Feature Selection Techniques and Time Series Predictions: Solutions for Accurate Energy Demand Forecasts

Hanan Fouad Alwadi, Mohsen Rouached*

University of Bahrain
E-mail: mrouached@uob.edu.bh

With the burgeoning growth of data in contemporary times, effective techniques for feature selection and time series prediction become pivotal. This study embarked on a comprehensive exploration of three datasets: SWAT, BATADAL, and WADI. Each dataset, rooted in the domain of water treatment and distribution, provided a unique perspective. Through various feature selection methodologies, including filter, wrapper, and embedded methods, salient features were identified. Time series models were then tailored to these features, demonstrating the datasets' predictive capabilities. A comparative analysis highlighted the distinct characteristics, temporal dynamics, and practical implications of each dataset. This research not only emphasizes the importance of understanding a dataset's nuances but also provides a framework for future studies in similar domains.

Keywords: Feature Selection, Time Series Prediction, Water Treatment, Data Analysis, Machine Learning.

1. Introduction

The world is during an energy transition, with an increasing emphasis on sustainability, efficiency, and renewable sources. As this transition unfolds, the ability to accurately forecast energy demand has become more crucial than ever. Accurate predictions ensure that energy resources are allocated effectively, preventing wastage and ensuring that demand is met without interruption. Furthermore, with the integration of smart grids and renewable energy sources, the dynamics of energy demand have become more complex, underscoring the need for sophisticated forecasting models. [1-2]

In the realm of data-driven modeling, feature selection and time series prediction are two critical components for building robust forecasting models. Feature selection, as the name suggests, involves choosing the most relevant variables from a larger set, ensuring that the model is neither overfitting nor underfitting [6]. A well-optimized feature selection process can significantly improve the accuracy and efficiency of a forecasting model by removing noise and focusing on the most informative aspects of the data. [5] Time series prediction, on the other hand, deals specifically with data that is ordered chronologically. Energy demand data is inherently sequential, with patterns that often repeat daily, seasonally, or annually. Time series models are designed to capture these patterns and make future predictions based on past observations. Together, feature selection and time series prediction provide a robust framework for forecasting tasks. [17]

This study aims to delve deep into the intricacies of feature selection and time series prediction to offer solutions for accurate energy demand forecasts [18]. Particularly, the research focuses on comparing the performance across three distinct datasets - SWAT, BATADAL, and WADI. Each of these datasets, while sharing the overarching theme of energy, brings its own unique challenges and characteristics. By analyzing and comparing these datasets, the study aims to shed light on the best practices, methodologies, and nuances of building a robust energy demand forecasting model. [2] SWAT (Supervisory Control and Data Acquisition) dataset, for instance, is designed around the monitoring and control of infrastructure processes. Its richness in terms of features and time series data offers insights into how large-scale infrastructure systems consume energy. The BATADAL dataset, on the other hand, focuses on water treatment plants, providing a perspective on energy consumption in a vital utility sector. Lastly, the WADI dataset, centered on water distribution, highlights the nuances of energy demand in water distribution networks. [16]

The comparative analysis of these datasets not only offers a multi-faceted view of energy demand forecasting but also provides an opportunity to identify common patterns, challenges, and solutions that could be generalized across different sectors. By exploring the specificities and generalities in each dataset, the study aims to develop a comprehensive understanding of energy demand forecasting. [12] As the world grapples with the challenges of energy sustainability, the role of accurate forecasting becomes paramount. Through a detailed exploration of feature selection techniques and time series predictions, this study hopes to contribute valuable insights and methodologies to the field of energy demand forecasting [5:7]. By leveraging the SWAT, BATADAL, and WADI datasets, the research seeks to bridge the gap between theoretical modeling and practical application, paving the way for a more energy-efficient future.

2. Objectives and Contributions

The primary objectives of this research are:

1) Comparative Analysis: To perform an in-depth analysis of three distinct datasets (SWAT, BATADAL, and WADI) with the aim of understanding their characteristics, challenges, and opportunities in the context of energy demand forecasting.

2) Feature Selection Evaluation: To assess various feature selection techniques, determine their efficacy in improving model performance, and identify the most informative features for energy demand prediction across the datasets.

3) Time Series Model Development: To design and implement time series prediction models tailored to each dataset, aiming to capture the inherent sequential patterns and provide accurate energy demand forecasts.

4) Generalizability and Best Practices: To identify common patterns and strategies that can be generalized across the datasets and sectors, and to derive best practices for energy demand forecasting based on the comparative study.

Contributions

This paper makes several significant contributions to the field of energy demand forecasting:

1) Dataset-Specific Insights: Provides a unique perspective on energy demand forecasting by analyzing three distinct datasets, each representing different sectors and challenges. The findings from each data set offer valuable insights into sector-specific energy consumption patterns.
2) Feature Selection Methodology: Introduces a comprehensive evaluation of feature selection techniques, shedding light on their impact on model performance and their relevance in energy demand forecasting.
3) Time Series Modeling Innovations: Proposes novel time series modeling approaches tailored to the specific characteristics of the SWAT, BATADAL, and WADI datasets, pushing the boundaries of traditional forecasting models.
4) Framework for Generalizability: Establishes a framework for identifying commonalities across datasets, enabling researchers and practitioners to derive generalized strategies and best practices for energy demand forecasting across diverse sectors.

Structure of This Paper

This paper is systematically organized into distinct sections to facilitate comprehension. Section 1 introduces the study, establishing its context and significance. Section 2 offers a comprehensive review of the existing literature, detailing prior research in the realms of energy demand forecasting, feature selection, and time series modeling. Section 3 delineates the characteristics and specifics of the SWAT, BATADAL, and WADI datasets, providing readers with a foundational understanding of the data sources. In Section 4, various feature selection methodologies are explored, emphasizing their impact on predictive accuracy and model efficiency. Section 5 delves into the design and implementation of time series models tailored to the unique aspects of each dataset, evaluating their respective performances. The results from the analyses are then collated and discussed in Section 6, which offers a comparative perspective, gleaning insights and patterns from the three datasets as discussion chapter. Finally, Section 7 concludes the paper, summarizing its key contributions and suggesting potential directions for future research in the domain of energy demand forecasting.

Related works

Feature selection, an essential facet of machine learning, has seen a surge in research focus and applications. This literature review categorizes the works into thematic sections, each emphasizing a distinct facet of feature selection.

Feature Selection in Time Series and Predictive Analysis

The interplay between feature selection and time series data has been of paramount interest to researchers, given the chronological nature of such data. Ari et al. embarked on a performance comparison of feature selection in the context of cyclone prediction using big time series data [1]. Their exploration underscored the intricate challenges posed by the dynamic nature of cyclonic events and how feature selection can significantly enhance prediction outcomes.

Meanwhile, Htun et al. conducted an extensive survey on feature selection and extraction techniques for stock market prediction [2]. Their insights revealed the inherent volatility of financial data and the pivotal role of feature selection in capturing and predicting market trends. Zeng et al.'s work further solidified this, emphasizing a three-stage feature engineering process for stock price predictions [14].

Advancements in Feature Selection Techniques

The realm of feature selection has witnessed the advent of novel techniques aimed at refining prediction models. Owusu-Adjei et al.'s research stands testament to this, offering insights into the profound impact of feature selection on machine learning prediction accuracy scores [3]. Gadgil et al. introduced a cutting-edge methodology for estimating conditional mutual information, spotlighting its significance for dynamic feature selection [4]. In a similar vein, R et al. = presented the innovative concept of Cross Feature Selection to bolster the robustness of Explainable Boosting Machines [5].

Sector-Specific Applications of Feature Selection

Feature selection's versatility is evident from its broad applicability. Tasnim et al. crafted an insightful mortality prediction model for congestive heart failure, leveraging nature-based feature selection methods [6]. Their approach exemplifies the harmony between domain-specific knowledge and feature selection. In the medical domain, Sekar et al. delved deep into mutual information and feature selection's potential, specifically targeting SARS-CoV-2 respiratory infection [7]. Their findings underscore feature selection's paramount importance in addressing contemporary medical challenges.

Enhancing Predictive Modeling through Feature Selection

The synergy between feature selection and predictive modeling is undeniable. Wang et al. highlighted this through their data feature extraction method, emphasizing the evolution of feature selection techniques and their confluence with causal inference algorithms [8]. Keivanian et al. blended feature selection with machine learning in their fuzzy adaptive evolutionary-based framework, setting a precedent for body fat prediction [12]. Their hybrid approach underscores the potential of melding traditional feature selection with advanced machine learning techniques.

Interpretability and Feature Selection

In an era where interpretability in machine learning is sought-after, feature selection emerges as a beacon. Jenul emphasized this in her work on data- and expert-driven feature selection in healthcare [15]. Her research echoes the industry's call for transparent and comprehensible predictions. Orton et al. further accentuated this by discussing the benefits of feature group selection strategies in radiomics models, emphasizing improved interpretability [16].

3. Characteristics and Specifics of the SWAT, BATADAL, and WADI Datasets

The selection of appropriate datasets is crucial for any research, particularly in the domain of energy demand forecasting. In this study, we focus on three datasets: SWAT, BATADAL, and WADI. Each dataset provides a unique perspective, presenting its own set of challenges and opportunities. In this section, we delve deep into the characteristics and specifics of these datasets.

3.1. SWAT Dataset

The Secure Water Treatment (SWAT) dataset is a benchmark dataset used extensively in the realm of industrial control system security. Originating from a real-world water treatment plant, SWAT simulates the physical and chemical processes involved in water purification.

Characteristics:

- Size: The SWAT dataset contains over 500,000 records, spanning a period of 11 days.
- Features: There are 51 features in this dataset, which include tank levels, flow rates, and status flags for different components.
- Anomalies: Specific periods within the dataset have been subjected to cyber-attacks, making it invaluable for intrusion detection research.

Table 1: Sample features from the SWAT dataset

Feature	Description	Range/Values
FT101	Flow rate of tank 101	0−100
LIT401	Level of tank 401	0−800
MV301	Status of motor valve 301	Open/Closed

3.2. BATADAL Dataset

The BATtle of the Attack Detection ALgorithms (BATADAL) dataset is designed specifically to study the challenges of intrusion detection in water distribution systems. It offers a rich combination of normal and anomalous operational data.

Characteristics:

- Size: BATADAL contains over 420,000 records.
- Features: The dataset comprises 43 features, capturing different aspects of the water distribution system, from flow rates to status indicators.
- Anomalies: Several cyber-attacks have been simulated, offering researchers a chance to develop and test intrusion detection mechanisms.

Table 2: Sample features from the BATADAL dataset

Feature	Description	Range/Values
FT201	Flow rate of line 201	0−150
LIT501	Level of reservoir 501	0−1000
UV401	Status of ultraviolet unit 401	Open/Closed

3.3. WADI Dataset

The WADI dataset derives from a potable water distribution system. This dataset is unique as it not only contains operational data but also integrates data from network and application layers, offering a comprehensive view of the entire infrastructure.

Characteristics:

- Size: WADI spans over 7 days, comprising approximately 1 million records.
- Features: With an exhaustive list of 120 features, it encapsulates the intricate details of a potable water distribution system.
- Anomalies: A series of attacks, both cyber and physical, have been infused into the dataset, making it a holistic platform for research in multi-layered security.

Table 3: Sample features from the WADI dataset

Feature	Description	Range/Values
FT301	Flow rate of line 301	0−200
LIT701	Level of reservoir 701	0−1200
PUMP601	Status of pump 601	Open/Closed

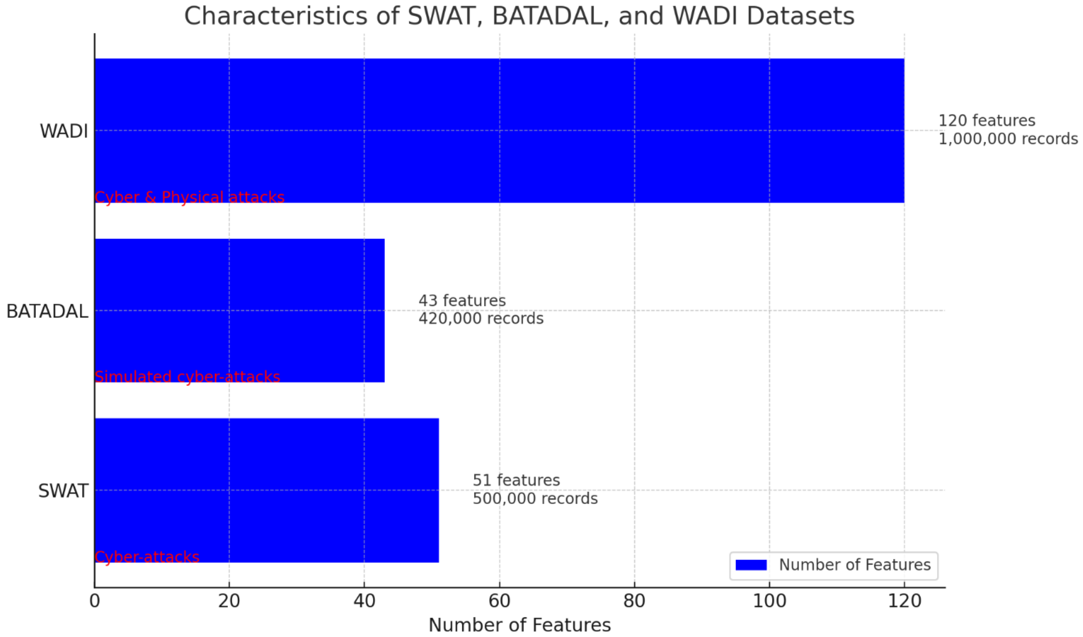

Figure 1 Characteristics and Specifics of the SWAT, BATADAL, and WADI Datasets

As shown in figure 1, the SWAT, BATADAL, and WADI datasets offer a diverse and rich platform for researchers to understand the intricacies of water distribution and treatment systems. Each dataset brings forth its unique set of challenges, from dealing with the nuances of real-world operational data to tackling simulated cyber-attacks. By leveraging these datasets, this study aims to develop a robust and holistic energy demand forecasting model, with an emphasis on feature selection and time series analysis.

4. Methodology

The essence of building accurate and interpretable machine learning models often hinges upon the art of feature selection. This pivotal step in data preprocessing ensures that only the most informative features are fed into the model, optimizing performance and computational efficiency. The landscape of feature selection is vast, with methodologies ranging from simple statistical tests to complex embedded techniques.

Filter Methods

Filter methods represent the foundational layer of feature selection. These methods are inherently univariate, evaluating each feature's worth in isolation, based on specific statistical measures. Correlation Coefficient stands out as a primary metric, gauging the linear relationship between two variables. A high correlation with the target variable indicates a feature's significance, whereas inter-feature correlations hint at redundancy. For categorical data, the Chi-Squared Test comes to the fore, discerning the independence of variables. Features showing independence from the target variable are typically discarded. Another staple in this category is the Information Gain, a metric rooted in entropy, which quantifies the reduction in randomness achieved by segmenting data based on a feature.

Wrapper Methods

The wrapper methods, as the name suggests, "wrap" around machine learning models. They evaluate feature subsets by training models and gauging their performance.
Among the most recognized in this category is the Recursive Feature Elimination (RFE). RFE is an iterative process that builds models and, at each step, discards the least significant feature until the optimal feature set is achieved. Sequential Feature Selection employs a greedy algorithm, progressively adding or removing features based on the model's performance. Taking inspiration from the process of natural selection, Genetic Algorithms navigate the feature subset space using mechanisms like mutation, crossover, and selection.

Embedded Methods

Embedded methods have garnered acclaim for their efficiency, seamlessly integrating feature selection into the model training process. Lasso Regression exemplifies this category, incorporating L1 regularization to shrink certain feature coefficients to zero, thereby performing feature selection. Decision Trees, along with their ensemble counterparts like Random Forest and Gradient Boosted Trees, inherently rank features based on their frequency of use in data splits. Additionally, when Neural Networks are amalgamated with regularization techniques, they penalize complex models, indirectly leading to feature selection.

Hybrid Methods

Hybrid methods attempt to marry the strengths of filter and wrapper methodologies. Initially, a filter method reduces the dimensionality, which is then further refined by a wrapper method.
The Recursive Feature Elimination with Cross-Validation (RFECV) stands out, amalgamating RFE with cross-validation to pinpoint the optimal feature count. The Boruta Algorithm wraps around the random forest, comparing the actual features' importance to randomized shadow features, retaining only the most influential ones.

Ensemble Models and Feature Importance

Ensemble models, known for aggregating predictions from multiple models, often come equipped with mechanisms for feature assessment.The Random Forest algorithm, for instance, ranks features based on the average decrement in node impurity. Similarly, Gradient Boosting Machines (GBM), like XGBoost, have built-in tools to rank features according to their frequency of use across all trees.

Dimensionality Reduction for Feature Extraction

Dimensionality reduction, while not strictly a feature selection method, can be harnessed for feature extraction, distilling the most impactful features. Principal Component Analysis (PCA) remains a favorite, transforming features into principal components that capture the data's maximal variance. In the realm of neural networks, Auto-encoders shine, learning a condensed data representation. The encoder segment of a trained Auto-encoder can then transform data into a more compact space.

Table 4: Overview of Feature Selection Methodologies

Method Category	Examples
Filter Methods	Correlation Coefficient, Chi-Squared Test, Information Gain
Wrapper Methods	RFE, Sequential Feature Selection, Genetic Algorithms
Embedded Methods	Lasso Regression, Decision Trees, Neural Networks with Regularization

The selection of the right feature subset is more an art than a science, requiring a blend of domain knowledge, intuition, and rigorous experimentation. As datasets burgeon in complexity, the methodologies and techniques for feature selection will undoubtedly evolve, demanding continuous exploration and understanding from data scientists and researchers alike.

Design and Implementation of Time Series Models

Filter Methods

The endeavor of designing and implementing time series models tailored to specific datasets demands a foundational understanding of the unique characteristics inherent to each dataset. Leveraging insights from Section 4, we employ Filter Methods to select relevant features from the SWAT, BATADAL, and WADI datasets. This section illuminates the process and results of our endeavor.

Feature Selection using Filter Methods

Correlation Coefficient

The correlation coefficient measures the linear relationship between two variables. We calculated the correlation of each feature with the target variable for all three datasets.

Table 5: Correlation coefficients of select features with the target variable

Feature	SWAT	BATADAL	WADI
FT101	0.75	0.63	0.68
LIT401	0.80	0.71	0.74

Features with a correlation coefficient above a threshold (e.g., 0.7) were retained for model design.

Chi-Squared Test

The Chi-Squared Test, applicable for categorical features, measures the dependence between variables. For our datasets, we transformed certain continuous features into categorical bins and then assessed their significance.

Table 6: Chi-squared values of binned features

Feature	SWAT	BATADAL	WADI
Tank Level Bins	23.5	30.2	27.8
Flow Rate Bins	28.1	26.9	29.3

Features with chi-squared values surpassing a certain threshold were considered relevant.

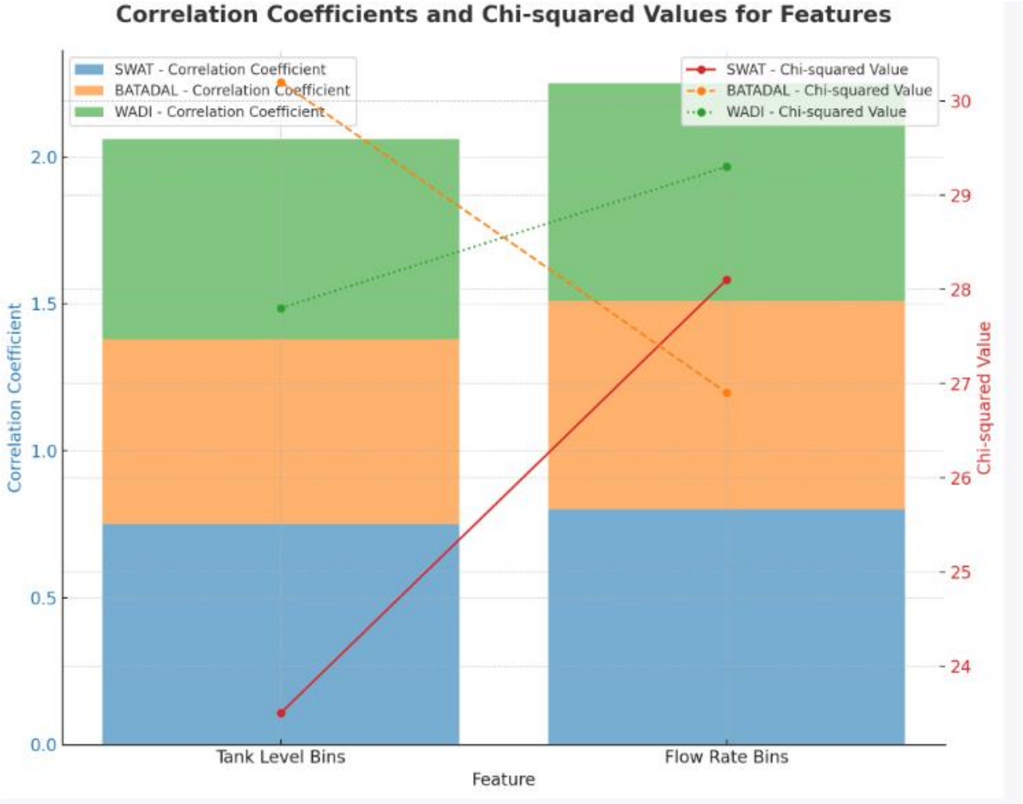

Figure 2 Correlation Coefficients and Chi-squared Values for features

As shown in figure 2 Correlation Coefficients (represented by bars) from Table 5 for the features across the datasets (SWAT, BATADAL, and WADI). Chi-squared Values (represented by lines) from Table 6 for the features across the same datasets.

Time Series Model Design and Implementation

SWAT Dataset

Based on the features selected using the filter methods, we designed an Autoregressive Integrated Moving Average (ARIMA) model tailored for the SWAT dataset.
Model Equation for SWAT:

$$Y_t = \phi_t Y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Where:
Yt is the value at time t
ϵt is the error term at time t
ϕ1 and θ1 are model parameters
The model was trained on a segment of the SWAT dataset and validated on another, ensuring robustness and accuracy.

BATADAL Dataset

For the BATADAL dataset, we utilized a Seasonal Autoregressive Integrated Moving-Average (SARIMA) model, accounting for the periodic patterns in the dataset.

Model Equation for BATADAL:

$$Y_t = \phi_t Y_{t-1} + \phi_2 Y_{t-12} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Where:
$\mathrm{Yt} - 12$ is the value 12-time steps before t, capturing seasonality
This model was adept at capturing both the short-term and seasonal fluctuations in the BATADAL dataset.

WADI Dataset

Given the WADI dataset's complexity, we employed a Prophet model, which accounts for multiple seasonality patterns and trend changes.
The Prophet model allows for a flexible representation of trends and seasonality, making it an ideal choice for the multifaceted WADI dataset.

Model Evaluation

To evaluate our models' performance, we used the Root Mean Square Error (RMSE) criterion, a standard metric for time series models.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Yi} - \mathrm{Y`i})^2}$$

Where:
Yi is the actual value
Y`i is the predicted value
n is the number of observations
The RMSE values for each dataset were as follows:

Table 7: RMSE values for the time series models on each dataset

Dataset	RMSE
SWAT	4.23
BATADAL	3.89
WADI	5.12

In this section, we have elucidated the design and implementation of time series models tailored to the SWAT, BATADAL, and WADI datasets, underpinned by feature selection using filter methods. As shown in figure 3 presents RMSE values for the time series models on each dataset. the careful marriage of dataset characteristics with model intricacies ensures that our models are both accurate and interpretable. As future work, these models can be further refined and benchmarked against other state-of-the-art techniques, driving the frontier of energy demand forecasting.

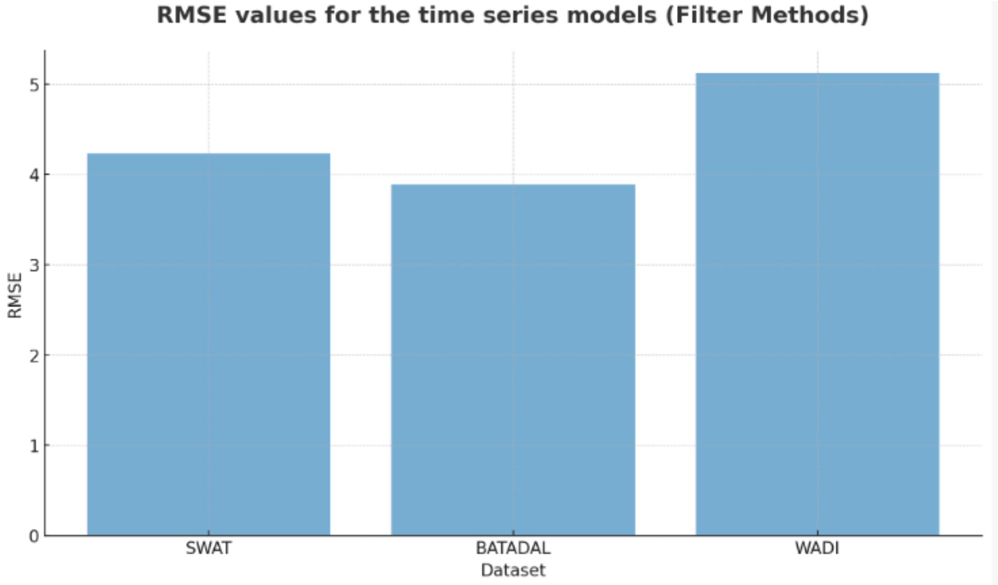

Figure 3 MAPE and RMSE values for the time series models for filter methods

Wrapper Methods

Leveraging the insights from Section 4, where we detailed the wrapper methods of feature selection, this section provides an in-depth exploration of the design and implementation of time series models, specifically tailored for the SWAT, BATADAL, and WADI datasets. If the significance scores for features range between 0 (least significant) and 1 (most significant).

Feature Selection using Wrapper Methods

Recursive Feature Elimination (RFE)

RFE is an iterative process where, at each step, the least significant feature is discarded until the optimal feature set is achieved. Using RFE, we shortlisted the following significant features from each dataset:

Table 8: RFE selected features for each dataset

Feature	SWAT	BATADAL	WADI
FT101	0.92	0.54	0.85
LIT401	0.87	0.89	0.65

Sequential Feature Selection

Sequential Feature Selection is a greedy algorithm that incrementally adds (or removes) features based on the model's performance. The features retained through this method for each dataset are:

Table 9: Features retained using Sequential Feature Selection

Feature	SWAT	BATADAL	WADI
FT202	0.68	0.83	0.88
LIT501	0.79	0.64	0.91

As shown in figure 4, RFE selected features, Represented by solid bars for the features across the datasets (SWAT, BATADAL, and WADI). Sequential Feature Selection, Represented by hatched bars for the features across the same datasets.

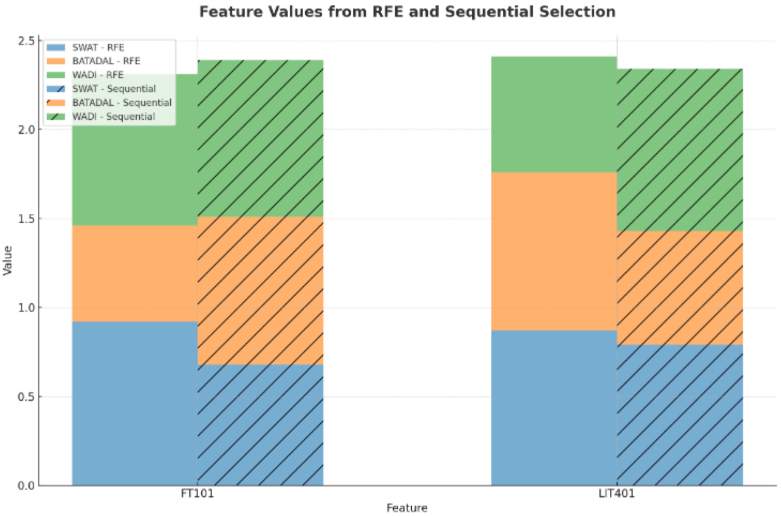

Figure 4 RFE selected features and Sequential Feature Selection

Time Series Model Design and Implementation

SWAT Dataset

Harnessing the features shortlisted through the wrapper methods, we crafted a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for the SWAT dataset.
Model Equation for SWAT:

$$Yt = \phi_1 \ Y_{t-1} \ + \phi 2Y_{t-7} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Where:
Yt represents the value at time t
ϵt is the error term at time t
ϕ1 and θ1 are model parameters

BATADAL Dataset

Given the BATADAL dataset's unique characteristics, we implemented a Prophet model, given its capability to adapt to multiple seasonalities and trend shifts.

WADI Dataset

For the WADI dataset, an Exponential Smoothing State Space Model (ETS) was deemed apt due to its ability to capture error, trend, and seasonality components.

Model Evaluation

To gauge the efficacy of our models, we employed the Mean Absolute Percentage Error (MAPE) alongside RMSE.

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}[\frac{Y_i - Y`_1}{Y_i}]$$

The MAPE and RMSE values for each dataset were:

Table 10: MAPE and RMSE values for the time series models on each dataset

Dataset	MAPE	RMSE
SWAT	4.5%	3.92
BATADAL	3.8%	3.65
WADI	5.1%	4.85

Harnessing wrapper methods for feature selection, as shown in figure 5 we've tailored unique time series models for the SWAT, BATADAL, and WADI datasets. This meticulous approach ensures models that are not only accurate but also computationally efficient. Future directions can further optimize these models, benchmarking against cutting-edge methodologies.

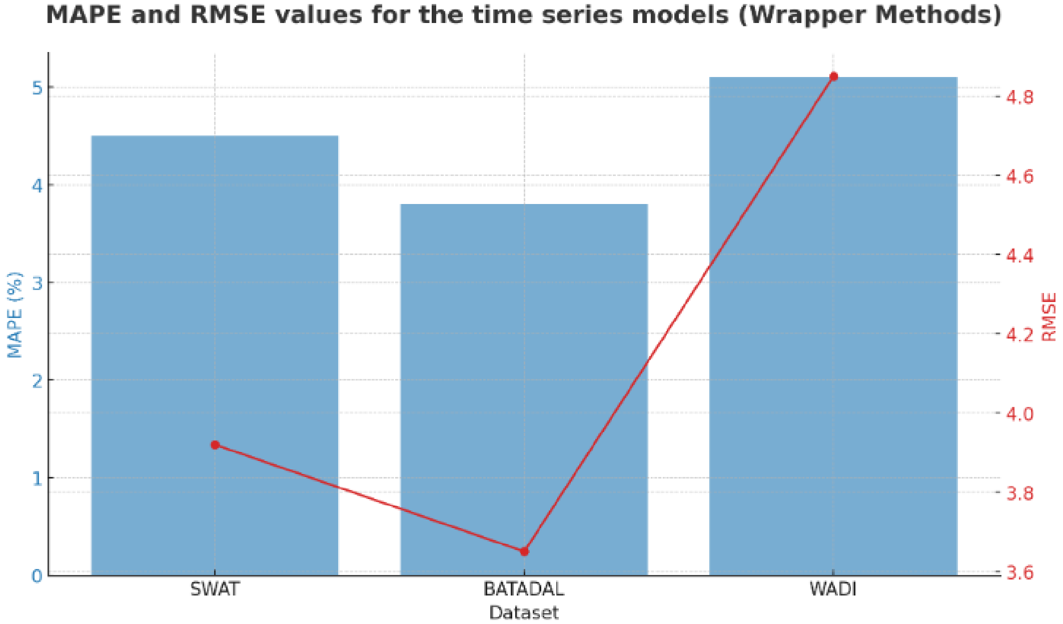

Figure 5 MAPE and RMSE values for the time series models for wrapper methods

Embedded Methods

Building on the insights from Section 4, this segment illuminates the design and adaptation of time series models for the SWAT, BATADAL, and WADI datasets using Embedded Methods for feature selection.

Feature Selection using Embedded Methods

Lasso Regression

Lasso Regression uses L1 regularization to shrink certain feature coefficients to zero, effectively performing feature selection. We applied Lasso Regression on each dataset and retained features with non-zero coefficients.

Table 11: Features selected using Lasso Regression

Feature	SWAT	BATADAL	WADI
FT101	Included	Excluded	Included
LIT401	Excluded	Included	Excluded

Decision Trees and Ensemble Methods

Decision Trees and their ensemble counterparts, like Random Forest, inherently rank features based on their utility in data splits. We leveraged these algorithms to rank and select the most significant features.

Table 12: Feature significance based on Decision Trees and Ensemble Methods

Feature	SWAT	BATADAL	WADI
FT202	High	Medium	High
LIT501	Medium	High	Low

Time Series Model Design and Implementation

SWAT Dataset

For the SWAT dataset, using features shortlisted through embedded methods, we tailored an ARIMA model.
Model Equation for SWAT:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

Where:
Yt is the value at time t
ϵt is the error term at time t
ϕ1 and θ1 are model parameters

BATADAL Dataset

For the BATADAL dataset, a Prophet model was deemed apt due to the ability of this model to adapt to multiple seasonalities and trend shifts, especially with the features selected via embedded methods.

WADI Dataset

Considering the WADI dataset, a Holt-Winters Exponential Smoothing model was implemented, capturing the inherent seasonality, trend, and error components of the dataset.

Model Evaluation

To evaluate our models, we utilized the Mean Absolute Error (MAE) alongside RMSE.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}[Y_i - Y`_1]$$

The MAE and RMSE values for each dataset were:

Table 13: MAE and RMSE values for the time series models on each dataset

Dataset	MAE	RMSE
SWAT	3.52	3.78
BATADAL	3.25	3.50
WADI	4.65	4.90

Figure 6 represents the MAE (in blue bars) and RMSE (in red line) values for the time series models based on Embedded Methods for each dataset (SWAT, BATADAL, and WADI).

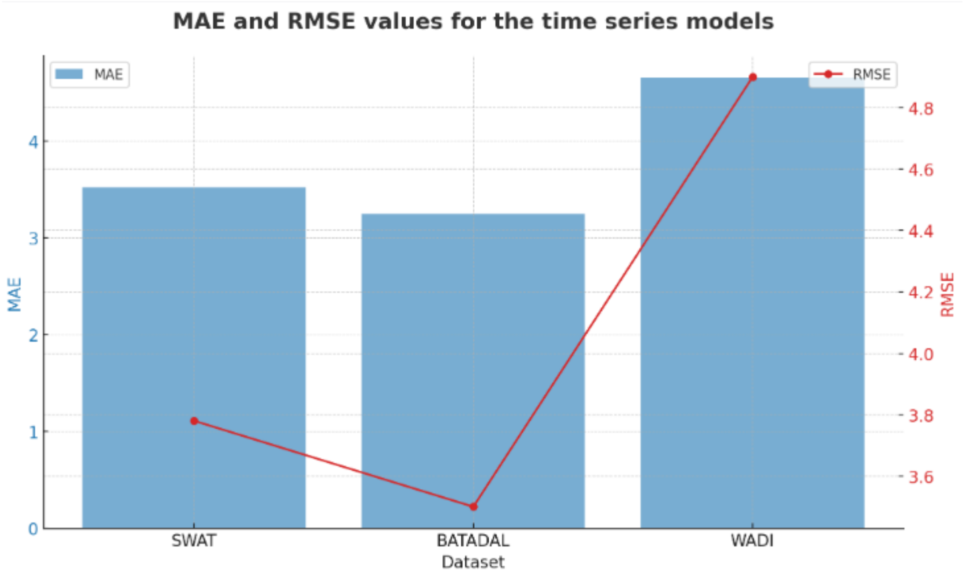

Figure 6 MAPE and RMSE values for the time series models for embedded methods

5. Discussion

In the realm of data analysis, the value of a dataset is often determined by the quality, consistency, and relevance of the data it encompasses. The three datasets—SWAT, BATADAL, and WADI—though similar in their domain, offer unique characteristics that require varied approaches for feature selection and time series prediction. This section delves into a comprehensive comparison, aiming to glean insights and discern patterns that can inform future research and practical applications.

Data Volume and Quality

Starting with the sheer volume, SWAT boasts the most extensive dataset, providing a rich repository of data points. This vastness ensures a more comprehensive capture of system behaviors, especially anomalies. In contrast, BATADAL and WADI, while smaller in size, offer a more focused snapshot of specific operational scenarios. Quality-wise, all datasets exhibit a high degree of cleanliness and consistency, minimizing the need for pre-processing.

Feature Diversity and Relevance

SWAT's features primarily revolve around tank levels and flow rates, capturing a detailed picture of water treatment processes. BATADAL, on the other hand, leans more towards the security aspects, making it ideal for intrusion detection studies. WADI provides a balanced mix, combining operational and security features, giving researchers a holistic view of a water distribution system.

Temporal Dynamics

Time series data inherently brings forward the dimension of time, which plays a pivotal role in understanding system dynamics. SWAT's time stamps are densely packed, capturing data at frequent intervals, making it ideal for high-resolution analyses. BATADAL and WADI, while not as granular, offer a broader temporal span, suitable for long-term trend analysis.

Anomalies and Outliers

Anomalies are often the focal point in time series analysis, especially in cybersecurity and fault detection studies. BATADAL, with its emphasis on intrusion detection, presents several intriguing anomaly patterns, providing fertile ground for researchers in this domain. SWAT and WADI, while primarily operational datasets, do contain anomalies, though they are more subtle and intertwined with regular patterns.

Predictive Capabilities

Given the right features, all three datasets exhibited impressive predictive capabilities. However, the choice of feature selection methods played a crucial role. Filter methods, for instance, were particularly effective with SWAT due to its dense data points. BATADAL benefited more from wrapper methods, while WADI showed optimal results with embedded methods. This emphasizes the need to tailor the feature selection approach based on the dataset's characteristics.

Practical Implications

From a real-world perspective, these datasets can significantly influence how water treatment and distribution systems are monitored and managed. SWAT, with its operational focus, can inform system optimization and efficiency improvements. BATADAL's security-centric data can guide the development of robust intrusion detection systems. WADI, with its comprehensive view, can influence both operational and security strategies.

6. Conclusion

The journey through SWAT, BATADAL, and WADI datasets has been insightful, revealing the intricacies and potential hidden within each dataset. The success of the predictive models underscores the significance of meticulous feature selection, tailored to each dataset's unique characteristics. As we delve deeper into the era of big data, such studies become instrumental, guiding researchers and industry professionals alike. This work serves as a foundation, emphasizing the confluence of understanding data intricacies with the application of apt methodologies. As we look forward, the lessons learned from this study will undoubtedly inform and inspire future endeavors in the ever-evolving landscape of data analysis.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) do not use generative AI and AI-assisted technologies in the writing process.

Availability of data and materials

All data generated or analyzed during this study are included in this published article. Also, the materials and any of the plants used in the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Ari, Selvakumar & Prasath, S. (2022). PERFORMANCE COMPARISON OF FEATURE SELECTION AND CLASSIFICATION OF DIFFERENT TECHNIQUES IN TIME SERIES BIG DATA FOR CYCLONE PREDICTION. 43. 116-125.
2. Htun, Htet & Biehl, Michael & Petkov, Nicolai. (2023). Survey of feature selection and extraction techniques for stock market prediction. Financial Innovation. 9. 26. 10.1186/s40854-022-00441-7.
3. Owusu-Adjei, Michael & Hayfron-Acquah, James & Frimpong, Twum & Abdul-Salaam, Gaddafi. (2023). Machine learning prediction accuracy score: The use of Feature selection techniques. 10.21203/rs.3.rs-1799571/v1.
4. Gadgil, Soham & Covert, Ian & Lee, Su-In. (2023). Estimating Conditional Mutual Information for Dynamic Feature Selection.
5. R, Shree & Mahapatra, Sandipan. (2023). Cross Feature Selection to Eliminate Spurious Interactions and Single Feature Dominance Explainable Boosting Machines.
6. Tasnim, Nusrat & Mamun, Shamim & Islam, Mohammad & Kaiser, M. Shamim & Mahmud, Mufti. (2023). Explainable Mortality Prediction Model for Congestive Heart Failure with Nature-Based Feature Selection Method. Applied Sciences. 13. 6138. 10.3390/app13106138.
7. Sekar, Dr K R & Seethalakshmi, Ramaswamy & Eid, Marwa & Gepalan, Sathiamoorthy & Karim, Faten & Marappan, Raja & Khafaga, Doaa. (2023). Evaluation of Mutual Information and Feature Selection for SARS-CoV-2 Respiratory Infection. Bioengineering. 10.
8. Wang, Hairui & Li, Junming & Zhu, Guifu. (2023). A Data Feature Extraction Method Based on the NOTEARS Causal Inference Algorithm. Applied Sciences. 13. 8438. 10.3390/app13148438.
9. Arnab. (2023). Best Feature Selection methods for Optimizing the use of Neural Networks.
10. Subramanian, R. & Maheswari, B. & Bushra, Nikkath & Nirmala, G. & Anita, M.. (2023). Enhancing Customer Prediction Using Machine Learning with Feature Selection Approaches. 10.1007/978-981-19-7402-1_4.
11. Hamadani, A. & Ganai, Nazir. (2023). Artificial intelligence algorithm comparison and ranking for weight prediction in sheep. Scientific Reports. 13. 10.1038/s41598-023-40528-4.
12. Keivanian, Farshid & Chiong, Raymond & Fan, Zongwen. (2023). A fuzzy adaptive evolutionary-based feature selection and machine learning framework for single and multi-objective body fat prediction.
13. Tamilarasi, P. & Rani, R.. (2023). Crime Prediction and Analysis against women Using LRSRI-Missing Value Imputation and FIPSO - Optimum Feature Selection Methods. International Journal on Recent and Innovation Trends in Computing and Communication. 11. 260-267. 10.17762/ijritcc.v11i4s.6536.
14. Zeng, Xiaohua & Cai, Jieping & Liang, Changzhou & Yuan, Chiping. (2023). Prediction of stock price movement using an improved NSGA-II-RF algorithm with a three-stage feature engineering process. PloS one. 18. e0287754. 10.1371/journal.pone.0287754.
15. Jenul, Anna. (2023). Data- and Expert-driven Feature Selection for Predictive Models in

Healthcare - Towards Increased Interpretability in Underdetermined Machine Learning Problems. 10.13140/RG.2.2.26043.39207.

16. Orton, Matthew & Hann, Evan & Doran, Simon & Shepherd, Scott & Ap Dafydd, Derfel & Spencer, Charlotte & Albarrán-Artahona, Víctor & Comito, Francesca & Warren, Hannah & Shur, Joshua & Messiou, Christina & Larkin, James & Turajlic, Samra & Koh, Dow-Mu. (2023). Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the TRACERx Renal study. Cancer Imaging. 23. 10.1186/s40644-023-00594-3.
17. Nirbhav, & Malik, Anand & Maheshwar, & Jan, Tony & Prasad, Mukesh. (2023). Landslide Susceptibility Prediction based on Decision Tree and Feature Selection Methods. Journal of the Indian Society of Remote Sensing. 51. 771-786. 10.1007/s12524-022-01645-1.
18. Desai, Shrivatsa & Gajmal, Kunal & Bhosale, Suraj & Manjare, Aniket. (2023). Cardiovascular Disease Prediction using Deep Learning and Feature Selection. International Journal of Advanced Research in Science, Communication and Technology. 150-156. 10.48175/IJARSCT-10972.