Condition-Based Maintenance of a Circulating Water System of a Canadian Nuclear Power Plant using Machine Learning and Statistical Tools

Chaitee Godbole[a*], Congjian Wang[a], Vivek Agarwal[a], Diego Mandelli[a], Mohammad Movassat[b], Brian Mori[b], Daniel Liang[b], Eddy Nur[b], Amir Birjandi[b], Bryan Lobo[b], Natalia Murcia Jacome[b]

[a]Idaho National Laboratory, Idaho Falls, USA
[b]Ontario Power Generation, Toronto, Canada

Abstract: Canada Deuterium Uranium pressurized-heavy-water reactors (PHWR) are a type of nuclear power plant that generate clean and reliable energy. The scope of this work is to automate data analysis methodologies to inform a condition-based maintenance strategy of a circulating water system (CWS) of a PHWR. The multiunit CWS provides a continuous supply of water to cool steam condensers, even during transient scenarios, thereby improving the thermal efficiency. This work aims to develop a machine learning (ML) based approach to detect anomalies in heterogeneous data of a CWS in a PHWR to help inform a predictive maintenance strategy. The heterogeneous data include textual and numeric time series data for a PHWR. Natural-language-processing (NLP)-based models are used to analyze textual data contained in work orders and operator logs and an event-timeseries correlation detection method is applied to assist anomalies diagnoses for CWS. An ML model Robust Linear Model (RLM) is also used to remove the seasonal variations in the system variable distributions based on distributions of environmental variables. A machine learning model, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), trained on both original data and data without any seasonal variations will then be used to detect if an anomaly exists. Thus, by moving to an automated methodology to detect, classify, and forecast anomalies, the maintenance strategy would be based on component condition instead of a time-based schedule.

Keywords: Circulating Water System, Density-Based Spatial Clustering of Applications with Noise, Robust Linear Model, Natural Language Processing

1. INTRODUCTION

It is highly crucial to move to clean and reliable carbon-free energy, and nuclear energy from nuclear power plants (NPPs) are a frontrunner to achieve this goal. A common type of nuclear power plant is a pressurized-heavy-water reactor (PHWR) that uses heavy water as moderator and natural uranium as the nuclear fuel due to the small thermal neutron absorption cross section of heavy water. Commercial heavy water reactors in Canada are termed as Canada Deuterium Uranium (CANDU) pressurized heavy water reactors (PHWR). It is essential to ensure that the functioning of the PHWR is safe and reliable while trying to reduce overall plant costs. There is thus a large shift toward digitizing the operations and maintenance of plant components to reduce maintenance costs without compromising the safety and reliability of the PHWR. Thus, it is essential to move to a condition-based maintenance strategy to not only reduce the overall costs but to avoid unnecessary maintenance.

Thus, the objective of this research is to automate data analysis through fault detection to help inform a condition-based predictive maintenance strategy of circulating water system (CWS) of a PHWR. This is achieved by utilizing machine learning (ML) methods to detect anomalies in heterogenous data provided by a multiunit PHWR along with natural language processing (NLP) for textual data analysis from work orders and operator logs. ML algorithms used in this work are Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect anomalies in plant CWS data. This trained DBSCAN algorithm can then be used to detect faults in any new data received from the PHWR. Seasonal variations in the system variables are also removed by building a statistical robust linear model (RLM) that captures the direct correlation of system variables with environmental variables. This correlation is then used to subtract out the seasonal variations in system variables. A recently developed technical language processing (TLP) approach [1] based on NLP is utilized to extract PHWR CWS equipment reliability information from PHWR textual data. Such extracted data can provide system engineers with insights into anomalous behaviors or degradation trends as well as the possible causes behind them to predict their direct consequences. Especially, correlating this extracted information with numerical data, such as sensor data and plant power outputs, can assist system engineers in

identifying the possible causes regarding anomalous behaviors in numerical data as well as quantifying structure, system, and component (SSC) health. For example, this correlation can help show if the anomalies were due to plant outages or due to a degradation or fault of a PHWR CWS. Thus, by moving to automated fault detection, the unnecessary maintenance of PHWR CWSs can be mitigated and maintenance can be based on the condition of the CWSs as compared to a time-based maintenance schedule, which can help reduce overall maintenance costs.

The research presented in [2] utilized a variety of a multilayered perceptron (a type of ML algorithm) to detect degradations in a three-phase induction motor. Another form of an ML algorithm is the radial basis function neural network which has been used as an anomaly detection method to identify cracks in gears [3] as well as identifying faults in induction motor bearings [4]. Previous work [5] [6] [7] used ML algorithms for fault detection and classification for pumps in circulating water system to move to a risk-informed, condition based predictive maintenance method to reduce operating costs while maintaining reliability and safety of commercial nuclear power plants. This work used XGBoost (eXtreme Gradient Boosting) [8] to detect faults in the circulating water system as well as to classify the type of fault and degradation mechanism. The diagnostic model was a binary classifier XGBoost to detect if the circulating water system was healthy or unhealthy while the prognostic model was a multiclass classifier XGBoost used to identify which fault occurred if system was detected unhealthy. This work also used topic analysis to extract important words/information including equipment condition from work orders using latent Dirichlet allocation. Natural language processing (NLP) was then used to classify equipment condition events of work orders as either related or unrelated to the degradation of the equipment by conducting text characterization using convolutional neural networks.

The research presented in this paper discusses the details of the PHWR CWS in section 2 while introduction on ML algorithms and statistical models and the methodology behind using them is explained in section 3. Section 4 explains the results achieved through DBSCAN, and NLP followed by conclusions and future work.

2. CIRCULATING WATER SYSTEM

The CWS is an essential system for the safe and reliable operation of PHWRs. It supplies cooling water from a heat sink to the steam turbine condensers. The cooling water returns to the heat sink after heat removal by condensers. The CWS provides a continuous supply of water to cool steam condensers, even during transient scenarios, thereby improving PHWR thermal efficiency. It works to cool the discharged heated water from the multiunit condensers to an acceptable temperature in the heat sink through a continuous supply of water from a water source under various load and transient conditions to ensure PHWR thermal efficiency. Major components of the CWS system are depicted in Figure 1. The CWS system has a continuous water supply filtered using multiple screens and traveling screens to prevent debris and unwanted items in the water. Screen wash pumps provide spray water to remove accumulated debris on the screens. The CWS also contains a vacuum priming system that removes any air from the CWS. An important component of the CWS is the CWS pumps that ensure there is sufficient head to supply the condenser and circulate water through the CWS. The water from the pumps then goes to condensers to remove heat from the steam collected from the turbines and convert it to condensate.

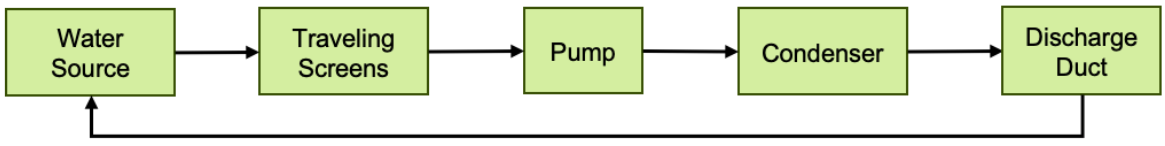


Figure 1. Major components of a CWS

3. MACHINE LEARNING AND STATISTICAL MODELING

ML has seen an exponential rise in use in a large variety of fields, including the nuclear industry, and has been used widely for real-time monitoring for predictive maintenance, thermal hydraulic computations, and nuclear design [9] [10]. ML algorithms have the capability to behave as a universal approximator for both linear and nonlinear relations when the physics may be unknown or complicated to model and have high computational speed even for large quantities of data. ML can be classified as either supervised, semisupervised, or unsupervised based on the model training regime. Supervised learning algorithms are used when the data

available has both input and output features, and the ML algorithm learns the relation between input and output. Unsupervised learning algorithms are used when there is no known target or output feature, and the ML algorithm works on learning structures, clusters or patterns in the input data. A semisupervised learning algorithm is a combination of both supervised and unsupervised learning algorithms where you have a lot of unlabeled data and a few labeled data.

This work uses DBSCAN, which belongs to the unsupervised ML category, to detect anomalies in data. DBSCAN is used on a large duration of data containing four system variables or data features. It detects anomalies and irregularities in the data by clustering all normal data points together and capturing and clustering all anomalies and irregularities as an anomaly cluster. The data used for DBSCAN is further broken down into smaller chunks to ensure optimal clustering by DBSCAN and to accurately capture all anomalies within the data without the need for hyperparameter tuning over each individual year. Seasonal variations in the system variables are also removed by building a statistical RLM that captures the direct correlation of system variables with environmental variables, which is captured by computing system variables as a function of environmental variables. This function is then used to subtract the seasonal variations in system variables by subtracting the function outputs with the real system variable values. The TLP approach developed by the coauthors [1] is used here to identify SSC entities and their corresponding health statuses, while a statistical testing approach based on maximum mean discrepancy (MMD) [11] is employed to correlate the extracted information with the numerical anomalies identified by DBSCAN.

3.1. DBSCAN

DBSCAN [12] [13] is a commonly used clustering algorithm belonging to the unsupervised ML category widely used for detecting outliers and anomalies in data [14]. It uses proximity parameters to cluster data points that are close together. The DBSCAN algorithm requires two parameters: epsilon, which specifies the distance between two points for them to be considered neighbors, and minimum points, which states that that there should be at least those set number of minimum points at epsilon to be considered part of cluster. Epsilon can be considered a radius for two-dimensional data, and minimum points then finds all the data points that have data points equal to minimum points within the radius epsilon and categorizes them together. Figure 2 shows the methodology of clustering by DBSCAN when minimum points is set at 4. All points belonging to a cluster are termed core points, shown by red dots in Figure 2. A core point is any point that has minimum data points at a radius epsilon around it. All data points belonging to a cluster and are around a core point but cannot themselves be termed a core point are called border points, depicted by yellow circles in Figure 2. All border points belong to the same cluster as the core point surrounding it at distance epsilon. Any data point not satisfying the minimum data points at distance epsilon criteria are then categorized as an anomaly data point by the DBSCAN algorithm, as detected by blue circle in Figure 2.

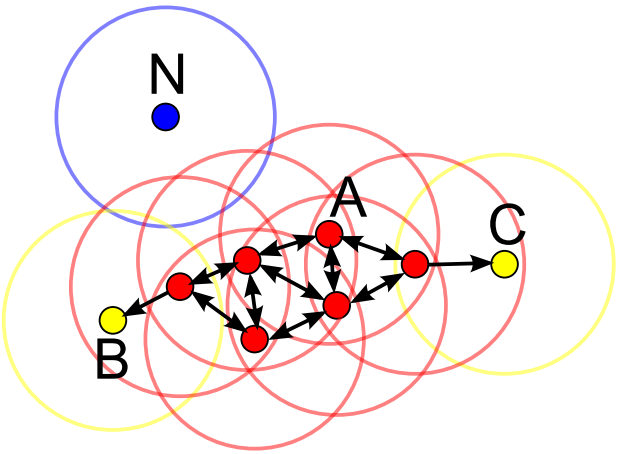


Figure 2. DBSCAN methodology for clustering data points [15]

This is how DBSCAN is utilized in this work to capture anomaly data points in system variables for the PHWR CWS. Epsilon and minimum data point parameters are manually tuned to capture the optimal values that work for data for a large duration of data. Data are split into smaller chunks to ensure the robustness and generalizability of DBSCAN without the requirement for hyperparameter tuning, and DBSCAN is applied to capture anomalies and normal data points.

3.2. RLM

Traditional regression models aim at finding a relationship between an independent variable and a dependent variable. Commonly used regression models include a linear regression model that aims to find the best hyperplane or line that accurately describes the linear relationship or function between the input variables and the output. A drawback of linear models is that they do not handle outliers well and can result in the model being biased towards those outliers. RLMs aim at overcoming this drawback by being able to handle outliers present in the data. This work uses an RLM with the Huber loss [16], which is an example of a robust regression algorithm that assigns less weight to observations identified as outliers. A Huber loss identifies outliers through the usage of residuals. If a data point is considered normal and not an outlier, the least square function is used to fit the linear model or else an absolute loss function is applied in the case of outlier data points.

This work uses a robust linear model with Huber loss to find the linear model that computes system variables as a function of environmental variables. The goal is to remove all the seasonal effects in system variable data distributions, which are easily captured in the environmental variable distributions using the RLM model to compute the function shown in Equation 1. Y_{system}^{RLM} indicates the system variables as computed by the RLM model that computes the linear function as a function of environment variables denoted by $X_{environment}$.

$$Y_{system}^{RLM} = RLM(X_{environment}) \tag{1}$$

Equation 1 thus computes system variables as a function of environment variables. To remove the seasonal variations in the system variables, Equation 2 is applied. Y'_{system} indicates the system variables without any seasonal variations and is computed by subtracting Y_{system}^{RLM}, which is the output from the RLM model that computes the relationship between system variables and environment variables from Y_{system}, which denotes the original values of system variables including seasonal variations.

$$Y'_{system} = Y_{system} - Y_{system}^{RLM} \tag{2}$$

3.3. NLP Textual Data Analysis

In this work, TLP is utilized to extract knowledge about SSC statuses. Figure 3 illustrates the steps to perform a TLP analysis with the information from textual data, such as work order reports and operator shift logs. The TLP methods developed in [1] could be applied to recognize SSCs (entities) of the CWS and identify their corresponding health statuses, the date of the event, and possible causes for the event. We then utilize an MMD-based statistical testing approach to correlate the extracted information with system responses for the anomaly diagnosis, such as existence of correlation, and temporal order. The statistical testing approach is first proposed in [17] and then extended by the coauthors to correlate numerical data and textual equipment reliability data in their recent report [11].

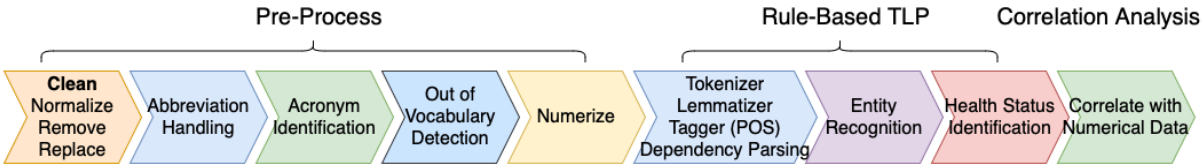


Figure 3. TLP analysis flow to extract the knowledge that may be possibly causes for anomalies

4. RESULTS

All variables in the results of this research work are anonymized to protect private/sensitive information. This is done by normalization of all numeric data and by denoting all variables as system variables X1, X2, X3 and X4 and environment variable. This section presents the results achieved from DBSCAN, RLM, and NLP. Section 4.1describes DBSCAN results when applied to chunks of data to capture the anomalies present in the data. Section 4.2 presents RLM results to remove the seasonal variations in the data. DBSCAN is then reapplied to the data without seasonality to show the robustness of the DBSCAN methodology. NLP is then applied to understand the correlation of numeric data with textual data and results are presented in section 4.3.

4.1. DBSCAN Results

DBSCAN is applied to capture anomalies present in four system variables of a CWS in a PHWR that are denoted by X1, X2, X3, and X4. DBSCAN consists of two hyperparameters, epsilon and minimum points. After manual hyperparameter tuning, the optimal values of epsilon and minimum points that work on various sections of the dataset containing four system variables are 1 and 2,000, respectively. Figure 4 shows the application of DBSCAN for a certain subset of data for all four system variables, X1, X2, X3, and X4, as shown in Figure 4 (a), (b), (c), and (d), respectively. The x-axis of Figure 3 depicts a variable of time. All the red data points are what the DBSCAN clusters as normal data points or core and border points. Blue data points are the outliers detected by DBSCAN. These anomalies match accurately with the true anomalies seen in the system variable data for the CWS as these anomalies have a clear trend that is out of the normal ranges of the system variable ranges depicting an outage or maintenance planned on the CWS.

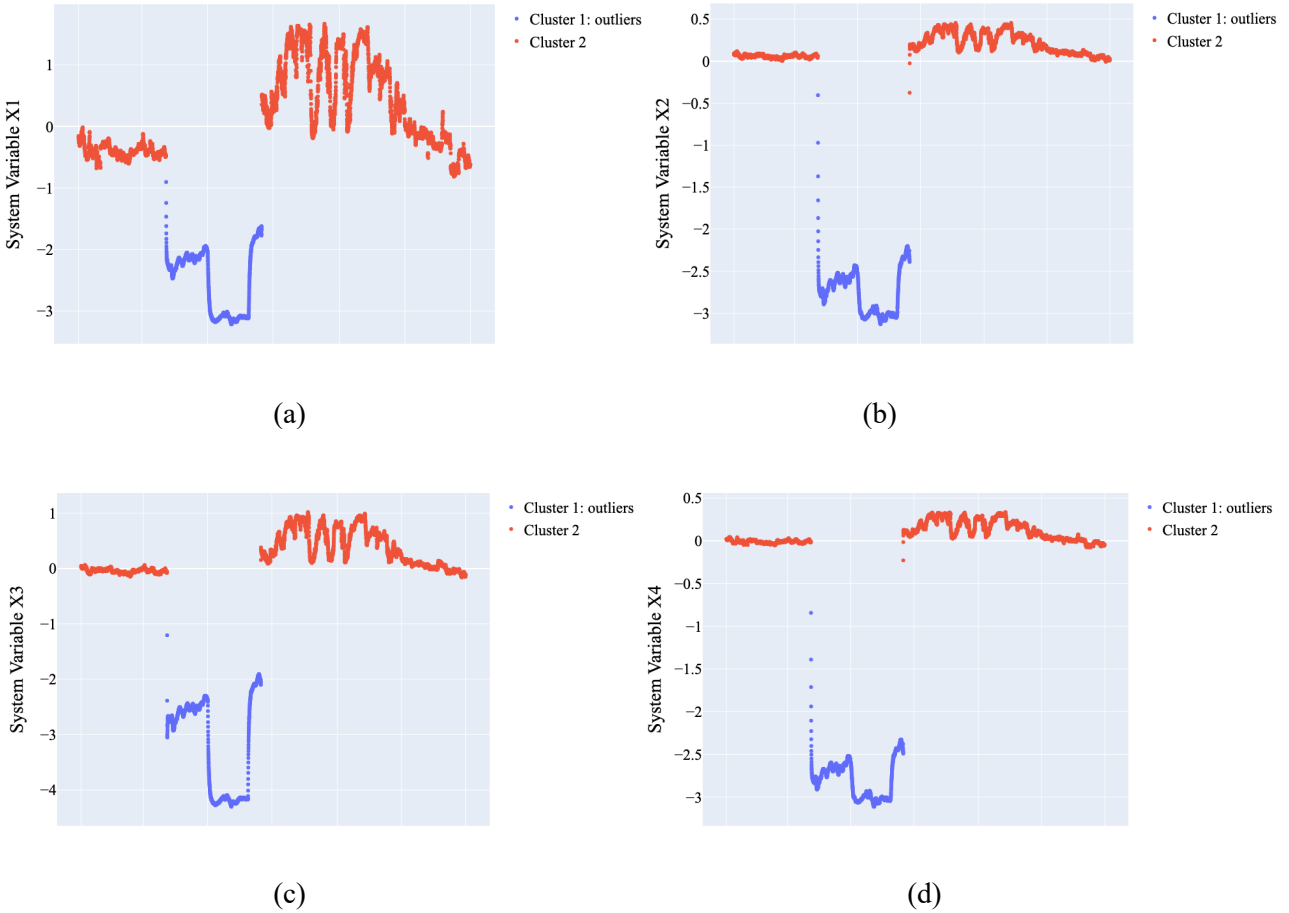


Figure 4. DBSCAN results for certain subset of the entire data set for (a) system variable X1, (b) system variable X2, (c) system variable X3, and (d) system variable X4

Similarly, Figure 5 shows the application of DBSCAN for a different subset of data for all four system variables, X1, X2, X3, and X4, as shown in Figure 5 (a), (b), (c) and (d), respectively. The x-axis of Figure 5 depicts a variable of time. All the red data points are what the DBSCAN clusters as normal data points or core and border points. Blue data points are the outliers detected by DBSCAN. These anomalies match accurately with the true anomalies seen in the system variable data for the CWS as can be seen through system textual information and can be seen visually as these anomalies have a clear trend that is out of the normal ranges and normal behavior/distribution of the system variable trends.

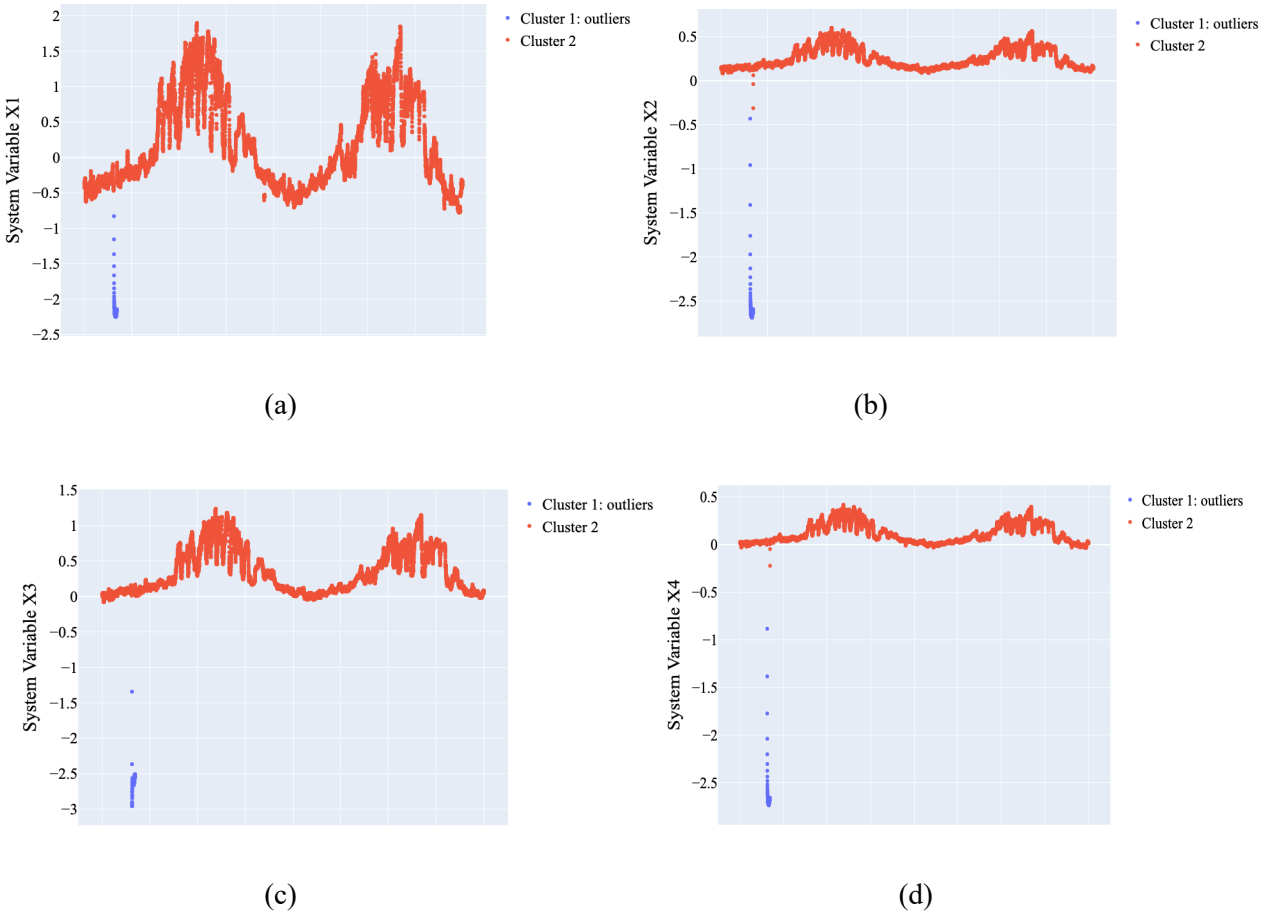


Figure 5. DBSCAN results for a different subset of the entire data set for (a) system variable X1, (b) system variable X2, (c) system variable X3, and (d) system variable X4

Thus, in both Figure 4 and Figure 5, when the trends of system variables saw anomalies depicting faults or maintenance activities, DBSCAN was accurately able to capture all the anomaly data points and label them as an anomaly.

4.2. RLM and DBSCAN Results

RLM is used to compute the relationship of system variables X1, X2, X3, and X4 as a function of the environment variable. Figure 6 shows the relationship of system variables with environment, which is linear except for a few regions with considerable deviations from the linear behavior. This deviation shows the existence of anomalies within the system variable data due to maintenance or outages as these deviations are not visible in the environment variable. Due to the linear variation of system variables with environment variables and due to the robustness of RLM to handle anomalies, RLM will successfully determine the relationship and negate these anomalies.

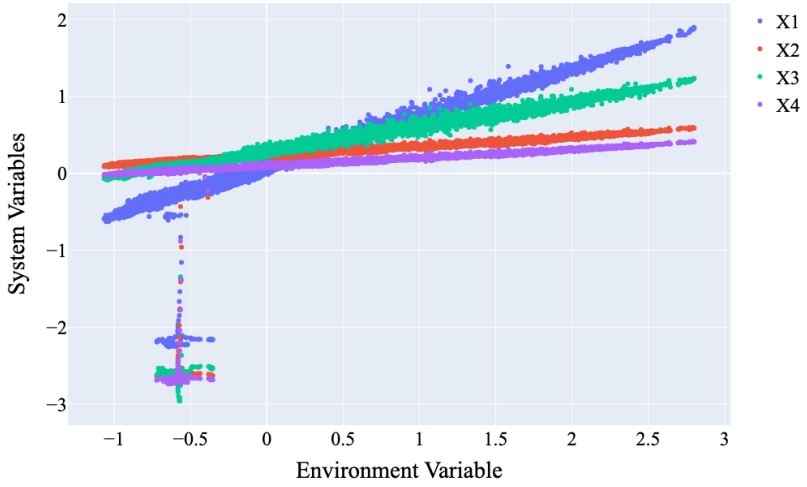


Figure 6. Relationship of system variables with the environemnt variable

Figure 7 shows the RLM prediction results as shown in Equation 1 along with the residual on the secondary y-axis as shown in Equation 2. The RLM prediction is shown by the red dotted line, and the residual prediction on the secondary y-axis is shown in green. This shows that the RLM results match accurately with the true variations of system variables and environment variables except for the anomalies. Thus the resudials are a good measure to capture the anomalies present in the data.

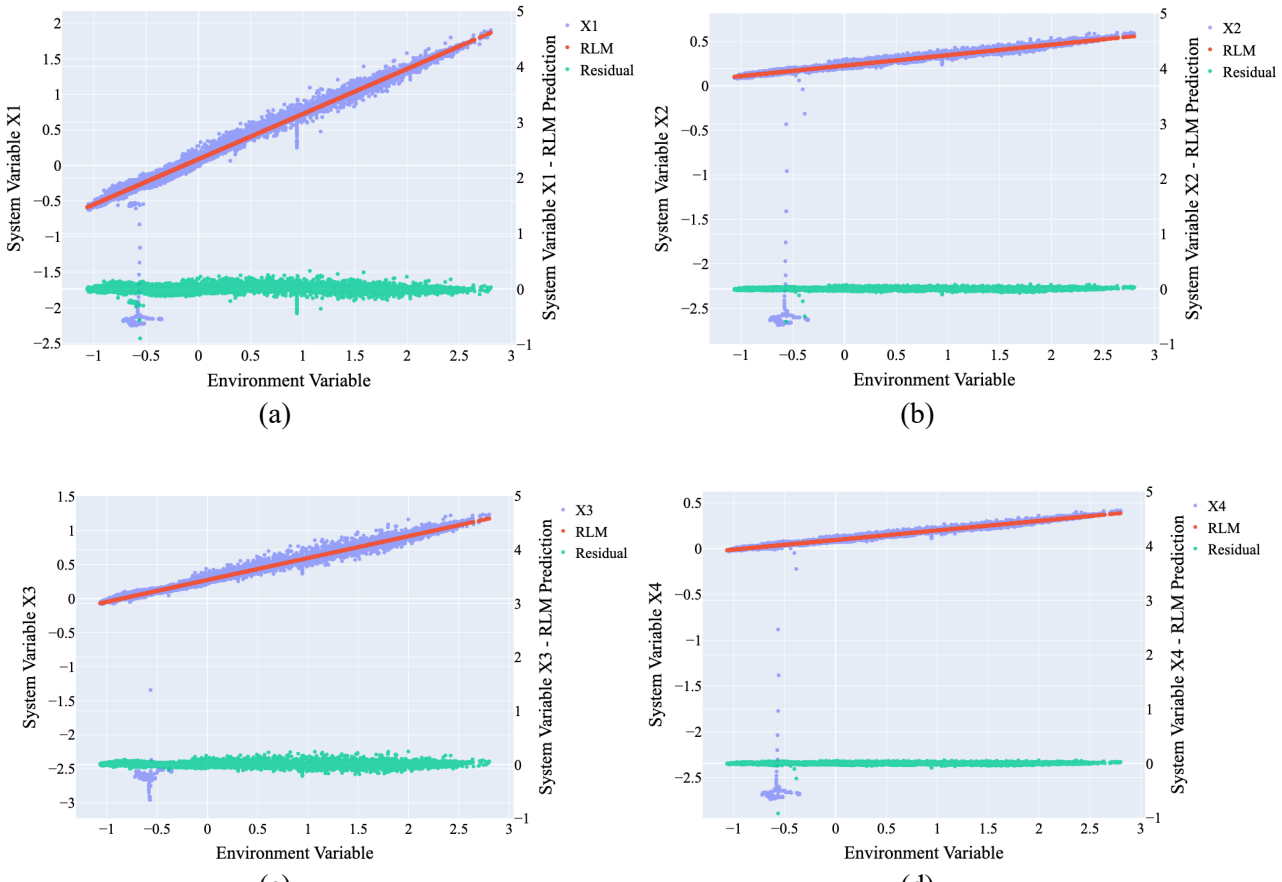


Figure 7. RLM predictions and residuals for system variables (a) X1, (b) X2, (c) X3, and (d) X4

DBSCAN is then applied on the residual values to capture the anomalies. The DBSCAN used for original data is used without any hyperparameter tuning needed with epsilon 1 and minimum points as 2,000. Figure 8 shows the DBSCAN results on the residuals computed using RLM for all of the four system variables.

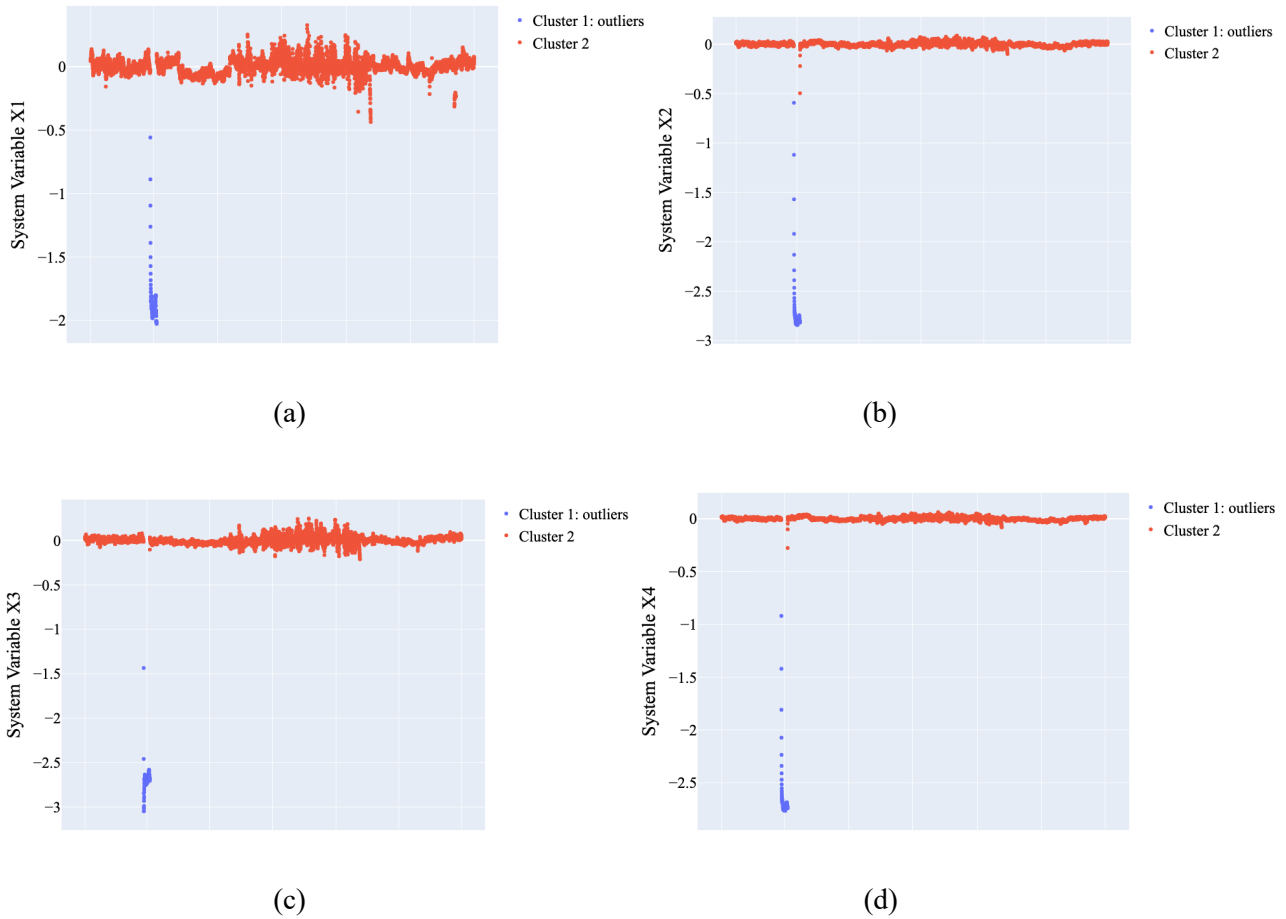
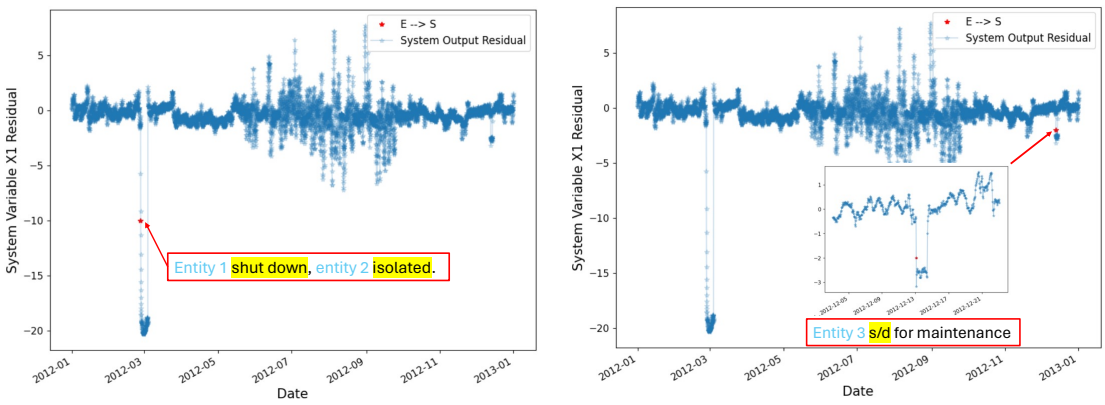


Figure 8. DBSCAN results for residuals computed by RLM to remove seasonal variations for system variables (a) X1, (b) X2, (c) X3, and (d) X4

Thus, DBSCAN was successfully able to predict the anomalies without the need for hyperparameter tuning for different datasets and different distributions, showing the robustness of the algorithm. DBSCAN was successfully able to capture and categorize all anomalies in the original dataset as well as the dataset with all seasonal variations removed based on environment variables which coincides with the anomalies seen in textual information on the CWS.

4.3. TLP and Event-Timeseries Correlation Results

As mentioned before, TLP is applied to PHWR work order reports and operator shift logs to identify event information. As illustrated in the red box in Figure 9, the SSC entities are identified and highlighted in blue, while their statuses are identified and highlighted in yellow. Through the MMD statistical testing, some of the identified events are correlated with the system variable output and their temporal orders are also determined. For example, E → S indicates the event is happening before the anomalies in the time series, and the sub-time series before this event is similar to the normal signal.



(a) (b)

Figure 9. Correlate events identified by TLP from shift logs with system variable X1, where E → S indicates the identified event may be the cause of the anomaly in the time series

5. CONCLUSIONS

There is an imperative move towards clean energy generation through NPPs. PHWRs are a promising source of such clean and reliable nuclear energy. Due to the rising need for digitizing the operation and maintenance of PHWR plant components to reduce maintenance costs, this work aims at moving to a condition-based maintenance strategy by using ML and statistical models to capture anomalies present in four system variables of a CWS in a PHWR. This is done by using DBSCAN as an unsupervised clustering algorithm to capture all anomalies present in the data. Seasonal variations are also removed from the system variables by using an RLM model to compute the system variables as a linear function of environment variables. The output from the RLM is then subtracted from the original system variable values to obtain new values of system variables that do not contain any seasonal variations. DBSCAN is then applied again to this data to show its robustness of DBSCAN on different data distributions. It was able to predict anomalies for both the original data and the dataset with all seasonal variations removed using RLM, thereby showing its robustness and its ability to handle multiple datasets. The TLP approach is able to extract events that happened in the textual reports, and the MMD based approach could correlate the identified events with anomalies in the system variables. This work thus successfully implements the usage of ML and statistical modeling to capture anomalies in the data of a CWS in a PHWR so as to inform a condition-based maintenance strategy.

6. FUTURE WORK

Future work consists of building a predictive model built on the predictions of the DBSCAN algorithm. All anomalies detected by DBSCAN are labeled as 0 and all other normal data points are labeled as 1. These labels then convert the data from an unsupervised category to a supervised category. ML models belonging to the supervised learning category can then be trained using the DBSCAN outputs and categories. This supervised algorithm can then behave as a predictive model that can detect anomalies in any new data. The TLP approach will be further enhanced with model-based system engineering models to identify the root causes for the anomaly behaviors.

Acknowledgements

This work was funded by the U.S. Department of Energy Office of Nuclear Energy's Light Water Reactor Sustainability Program. Authors would also like to thank editor Alexandria Madden of Idaho National Laboratory for technical editing of the paper.

References

[1] C. Wang, D. Mandelli and J. Cogliati, "Technical Language Processing of Nuclear Power Plants Equipment Reliability Data," *Energies,* vol. 17, no. 7, 2024.

[2] V. N. Ghate and S. V. Dudul, "Optinal MLP neural classifier for fault detection of three phase induction motor," *Expert Systems with Applciations,* vol. 37, 2010.

[3] H. Li, Y. Zhang and H. Zheng, "Gear fault detection and diagnosis under speed-up condition based on order cepstrum and radial basis function neural network," *Journal of Mechanical Science and Technology,* vol. 23, no. 10, pp. 2780-2789, 2009.

[4] Z. Jin, Q. Han, K. Zhang and Y. Zhang, "An Intelligent Fault Diagnosis Method of Rolling Bearings based on Welch Power Spectrum Transformation with Radial Basis Function Neural Network," *Journal of Vibration and Control,* vol. 26, pp. 629-642, 2020.

[5] V. Agarwal, A. Gribok, K. M. Araseethota, J. A. Smith, N. J. Lybeck, V. Yadav, M. Yarlett, B. Diggans and H. Palas, "Integrated Risk-Informed Condition Based Maintenance Capability and Automated Platform: Technical Report 3," Idaho National Laboratory (INL) , Idaho Falls, 2022.

[6] D. Mandelli, C. Wang, V. Agarwal, L. Lin and K. A. Manjunatha, "Reliability modeling in a predictive maintenance context: A margin-based approach," *Reliability Engineering & System Safety,* vol. 243, 2024.

[7] R. M. Spangler, V. Agarwal and D. G. Cole, "A Hybrid Reliability Model Using Generalized Renewal Processes for Predictive Maintenance in Nuclear Power PLant CIrculating Water Systems," *IEEE Access,* vol. 11, pp. 136726-136740, 2023.

[8] XGBoost Developers, "XGBoost Documentation," Sphinx, 2024. [Online]. Available: https://xgboost.readthedocs.io/en/stable/. [Accessed 2024].

[9] X. Zhao, K. Shivran, R. Salko and F. Guo, "On the Prediction of Critical Heat Flux using a Physics-Informed Machine Learning-Aided Framework," *Applied Thermal Engineering,* vol. 164, no. 114540, 2020.

[10] C. Godbole, G. Delipei, X. Wu, M. Avramova and U. Rohatgi, "Machine Learning-based Prediction of Departure from Nucleate Boiling Power for the PSBT Benchmark," in *Advances in Thermal Hydraulics (ATH 2022)*, 2022.

[11] D. Mandelli, C. Wang and V. Agarwal, "Development of Analysis Methods that Integrate Numeric and Textual Equipment Reliability Data," No. INL/RPT-23-74530-Rev000. Idaho National Laboratory (INL), Idaho Falls, ID (United States), 2023.

[12] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databased with noise," *kdd,* vol. 96, no. 34, pp. 226-231, 1996.

[13] T. M. Thang and J. Kim, "The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters," in *2011 International Conference on Information Science and Applications*, 2011.

[14] M. Çelik, F. Dadaşer-Çelik and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *International Symposium on Innovations in Intelligent Systems and Applications*, IStanbul, Turkey, 2011.

[15] B. C.-. O. Work, *CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17045963.*

[16] P. J. Huber, "Robust Statistics," John Wiley and Sons, Inc., Net York, 1981.

[17] C. Luo, J.-G. Lou, Q. Lin, Q. Fu, R. Ding, D. Zhang and Z. Wang, "Correlating events with time series for incident diagnosis," in *Proceedings of the 2-th ACM SIGKDD international conference on Knowledge discovery and data minind*, 2014.