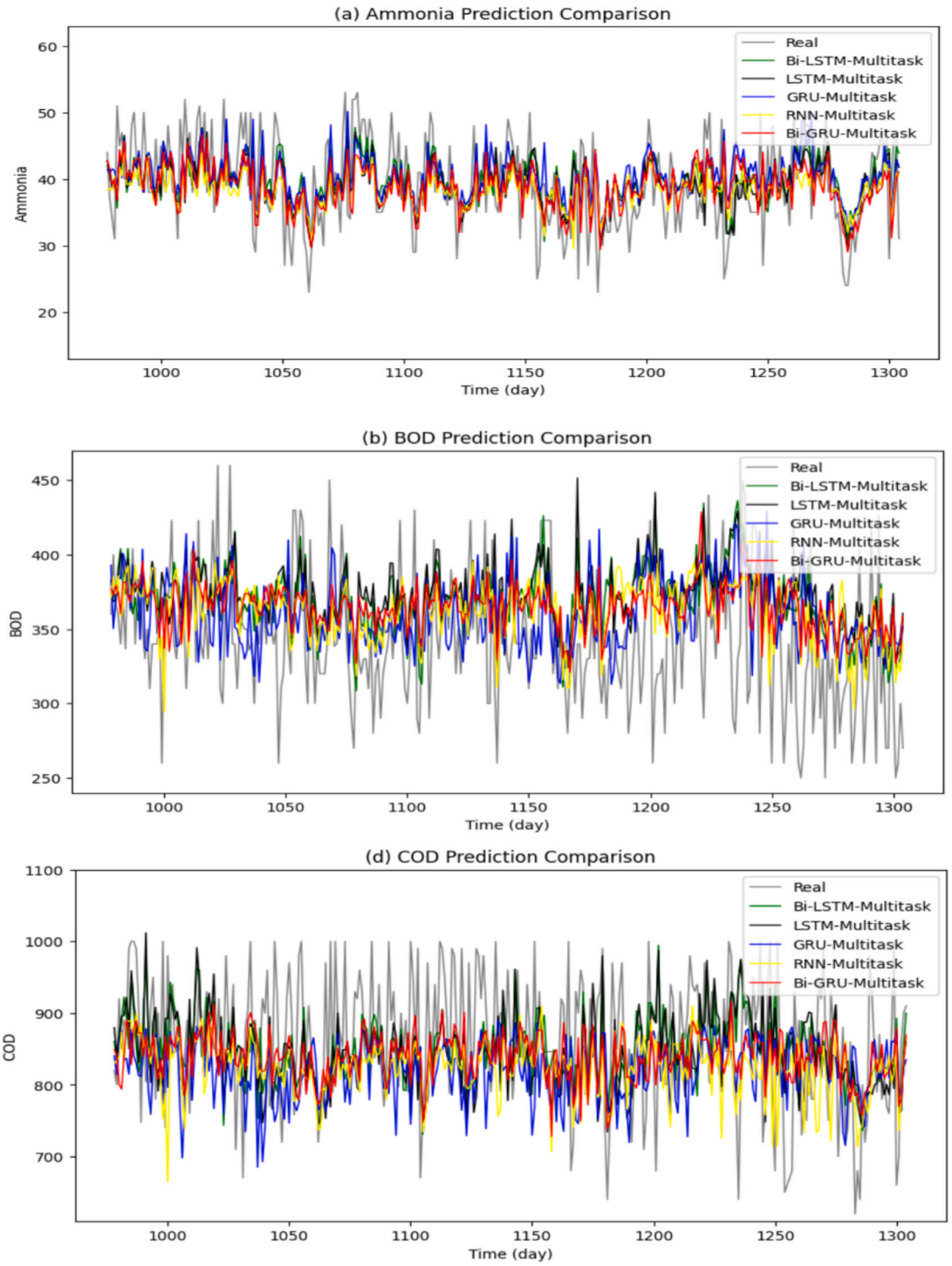

Fig. 12. Comparison between actual data and predicted data of models for wastewater quality with optimized data.

energy pricing dynamics, or chemical usage efficiency. For practical deployment in business-critical infrastructure, it is important to align predictions with actionable outcomes that reflect financial constraints. Future work will consider incorporating cost-aware optimization objectives or reinforcement learning approaches that allow WWTPs to minimize operational costs while maintaining quality standards.

Another limitation of the current study is that the proposed framework has been developed and validated using data from a single WWTP located in Melbourne, Australia. While the dataset is rich and diverse in terms of temporal coverage and multivariate inputs, it does not capture the operational heterogeneity, infrastructural differences, or climatic variability found in WWTPs across different geographical regions. This may limit the generalizability of the model to other sites with different treatment technologies, influent compositions, or weather profiles. Although cross-site validation was beyond the scope of this study due to data access constraints, we recognize its critical importance for broader applicability.

Future work will therefore focus on evaluating the transferability of the model using datasets from multiple WWTPs, and incorporating transfer learning or domain adaptation techniques to enable effective generalization with minimal retraining. In addition, we plan to perform sensitivity analyses to systematically assess how variations in input

Table 6
Comparison with literature studies in predicting energy consumption.

Papers	Methods	Transformer models	RMSE	MAE	MAPE
Bagherzadeh et al. (2021b)	GBM	No	33.9	26.9	–
	RF	No	34.8	27.7	–
	ANN	No	39.8	32.1	–
	RNN	No	37.3	29.3	–
	KNN	No	37.33	28.23	10.65
Alali et al. (2023)	XGBoost	No	37.14	28.5	10.81
	LightGBM	No	37.38	28.63	10.96
	GPRRQ	No	37.45	28.65	10.04
	GSVR	No	37.7	28.88	10.12
Harrou et al. (2023)	RNN	No	38.655	29.446	11.196
	GRU	No	38.683	29.771	10.88
	LSTM	No	37.759	28.698	10.738
	BiGRU	No	39.064	30.087	11
	BiLSTM	No	39.114	30.055	11.001
Our work	Transformer-hydraulic	Yes	33.1359	27.0887	9.7757
	Transformer-climate	Yes	33.6919	27.3345	9.7918
	Transformer-wastewater	Yes	33.1569	27.3378	9.9301
	The voting ensemble model	Yes	31.6163	24.3304	8.9026

features — such as temperature, rainfall, or inflow volume — affect the model's performance, further enhancing its robustness under diverse operational conditions. In addition, we will consider applying hyperparameter optimization techniques — such as grid search, random search, and Bayesian optimization — to select the best model architecture.

In addition, the study does not include a real-time deployment or scalability demonstration. Practical implementation in operational WWTPs requires integration with real-time data streams, compatibility with existing SCADA systems, and efficient model inference under limited computational resources. These engineering and systems-level considerations are critical for bridging the gap between research and industrial adoption. As part of future work, we aim to develop a lightweight, scalable deployment pipeline using tools such as TensorFlow Lite or ONNX to support real-time inference. Addressing system integration, fault tolerance, and latency issues will be central to ensuring that the proposed models can be reliably adopted in production environments.

Finally, the model has been developed and evaluated in an offline setting. However, deployment in live operational environments will require real-time integration with SCADA systems, sensor streams, and decision-support interfaces. Future extensions will involve building a complete predictive analytics pipeline capable of ingesting streaming data, executing lightweight inference (e.g., using ONNX or TensorFlow Lite), and offering user-friendly feedback to WWTP operators. A real-world implementation will also allow for the assessment of model usability, reliability, and user acceptance.

6. Conclusion

This paper has introduced a novel and unique prediction framework to optimize energy consumption and wastewater quality in WWTPs. The proposed framework incorporated predictions from hydraulic, wastewater, and meteorological data by incorporating a voting ensemble of transformer models. The findings showed performance improvement, with a 3% drop in RMSE to 31.62. Using Bi-GRU models with optimized data, the creative use of a multi-task DL model improves operational efficiency by concurrently predicting three important wastewater quality parameters, i.e., ammonia, BOD, and COD, with RMSE values of 6.1689, 48.0323, and 88.2214, respectively. Additionally, the proposed novel multi-task deep learning models effectively predicted wastewater quality indicators (ammonia, BOD, and COD) with superior performance metrics.

Improved model performance and strong predictions were the outcome of a thorough data-cleaning procedure that included z-score and KNN imputation techniques to guarantee the dataset's quality and dependability. The results demonstrated how flexible and resilient the proposed predictive framework has handled the dynamic and nonlinear interactions that are a part of WWTP operations. An important outcome of this study is the realization that although energy consumption forecasting and effluent quality prediction represent distinct objectives, they benefit from being addressed within a unified framework. The separate modeling of each task — using transformer ensembles for energy and multi-task Bi-GRU for water quality — allowed us to tailor architectures to the specific characteristics of each domain. At the same time, their integration supports coordinated operational insight, highlighting how predictive learning across multiple, interdependent plant functions can collectively enhance the sustainability and efficiency of WWTP management.

In summary, this study demonstrates that intelligently combining deep learning architectures — specifically transformer ensembles and multi-task recurrent networks — can address multiple interdependent objectives in wastewater treatment operations. The integration of these methods is not simply technical but serves a clear operational goal: to provide WWTP operators with a data-driven decision support system that optimizes energy consumption while ensuring effluent quality. The resulting framework improves predictive performance and reduces model complexity through shared representations, robust feature selection, and targeted preprocessing. This work moves beyond isolated prediction tasks to offer a scalable, interpretable, and sustainable solution for real-world WWTP management. Future work could explore expanding the predictive framework's applicability to other critical parameters, further improving its scalability and impact on sustainable urban development.

CRediT authorship contribution statement

Hager Saleh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Data curation. **Sherif Mostafa:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation. **Shaker El-Sappagh:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Abdulaziz AlMohimeed:** Methodology, Writing – review & editing, Investigation, Visualization. **Michael McCann:** Writing – review & editing, Writing – original draft, Investigation, Funding acquisition. **Saeed Hamood Alsamhi:** Writing – review & editing, Writing – original draft, Investigation. **Niall O'Brolchain:** Writing – review & editing, Writing – original draft. **John G. Breslin:** Writing – review & editing, Writing – original draft, Software, Investigation, Funding acquisition. **Marwa E. Saleh:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation.

Declaration of competing interest

All authors declare that they have no conflicts of interest.

Acknowledgments

This publication has emanated from research conducted with the financial support of Taighde Éireann - Research Ireland under Grant Numbers 12/RC/2289_P2 (Insight) and 21/FFP-A/9174 (SustAIn). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Furthermore, this publication has emanated from research conducted with the financial support of the Interreg North-West Europe project ResNRJwater (NWE0200163).

Data availability

All datasets used to support the findings of this study are available from the direct link in the dataset citation.

References

Abba, S.I., Elkiran, G., 2017. Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application. Procedia Comput. Sci. 120, 156–163.

Abdelgwad, M.M., Soliman, T.H.A., Taloba, A.I., Farghaly, M.F., 2022. Arabic aspect based sentiment analysis using bidirectional GRU based models. J. King Saud Univ.-Comput. Inf. Sci. 34 (9), 6652–6662.

Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A., Gandomi, A.H., 2023. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif. Intell. Rev. 56 (11), 13521–13617.

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

Alsulaili, A., Refaie, A., 2021. Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. Water Supply 21 (5), 1861–1877.

Alvi, M., Batstone, D., Mbamba, C.K., Keymer, P., French, T., Ward, A., Dwyer, J., Cardell-Oliver, R., 2023. Deep learning in wastewater treatment: a critical review. Water Res. 120518.

Baarimah, A.O., Bazel, M.A., Alaloul, W.S., Alazaiza, M.Y., Al-Zghoul, T.M., Al-muhaya, B., Khan, A., Mushtaha, A.W., 2024. Artificial intelligence in wastewater treatment: Research trends and future perspectives through bibliometric analysis. Case Stud. Chem. Environ. Eng. 100926.

Bagherzadeh, F., Nouri, A.S., Mehrani, M.-J., Thennadil, S., 2021a. 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., Nouri, A.S., Mehrani, M.-J., Thennadil, S., 2021b. 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.

Baki, O.T., Aras, E., Akdemir, U.O., Yilmaz, B., 2019. Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models. Desalination Water Treat. 157, 79–89.

Belletti, F., Chen, M., Chi, E.H., 2019. Quantifying long range dependence in language and user behavior to improve RNNs. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1317–1327.

Cardoso, B.J., Rodrigues, E., Gaspar, A.R., Gomes, Á., 2021. Energy performance factors in wastewater treatment plants: A review. J. Clean. Prod. 322, 129107.

Chikodili, N.B., Abdulmalik, M.D., Abisoye, O.A., Bashir, S.A., 2020. Outlier detection in multivariate time series data using a fusion of K-medoid, standardized euclidean distance and Z-score. In: International Conference on Information and Communication Technology and Applications. Springer, pp. 259–271.

Das, A., Kumawat, P.K., Chaturvedi, N.D., 2021. A study to target energy consumption in wastewater treatment plant using machine learning algorithms. In: Computer Aided Chemical Engineering. Vol. 50, Elsevier, pp. 1511–1516.

Hao, X., Liu, R., Huang, X., 2015. Evaluation of the potential for operating carbon neutral WWTPs in China. Water Res. 87, 424–431.

Harrou, F., Dairi, A., Dorbane, A., Sun, Y., 2023. Energy consumption prediction in water treatment plants using deep learning with data augmentation. Results Eng. 20, 101428.

Heddam, S., Lamda, H., Filali, S., 2016. Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: a comparative study. Environ. Process. 3, 153–165.

Lv, H., Chen, J., Pan, T., Zhang, T., Feng, Y., Liu, S., 2022. Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application. Measurement 199, 111594.

Ma, D., Yi, H., Lai, C., Liu, X., Huo, X., An, Z., Li, L., Fu, Y., Li, B., Zhang, M., et al., 2021. Critical review of advanced oxidation processes in organic wastewater treatment. Chemosphere 275, 130104.

Mekaoussi, H., Heddam, S., Bouslimanni, N., Kim, S., Zounemat-Kermani, M., 2023. Predicting biochemical oxygen demand in wastewater treatment plant using advance extreme learning machine optimized by Bat algorithm. Heliyon 9 (11).

Ming, Y., Cao, S., Zhang, R., Li, Z., Chen, Y., Song, Y., Qu, H., 2017. Understanding hidden memories of recurrent neural networks. In: 2017 IEEE Conference on Visual Analytics Science and Technology. VAST, IEEE, pp. 13–24.

Molinos-Senante, M., Sala-Garrido, R., Iftimi, A., 2018. Energy intensity modeling for wastewater treatment technologies. Sci. Total Environ. 630, 1565–1572.

Niu, Z., Zhong, G., Yue, G., Wang, L.-N., Yu, H., Ling, X., Dong, J., 2023. Recurrent attention unit: A new gated recurrent unit for long-term memory of important parts in sequential data. Neurocomputing 517, 1–9.

Oliveira, P., Fernandes, B., Analide, C., Novais, P., 2021. Forecasting energy consumption of wastewater treatment plants with a transfer learning approach for sustainable cities. Electronics 10 (10), 1149.

Oulebsir, R., Lefkir, A., Safri, A., Bermad, A., 2020. Optimization of the energy consumption in activated sludge process using deep learning selective modeling. Biomass Bioenergy 132, 105420.

Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia 4 (2), 1883.

Qambar, A.S., Al Khalidy, M.M., 2022. Prediction of municipal wastewater biochemical oxygen demand using machine learning techniques: a sustainable approach. Process. Saf. Environ. Prot. 168, 833–845.

Qiao, J., Zhou, H., 2018. Modeling of energy consumption and effluent quality using density peaks-based adaptive fuzzy neural network. IEEE/CAA J. Autom. Sin. 5 (5), 968–976.

Ramli, N.A., Hamid, M.A., 2018. Data based modeling of a wastewater treatment plant by using machine learning methods. J. Eng. Technol. 6 (1), 14–21.

Saleh, B.A., Kayi, H., 2021. Prediction of chemical oxygen demand from the chemical composition of wastewater by artificial neural networks. In: Journal of Physics: Conference Series. Vol. 1818, IOP Publishing, 012035.

Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D: Nonlinear Phenom. 404, 132306.

Shewalkar, A., Nyavanandi, D., Ludwig, S.A., 2019. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. J. Artif. Intell. Soft Comput. Res. 9 (4), 235–245.

Siami-Namini, S., Tavakoli, N., Namin, A.S., 2019. The performance of LSTM and BiLSTM in forecasting time series. In: 2019 IEEE International Conference on Big Data (Big Data). IEEE, pp. 3285–3292.

Srivastava, R.K., Shetti, N.P., Reddy, K.R., Aminabhavi, T.M., 2020. Sustainable energy from waste organic matters via efficient microbial processes. Sci. Total Environ. 722, 137927.

Torregrossa, D., Leopold, U., Hernández-Sancho, F., Hansen, J., 2018. Machine learning for energy cost modelling in wastewater treatment plants. J. Environ. Manag. 223, 1061–1067.

Torregrossa, D., Schutz, G., Cornelissen, A., Hernández-Sancho, F., Hansen, J., 2016. Energy saving in WWTP: Daily benchmarking under uncertainty and data availability limitations. Environ. Res. 148, 330–337.

Vinutha, H., Poornima, B., Sagar, B., 2018. Detection of outliers using interquartile range technique from intrusion dataset. In: Information and Decision Sciences: Proceedings of the 6th International Conference on Ficta. Springer, pp. 511–518.

Yang, Y., Qu, Z., Yan, Z., Gao, Z., Wang, T., 2024. Network configuration entity extraction method based on transformer with Multi-Head attention mechanism. Comput. Mater. Contin. 78 (1).

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 31 (7), 1235–1270.

Yusuf, J., Faruque, R.B., Hasan, A.J., Ula, S., 2019. Statistical and deep learning methods for electric load forecasting in multiple water utility sites. In: 2019 IEEE Green Energy and Smart Systems Conference. IGESSC, IEEE, pp. 1–5.

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.

Zheng, C., Zhao, L., Zhou, X., Fu, Z., Li, A., 2013. Treatment technologies for organic wastewater. Water Treat. 11, 250–286.

Zhou, X., Du, H., Xue, S., Ma, Z., 2024. Recent advances in data mining and machine learning for enhanced building energy management. Energy 132636.