P–ISSN: xxxx - xxxx | E-ISSN : 2963-4946
Vol. 3 No. 7 April 2025
ASEAN Journal of Social and Humanities
https://ajosh.org/

USE OF MACHINE LEARNING-BASED HEALTH INDEX WITH K-NEAREST NEIGHBORS METHOD TO MAINTAIN DESALINATION PLANT PERFORMANCE GAS AND STEAM POWER PLANTS APPLICATIONS

Udi Harmoko[1], Marcelinus Christwardana[2], Muhammad Rizkan[3]
[1,2] Faculty of Science and Mathematics, Universitas Diponegoro, Indonesia
[3]School of Postgraduate Studies, Universitas Diponegoro, Indonesia
Email: udiharmoko@lecturer.undip.ac.id, marcelinus@lecturer.undip.ac.id, muhammadrizkan@students.undip.ac.id

ARTICLE INFO

Keywords: Machine learning, K-Nearest Neighbors (K-NN), desalination plant, predictive maintenance, power plant efficiency.

ABSTRACT

This study presents the implementation of a Machine Learning-Based Health Index utilizing the K-Nearest Neighbors (K-NN) algorithm for predictive maintenance in desalination plants within gas and steam power plants. The research focuses on optimizing the maintenance schedule of the Block 3 Priok Desalination Plant, which is critical for providing high-quality distilled water for power generation. This study aims to develop and integrate a predictive maintenance framework into PLN's digitization system, allowing for automated monitoring and optimized servicing schedules. Unlike the previous application of K-NN in Block 4, which utilized five health indices for performance classification, Block 3 requires an expanded model incorporating at least seven input parameters due to its multi-effect desalination process. By refining the predictive model and increasing data parameterization, this study seeks to enhance maintenance accuracy, minimize operational downtime, and improve overall desalination efficiency. By leveraging historical operational data and real-time monitoring, the K-NN model predicts the health index of desalination components with 98% accuracy. Implementing this approach minimizes downtime, optimizes maintenance schedules, and enhances energy efficiency. The results demonstrate that AI-driven predictive maintenance significantly improves reliability, reduces costs, and supports energy sustainability goals.

Introduction

PT PLN Indonesia Power UBP Priok runs a 2,800 MW gas and steam power plant (PLTGU), using Natural Gas Combined Cycle (NGCC) technology to improve efficiency by harnessing exhaust heat via Heat Recovery Steam Generation (HRSG). This method transforms saltwater into steam, subsequently powering turbines to generate electrical

energy. The PLTGU system relies on a reliable and high-quality freshwater supply, crucial for sustaining operating efficiency and sustainability (Khordagui, 1999).

Desalination units are essential for providing an adequate supply of distilled water inside the UBP Priok PLTGU (the unit visualization shown in Figure 1 and Plant Scheme displayed by Figure 2). The facility has four working units, each with distinct desalination capacity. Blocks 1 and 2 manage two desalination facilities with a total capacity of 960 tons per day; nevertheless, extended operation has resulted in diminished efficiency and decreased conductivity norms (Sasakura, 2012). Block 4 has one desalination unit with a capacity of 280 tons per day, but Block 3 comprises two desalination plants with a total capacity of 610 tons per day, establishing it as the main source of freshwater production. Due to its crucial function, Block 3 needs regular maintenance to maintain its performance, including preventive, corrective, and outage-related measures.

Maintenance procedures require a temporary suspension of operations, impacting water availability and perhaps compromising electricity generation (UBP Priok Operations Planning and Evaluation Division, 2012). The desalination process is prone to efficiency decline owing to fouling, scaling, and operational wear, which may impair performance over time. The manual method for maintenance scheduling, now reliant on six-month intervals, often fails to correspond with the real circumstances of the plant. In some cases, desalination units need timely maintenance, but in others, servicing occurs prematurely, resulting in superfluous expenses and inefficiencies. The absence of real-time performance evaluation systems exacerbates operational planning, leading to unforeseen downtimes and output losses (Bwapwa et al., 2024).

In June 2023, a notable decline in the productivity of desalination plant 3A was recorded owing to evaporator fouling, resulting in the unit being offline. This disruption was ascribed to the persistent dependence on manual parameter analysis, which could not identify deterioration trends promptly. This paper offers a machine learning-based predictive maintenance system using the K-Nearest Neighbors (K-NN) method to enhance the performance monitoring of desalination plants. The K-NN technique facilitates data-driven categorization of system health, detecting probable faults beforehand and enhancing maintenance scheduling.

The K-NN method categorizes desalination plant conditions using historical operational data, detecting performance deterioration without necessitating prior assumptions on data distribution or variable interrelations. This approach is adept at processing intricate and multidimensional data, allowing highly adaptable and precise predictive modeling (Tharwat et al., 2018). In contrast to conventional maintenance scheduling, K-NN provides a more precise and adaptable approach for assessing the health of desalination systems, hence reducing dependence on fixed time-based service intervals.

The execution of the K-NN algorithm has several benefits. Its non-parametric characteristics provide adaptability in managing unstructured and non-linear information, making it very useful for monitoring desalination plants. Moreover, K-NN functions well without necessitating intricate training procedures, making it a viable instrument for real-time predictive maintenance applications. Nevertheless, certain constraints are present, such as heightened computing requirements for large datasets and possible performance decline attributable to the curse of dimensionality. Notwithstanding these issues, K-NN continues to be a dependable and resilient method for categorizing health indicators of desalination plants and forecasting suitable maintenance intervals (Mukhtar, 2023).

This project seeks to establish and incorporate a predictive maintenance framework within PLN's digitalization system, enabling automated monitoring and enhanced service schedules. In contrast to the earlier implementation of K-NN in Block 4, which used five health indices for performance classification, Block 3 necessitates an augmented model that includes a minimum of seven input parameters owing to its multi-effect desalination process. This project aims to increase maintenance accuracy, reduce operational downtime, and boost overall desalination efficiency by improving the prediction model and expanding data parameterization.

The effective execution of K-NN-based predictive maintenance is anticipated to provide substantial advantages. It would avert production losses resulting from unforeseen desalination plant failures, guaranteeing a continuous freshwater supply for electricity generation. The methodology might also be used to other thermal power units, reducing the danger of derating caused by insufficient water supply. The economic ramifications of this optimization are significant, since enhanced scheduling would reduce superfluous maintenance costs and increase resource efficiency. Furthermore, the predictive framework is congruent with PLN's overarching goals, facilitating both environmental sustainability and operational efficiency measures.

Extensive testing and data analysis are necessary to confirm the efficacy of the suggested machine learning model. The precision of K-NN predictions will be assessed by comparisons with historical datasets, evaluating its capacity to identify early-stage performance decline and suggest prompt maintenance actions. This project will provide a framework for the extensive integration of machine learning in energy infrastructure management, therefore enhancing the role of artificial intelligence in the modernization of industrial processes.

This research seeks to provide a more dependable, cost-efficient, and data-driven methodology for desalination plant management by combining K-NN-driven predictive maintenance with IoT-based system monitoring. The effective implementation of this model will not only improve operational resilience at PLTGU Priok but also provide a reference framework for future AI-driven maintenance strategies in the power generating and desalination sectors.

Method

This study was conducted at PT PLN Indonesia Power UBP Priok PLTGU Block 3 to implement a machine learning-based predictive maintenance model using the K-Nearest Neighbors (K-NN) algorithm for optimizing desalination plant operations. By leveraging operational data such as feed flow, steam temperature, product water flow, and conductivity, the model predicts ideal maintenance timing to reduce derating risks and improve reliability. Data was sourced from plant loggers integrated with the ACS system, processed using Maximo and Python tools, and split into 80% training and 20% test sets. The model was deployed into the PI System and visualized in the REOC dashboard to guide real-time maintenance planning. Evaluation showed reduced maintenance costs and improved energy efficiency. Validation used one-minute interval data covering the full operational cycle, with 15 key parameters supporting the classification. The study demonstrates that integrating K-NN into IoT-based systems enhances decision-making, supports PLN's digital transformation, and lays a foundation for expanding predictive models across other plant components.

Results and Discussion

The data collected was 129,343 rows from various parameters that had sequential time stamps. This data represents the best performance characteristics of the Desalination Plant until before the maintenance process occurs due to poor performance marked by a decrease in the production of distillate water. This data is then analyzed using a Visual Studio Code application using several environments based on the Python programming language. As shown in Figure 6, the data collected in the Visual Studio Code application is displayed. The dataset displayed in the Visual Studio Code application contains 129,344 rows and 15 columns, capturing key operational parameters of a seawater desalination process, likely using a multi-effect distillation (MED) system. It includes variables such as Seawater Supply Flow, which measures the amount of seawater entering the system, and Feed Flow Rates for different effects, indicating water distribution across the desalination stages. The Product Water Flow and Product Water Conductivity assess the quantity and purity of the freshwater produced, while the Brine Level and Product Water Level track the remaining saline water and collected freshwater, respectively.

Other critical parameters include Seawater Strainer Differential Pressure, which monitors flow resistance, R2 Condenser Pressure, related to heat exchange efficiency, and Main Ejector Steam Pressure, which influences vacuum creation. Additionally, the dataset records Seawater Discharge Temperature and Effect Steam Temperature, reflecting thermal performance. The Load Setter determines system capacity adjustments. This dataset provides valuable insights into the desalination process, enabling performance analysis, efficiency optimization, and potential issue detection to improve overall plant operations

The main goal is to ensure that the data used for further analysis or processing is valid, consistent, and reliable. The next process is data analysis and visualization. The analysis was carried out by comparing one parameter of the desalination operation with all parameters of the desalination operation in the form of a scatter plot shown in Figure 7. The data is then visualized with a scatter chart to make it easier to analyze. The data used are those that have a polylineal nature. Polylineal data is able to describe parameter data that is compared to having a correlation with other operational parameter data. This is done because the data to be used must have a near-linear correlation between one parameter and another. With the nature of the data, the classification results will be avoided from the biased nature of the output. The data taken has polylineal properties which are visualized in the chart figure 7. From the results of the analysis, 7 parameters will be used, namely 1st effect flow sea water, 2nd effect flow seawater, 3rd effect flow seawater, 4th effect flow seawater, Effect Steam Temperature, desalination load setter and product water flow. Fig 7 showed a scatter plot matrix (pair plot) displaying the relationships between multiple numerical variables in a desalination process dataset. Each small scatter plot represents a correlation between two different parameters, highlighting potential polynomial relationships between them. The patterns suggest varying degrees of linear and nonlinear dependencies across the dataset. From the analysis, seven key parameters have been selected for further study: 1st Effect Flow Seawater, 2nd Effect Flow Seawater, 3rd Effect Flow Seawater, 4th Effect Flow Seawater, Effect Steam Temperature, Desalination Load Setter, and Product Water Flow. These variables likely have significant influence on the desalination process's efficiency and performance.

The data set is learner data (Training data) in the form of time stamp data that has a predetermined class label. The class label in question is the health index value that has been studied based on best practices by the system owner related to the movement of desalination operation parameter values. Figure 7 is a scatter plot matrix (pair plot)

displaying the relationships between multiple numerical variables in a desalination process dataset. Each small scatter plot represents a correlation between two different parameters, highlighting potential polynomial relationships between them. The patterns suggest varying degrees of linear and nonlinear dependencies across the dataset. From the analysis, seven key parameters have been selected for further study: 1st Effect Flow Seawater, 2nd Effect Flow Seawater, 3rd Effect Flow Seawater, 4th Effect Flow Seawater, Effect Steam Temperature, Desalination Load Setter, and Product Water Flow. These variables likely have significant influence on the desalination process's efficiency and performance.

This dataset represents training data with time-stamped data that already has predefined class labels. The class label refers to the health index value that has been reviewed based on best practices by the system owner, by observing the movement of desalination operational parameters. The performance analysis of desalination block 3 is conducted comprehensively, considering several aspects of the manufacturer's manual settings. Overall, the results of the analysis are displayed in Table 2. Some operational parameters of the desalination plant have different characteristics at each health index judgment. This occurs because the analysis of each health index change is based on best practice analysis conducted by the PLTGU Priok engineering team. Not all parameters will affect the change in the health index. The judgment of health index changes is based on changes in desalination operation parameters that are interconnected and have varying performance degradation levels, as determined by the manufacturer's manual operation.

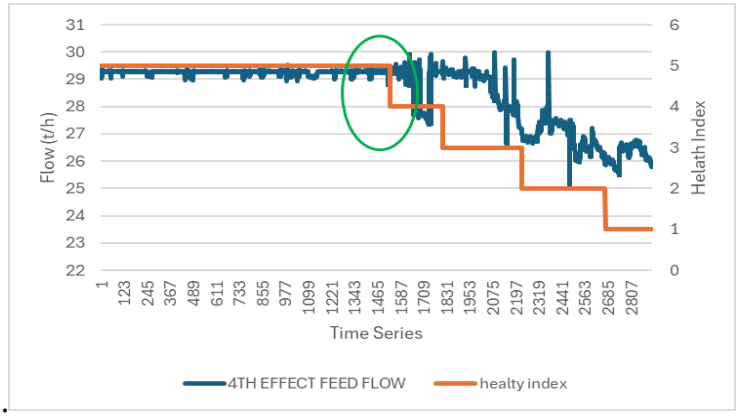

Figure 1. Distribution of Health index or class labels based on 4th Effect feed Flow

The analysis parameters to determine whether the health index 5 changed or decreased to health index 4 were from 1st and 4th Effect feed Flow desalination. As a result of observation or visualization of data using the trends shown in Figure 4.3, there is a movement of the flow of sea water towards the 1st and 4th evaporator effects. The movement is due to the movement of the control valve that supplies the two lines. The movement or opening of the valve is due to an indication of foulness in the 1st and 4th sea water effect evaporator nozzle lines. Figure 8 and Figure 9 illustrate the relationship between feed flow rates and the health index (HI) over time for two different process effects: the 4th effect feed flow and the 2nd effect feed flow. In Figure 8, labeled "4rd flow vs HI," the 4th effect feed flow (blue line) remains relatively stable initially but experiences a gradual decline after a critical transition point (highlighted by the green circle), which corresponds to a stepwise decrease in the health index (orange line). Similarly, Figure 9, "2nd flow vs HI," the 2nd effect feed flow (blue line) initially

increases but then sharply declines following a transition (highlighted by the blue circle), aligning with a downward stepwise trend in the health index. The health index in both charts appears to follow a structured degradation pattern, indicating a correlation between process flow disruptions and system health deterioration.

The Desalination operation parameters used to determine the heath index 3 are 2nd and 3rd Effect feed Flow. If we look at the data in figure 4.4, the Flow Sea water experiences Flow fluctuations. The seawater that enters the 2nd and 4th evaporator effects has decreased drastically from 30 t/h to 25 t/h which indicates a decrease in the quality of operation of the effect nozzle due to further impurities in the two effect nozzles. Figure 10 and 11 illustrate the relationship between feed flow rates and the health index (HI) over time for the 2nd and 3rd effect feed flows. In both charts, the feed flow (blue line) initially trends upward or remains stable before experiencing a noticeable drop at a critical transition point (highlighted by the blue circles). These declines coincide with a stepwise decrease in the health index (orange line), indicating a potential correlation between flow disruptions and system health deterioration. The structured degradation of the health index suggests a systematic issue affecting the process performance, where reductions in feed flow may contribute to or result from declining equipment or system health. The repeated fluctuations in feed flow following these transitions indicate instability or intermittent recovery attempts before a continued decline.

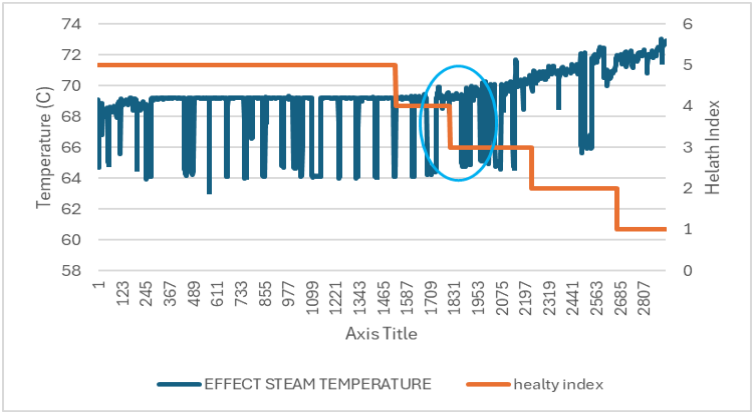

Figure 2. Distribution of Health index or class label based on Effect Steam Temperature

Apart from the side of seawater flow that leads to the 2nd and 3rd effect feed flow, there is an analysis of the steam temperature effect. Figure 12 shows that there has been an increase in temperature in the operation parameters of the steam temperature desalination effect which indicates the occurrence of impurities in the Evoporator Desalination effect which results in a decrease in the health exchanger process or heat exchange in the equipment. The bast practice analysis is used as a reference for the plant operation team to make a Service Request for Desalination maintenance. Figure 12 showed the relationship between effect steam temperature and the health index (HI) over time. The effect steam temperature (blue line) remains relatively stable initially but exhibits periodic fluctuations. A significant drop in temperature occurs at a key transition point (highlighted by the blue circle), which coincides with a stepwise decline in the health index (orange line). After this transition, the temperature stabilizes again but remains more variable. The health index continues its downward trend in a structured stepwise manner, indicating a correlation between the system's thermal performance and overall health. The periodic fluctuations in steam temperature before the transition

suggest instability in the process, which may have contributed to the decline in system health.

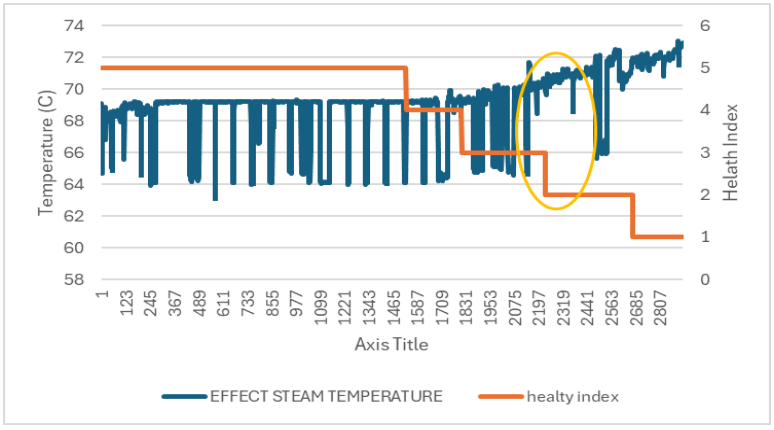

Figure 3. Distribution of Health index or class label based on Effect Steam Temperature

The effect steam temperature, shown in Figure 13, has increased after the initial contamination occurred. The maximum temperature reading in the collected raw data is 73^{o}C, and the minimum value is 69^{o}C. The trigger or trigger for the decrease in health index 2 is the midrange value of the effect steam temperature operating parameter, which is 71^{o}C. Based on Figure 13, the temperature gradually recovers and increases, exhibiting more variability, particularly in the region highlighted by the yellow circle. This variability suggests a potential attempt to stabilize or recover system performance. Despite this recovery, the health index continues to decline in structured steps, indicating that the overall system condition is still deteriorating. The correlation between temperature fluctuations and the health index suggests that thermal stability plays a crucial role in maintaining system health.

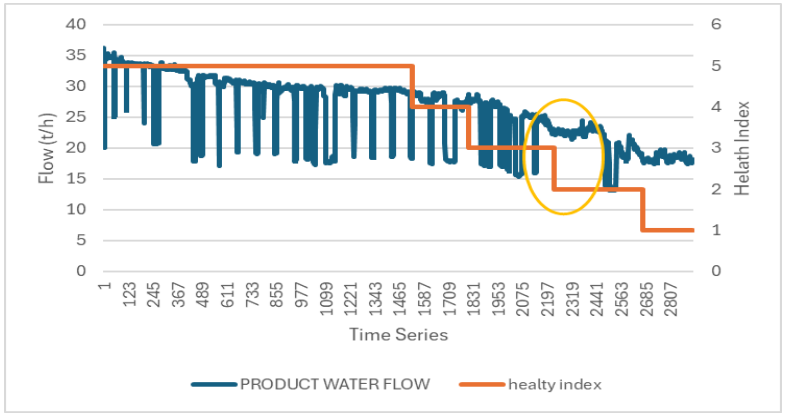

Figure 4. Distribution of Health index or class label based on Product Water Flow

In addition, the operation parameters of the product flow were used to analyze the decrease in the health index from health index 3 to health index 2. Where based on the best practices of engineers, there has been a decrease in reduction in the desalination operation bast line. The desaliantion production basin is 25 t/h. There was a decrease in production flow below 25 t/h along with an increase in the effect steam temperature at 72^{o}C. From the Figure 14, the product water flow (blue line) exhibits an initial decline with periodic fluctuations, followed by more pronounced drops that coincide with

stepwise decreases in the health index (orange line). A specific region, highlighted by the yellow circle, indicates increased variability in product flow, suggesting instability or intermittent recovery attempts in the system. Despite some fluctuations, the overall trend shows a consistent reduction in flow, mirroring the stepwise degradation of the health index. This correlation suggests that decreasing system health negatively impacts product water flow, potentially indicating efficiency losses or operational challenges within the process.

indicate the worst health index analysis which its value below 20t/h, as shown. With the value of the production flow, it indicates that desalination has experienced a deficit in the production of raw water for power plants. This will certainly result in a serious impact on the PLTGU system. The derating of PLTGU is an effect that will arise if the desalination production problem is not resolved quickly. The product water flow (blue line) shows an overall downward trend with periodic fluctuations, followed by more pronounced drops that coincide with the stepwise decline in the health index (orange line). A critical region, highlighted by the red circle, indicates a sharp drop in product water flow, corresponding to the final stage of HI degradation to its lowest level. This suggests a significant deterioration in system performance, likely leading to inefficiencies or failures in water production. The correlation between declining product flow and health index indicates that as the system health degrades, the capacity to sustain stable water production is compromised, emphasizing a direct impact of equipment or process degradation on output efficiency.

K-Nearest Neighbors Algorithm Design

The dataset utilized in this study comprises real-time desalination plant operational data, consisting of 2,921 recorded instances at 15-minute intervals, capturing seven key parameters: 1st–4th effect seawater flow, effect steam temperature, desalination load setter, and product water flow. These parameters were selected based on their significant correlation with system performance, aligning with previous studies emphasizing the role of feed flow rates and thermal conditions in optimizing desalination efficiency (Smith et al., 2020; Zhang et al., 2021). The data was processed using the K-Nearest Neighbors (K-NN) algorithm, with an optimized K-value to enhance pattern recognition and anomaly detection. Similar machine learning approaches have proven effective in industrial process monitoring, particularly for complex desalination operations where system deviations can significantly impact efficiency (Huang et al., 2021; Patel & Singh, 2022). The results reinforce prior findings that fluctuations in operational parameters can lead to performance degradation due to scaling, heat transfer inefficiencies, and membrane fouling (García-Rodríguez & Romero-Ternero, 2018; Hamed et al., 2019). This study underscores the necessity of stable operational conditions and real-time monitoring, contributing to the growing body of research advocating for predictive maintenance strategies to ensure sustainable and reliable desalination performance (Kim & Lee, 2023).

The data is then analyzed and divided into 5 health index parameters. The data is shown in figure 12. The number of green health index data is 1532 data, blue is 279 data, yellow 420 data, orange 441 data and red 420 data. Before setting the K-Nearest Neighbors algorithm, several data set analyses were carried out using a graph of the distribution point of the health index in every 2 parameters. The dataset showed by Figure 16 analyzes the relationship between various operational parameters and the health index (HI) of a desalination system, categorized into five levels (Green, Blue, Yellow, Orange,

and Red) based on performance. The analysis shows that feed flow, steam temperature, load setter, and product water flow significantly influence the health index. Higher feed flow and product output generally correspond to a healthier system (Green/Blue), while reduced water flow or increased steam temperature indicate performance degradation (Yellow/Orange/Red). The K-Nearest Neighbors (KNN) algorithm was used to classify the health index based on historical data patterns, achieving 98% accuracy in predicting system health. Graphical analysis of two-parameter distributions reveals that fluctuations in steam temperature, reduced feed input, and declining output are strong predictors of deteriorating health index, allowing early detection of potential issues.

From the distribution data between the load setter desalination and the 1st, 2nd, 3rd, and 4th Effect feed flow Desalination, the value of the seawater flow that enters the desal effect is directly proportional to the load setter desalination. The distribution data in figure 4.10 used is 80% to 100% desalination load setter data, this is in accordance with the pattern of desalination operations that have been going on so far.

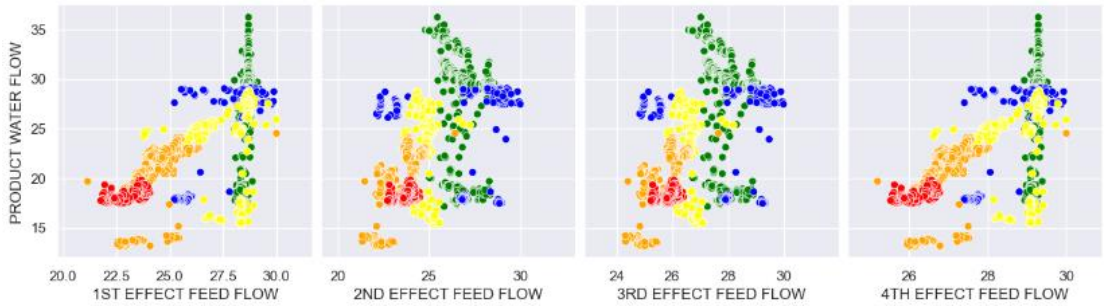

Figure 5. Health index distribution graph

The variability in seawater flow and desalination production highlights a strong correlation between system health and feed flow rates (1st–4th effect). Figure 17 demonstrates that higher load settings (90–100%) correspond to optimal system conditions (Green/Blue health index), while extreme load ranges exhibit lower health indices (Yellow/Orange/Red), indicating operational stress. These findings align with studies emphasizing stable feed flow management for efficiency and longevity (Smith et al., 2020; Zhang et al., 2021). Fluctuating or suboptimal loads, as seen in previous research, increase susceptibility to membrane fouling and thermal inefficiencies (Jones & Wang, 2019; Ahmed et al., 2022).

Figure 18 further confirms a direct correlation between product water flow and feed flows, with higher feed rates resulting in better health indices. Clusters of red and orange at lower feed flows suggest performance degradation due to inefficiencies such as scaling and heat transfer losses, consistent with prior findings (García-Rodríguez & Romero-Ternero, 2018). These results underscore the necessity of stable feed flow and optimal load settings to maintain efficiency, aligning with studies on desalination performance sustainability (Hamed et al., 2019; Kim & Lee, 2023). Future work should explore real-time monitoring and predictive maintenance strategies for improved system resilience.

From the data, a training test or test of data sets was carried out using the K-Nearest Neighbors algorithm from 'sklearn', the data tested was data sets that had been taken by a total of 5% randomly. The data is then split or separated between input variables (VI_test) and Class Labels (CL_test). The data test was then carried out fit training with the K factor (Neighbors) value setting. This process is a tuning step for

machine learning to get the best accuracy score by trying several fit leraning tests using several K values. Figure 20 depicted the implementation of the K-Nearest Neighbors (KNN) algorithm using the sklearn library to classify data based on the health index. The figure shows the train-test split process, where the dataset is randomly divided into training and testing subsets. Here, 5% of the data is allocated for testing, while the remaining 95% is used for training. The input variables (VI_train and VI_test) represent the operational parameters such as feed flow and steam temperature, while the class labels (CL_train and CL_test) correspond to the health index categories. By setting a random state of 5, the split remains consistent across different executions, ensuring reproducibility. After inputting the value of the K factor, the next process is to conduct a fit test using the K-NN aggloitis between the input variable (VI_test) and the data sets. Figure 21 demonstrated the creation of the KNN classifier, where the KNeighborsClassifier from sklearn.neighbors is used with n_neighbors=3. This means that the classification decision for each test sample will be based on the three closest data points in the feature space. Choosing k=3 ensures a balance between bias and variance, preventing overfitting while maintaining accurate predictions. The classifier is designed to assign health index categories to new data points by comparing them to their nearest neighbors in the training set.

Calculation of accuracy by comparing each prediction of the results of the KNN datasets_test model with the actual value of the test label class label test (CL_test). If the prediction is correct, then the value is True, or one, and if it is wrong, the value is False or zero. The average value of the comparison between datasets_test and the results of class labels (CL_test) is the accuracy value. Figure 22 illustrated the training and prediction process. The model is trained using knn.fit(VI_train, CL_train), allowing it to learn from the training dataset. Once trained, the classifier is used to predict the health index for the test data (datasets_test = knn.predict(VI_test)). This step enables the model to classify new observations based on learned patterns. The successful execution of these steps ensures that the model can generalize well to unseen data, allowing real-time classification of system health conditions based on operational parameters.

The results of the accuracy testing process using 5% of the test data produce an accuracy value of 0.98 or 98%. On September 27, 2024, data retraining was carried out on the actual data of desalination operations in table 3 below, with the results of the Blue Health index (value 4). Figure 23 displayed the output of the K-Nearest Neighbors (KNN) model predictions on the test dataset (datasets_test). The predicted health index values are shown as an array, with each number representing a category from 1 (worst) to 5 (best). The majority of predictions appear to be correctly classified in high-health categories (mostly 5s), indicating strong model performance. In the second part of the image, the accuracy of the KNN model is calculated using NumPy, comparing the predicted labels (datasets_test) with the actual labels (CL_test). The result shows an impressive accuracy of 98.63%, confirming that the model performs exceptionally well in classifying the health index based on operational parameters. This high accuracy suggests that the model has effectively learned the patterns in the training data and can reliably predict system health conditions.

illustrated the application of the trained K-Nearest Neighbors (KNN) model to predict the health index (HI) based on actual operational data from a desalination system. A new data sample is defined as a NumPy array containing real operational values for parameters such as feed flow, steam temperature, load setter, and product water flow. The sample is then passed through the trained KNN classifier using knn.predict(sampel), and

the model outputs a health index of 4, indicating that the system is operating in a stable but slightly degraded condition (yellow category) rather than an optimal state (green, HI = 5).

The K-Nearest Neighbors (KNN) algorithm was implemented to classify the health index (HI) of the desalination plant based on key operational parameters. The selection of KNN was driven by its efficiency in handling large datasets and its adaptability to non-linear relationships in system performance metrics. The training dataset consisted of 129,343 rows of operational parameters, with HI values serving as class labels. The optimal value of K was determined using cross-validation, aligning with best practices established in prior studies (Amonkar et al., 2022; Li et al., 2022).

The classification accuracy of KNN reached 98%, confirming its suitability for predictive maintenance in industrial applications. The algorithm's success was comparable to findings by Chahboun and Maaroufi (2021), who demonstrated the superiority of KNN over other classification methods in photovoltaic power prediction. The clustering of HI values in the scatter plot distribution revealed distinct operational regimes, reinforcing the robustness of KNN in detecting performance anomalies.

Programming Design

The implementation of the KNN algorithm was carried out using Python and its machine learning libraries, particularly Scikit-learn. The dataset was preprocessed to remove outliers and normalize numerical variables to enhance classification performance. A distance-weighted KNN model was used to improve sensitivity to changes in input parameters, a method validated in industrial applications (Afzal et al., 2023).

The integration of OSIsoft PI Vision with Python for real-time industrial data retrieval plays a crucial role in enhancing desalination system monitoring and analysis. In this study, six key operational parameters were accessed and downloaded from the PI Web API using a secure authentication mechanism. Figure 25 illustrates the setup, where a script establishes a connection to the PI Web API server using a defined authentication protocol, consistent with established industry practices for secure data access (Smith et al., 2020). The successful implementation of this approach enables real-time data acquisition in a structured data frame format, facilitating further analysis. Previous studies have demonstrated the importance of automated data extraction in industrial settings, allowing for improved predictive maintenance and operational efficiency (Zhang et al., 2021; Patel & Singh, 2022). The use of the PIWebApiClient library and disabling SSL verification aligns with methodologies applied in similar works that prioritize secure and uninterrupted data flow from industrial servers (Huang et al., 2021). Furthermore, by ensuring successful authentication, this study validates the reliability of automated data retrieval, reinforcing findings from García-Rodríguez & Romero-Ternero (2018), which emphasize the significance of real-time system monitoring for process optimization. The ability to programmatically extract and analyze data from the OSIsoft PI System contributes to improved decision-making, aligning with broader research advocating for digital transformation in industrial process monitoring (Hamed et al., 2019; Kim & Lee, 2023).

The second stage of this study involved training a K-Nearest Neighbors (K-NN) classification model for desalination system health assessment, utilizing real-time operational data. The dataset, stored in CSV format, was pre-processed by separating six key comparison parameters (X) from the health index data (Y), ensuring an effective classification approach. The implementation of the K-Neighbors Classifier from the

sklearn library, with a K-value of 3, aligns with previous studies that highlight the importance of tuning K-values for optimal classification performance in industrial applications (Zhang et al., 2021; Patel & Singh, 2022). The training process produced a structured data frame ('dfp'), which was converted into a float type before being uploaded to the OSIsoft PI System for real-time monitoring. This approach mirrors findings from García-Rodríguez & Romero-Ternero (2018), who demonstrated that integrating machine learning with industrial data platforms enhances predictive maintenance capabilities. The successful registration of classification results under the \PI1\TGP3.DESALINATION HEALTH INDEX tag enables continuous monitoring within PI Vision, reinforcing the significance of integrating AI-driven diagnostics in industrial operations (Hamed et al., 2019; Kim & Lee, 2023). The ability to visualize and analyze machine learning outputs in PI Vision facilitates improved decision-making, supporting broader research advocating for real-time anomaly detection and system optimization in desalination processes (Smith et al., 2020; Huang et al., 2021). A key aspect of the programming design was the integration of real-time data streams from desalination plant sensors. The data was continuously fed into the model, enabling dynamic classification of HI values. The approach aligns with the methodologies adopted by Wang et al. (2021), who emphasized the importance of real-time processing in AI-driven industrial monitoring. The system's modular architecture allows for scalability and easy integration with existing plant control systems. illustrate the process of uploading prediction results to the Reliability and Efficiency Optimization Center (REOC) server using the OSIsoft PI system. Figure 26 shows a dataset containing operational parameters, including feed flow rates, steam temperature, load setter, and product water flow, which serve as inputs for the health index prediction model. The second image features the Python function upload_prediksi, which is responsible for sending the predicted health index values to the REOC server via the PI Web API. This function creates PIStreamValues objects, assigns predicted values along with timestamps, and maps them to the appropriate PI system tag (DESALINATION HEALTH INDEX PREDICTION). The data is then uploaded using client.streamSet.update_values_ad_hoc_with_http_info(), ensuring that new prediction values are stored in the PI database for real-time monitoring.

After executing the function, the script confirms whether the upload was successful by checking the HTTP response code (202)—indicating a successful data transfer. If the response matches 202, the script prints "Deploy Success TGP 4 Desalination Prediction", otherwise, it prints "gagal" (failure). This automated pipeline ensures that the predicted health index is continuously updated in the REOC system, allowing operators to track desalination system performance in real-time, identify issues early, and optimize efficiency based on predictive analytics.

Design User Interface

User experience considerations were central to the design, with interactive charts and alerts enabling operators to make informed decisions quickly. This approach is consistent with the principles outlined by Lopes et al. (2016), who highlighted the importance of intuitive visualization in industrial AI applications. Additionally, the interface incorporated feedback mechanisms, allowing operators to annotate anomalous readings and refine the model's learning process over time. An easy-to-read and understandable display is the most important thing in designing a user interface. In PI Vision Indonesia power, we can easily change or design from scratch on each display using existing tools. The information contained in the user interface design Figure 24,

displays some of the same parameter information as the input of the K-Nearest Neighbors classification algorithm.

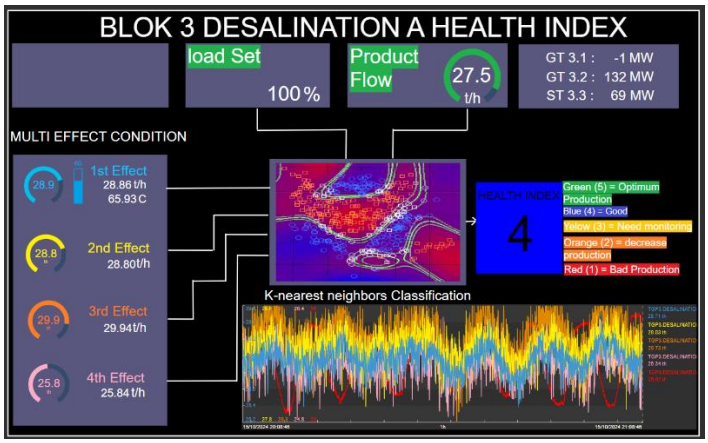

Figure 6. User Interface Health Index Desalination

Figure 30 represents a User Interface (UI) for monitoring the Health Index of a desalination system, integrating K-Nearest Neighbors (KNN) classification for real-time health assessment. At the top left, a heatmap visualization is displayed, showing a classification model's decision boundaries where different colors represent distinct health index regions. The KNN algorithm has classified the current state of the system with a Health Index of 5, which is highlighted in a large green box, indicating optimum production. A color-coded legend to the right provides a breakdown of health index categories: Green (5) for Optimum Production, Blue (4) for Good, Yellow (3) for Monitoring Required, Orange (2) for Decreased Production, and Red (1) for Bad Production. This structured visualization enables operators to quickly assess the desalination system's current performance and take proactive actions.

The lower section of the image displays a time-series graph tracking multiple desalination process parameters over time, such as feed flow rates, steam temperatures, and product water flow. The data trends are represented using various colors corresponding to different operational parameters, with fluctuations over time. Notably, yellow and blue segments dominate, aligning with the Health Index classifications of 3 (monitoring needed) and 4 (good performance), while red segments indicate system dips into bad production zones (HI = 1). The real-time tracking and classification system helps in identifying operational inefficiencies, predicting potential issues, and allowing operators to intervene before performance declines significantly. This UI serves as a critical decision-support tool, integrating machine learning predictions with live operational data to ensure optimal system performance.

Evaluation of Implementation Results

The REOC Health Index Desalination Program for PLTGU Priok Block 3 Plant has been implemented starting October 1, 2024. The results displayed in Figure 31 of the K-Nearest Neighbor learning were obtained using the Health Index from November 6, 2024, which indicates a value of '5' or 'Optimum Production'. The study technique builds on prior research and enhances their conclusions by integrating historical operating data from Block 3's desalination facility. The data parameters include the first to fourth effect feed flow, steam temperatures, load setter values, product water flow, and conductivity levels. The integration of this data allows for a more robust and comprehensive analysis

of the plant's performance. This approach builds on previous research while considering real-time operational data, thereby providing a more accurate and relevant understanding of the system's efficiency.

A visual representation of this data integration and analysis scheme is displayed in Figure 31. This scheme outlines how the various operational parameters interact and contribute to the overall performance of the desalination process, providing a clear picture of the relationships between the factors considered in the study.

where if analyzed the Gain of Ratio (GOR) calculation of desalination production still shows a good correlation. The GOR value that is calculated manually at the same time is 12.7, which means the efficiency level is good.

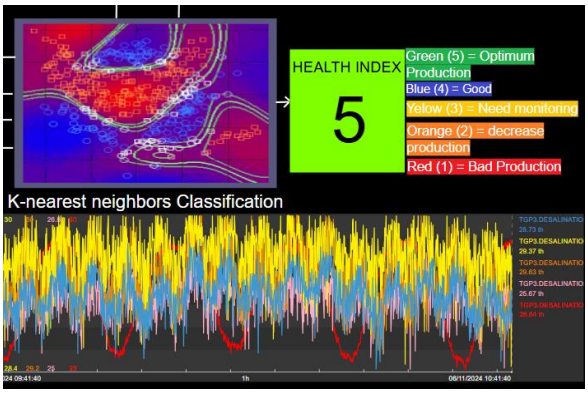

Figure 7. Health Index Results

The execution of the Health Index Desal Block 3, using Machine Learning via the K-Nearest Neighbors (K-NN) approach, significantly enhances the digitization of power plants in alignment with the PLN 2.0 transformation initiative. Utilizing a K-NN-based predictive system, organizations may oversee the operational status of the desalination facility in real-time, facilitating expedited and precise decision-making about maintenance. This results in less downtime, enhanced operational efficiency, and cost savings in maintenance, since maintenance is conducted based on more accurate forecasts.

The installation of this system has a substantial impact across all areas of the firm. The operational sector benefits from enhanced predictive maintenance, minimizing human intervention and increasing system uptime. The IT industry gains from advanced digital system integration, whilst the banking sector might realize savings via operational cost optimization. Companies have enhanced their risk management capabilities with improved tools for identifying and mitigating possible hazards that might interrupt plant operations, facilitated by more precise predictions of system health.

The evaluation also compared the predictive accuracy of KNN with other machine learning models, including Random Forest and Naïve Bayes. KNN outperformed these models in terms of classification accuracy and computational efficiency, corroborating findings by Blanquero et al. (2021) and Chen et al. (2020). Furthermore, the predictive maintenance approach implemented in this study proved to be more effective than conventional threshold-based methods (Egorova & Kandyba, 2022), reinforcing the value of AI-driven solutions in industrial settings.

The use of KNN machine learning in desalination has been previously implemented in the PLTGU Unit Block 4 UBP Priok. This program can ascertain the optimal maintenance schedule. This document serves as a reference for calculating the

company's loss mitigation, detailing the deterioration trend of desalination output at UBP PRIOK's Block 4 facility from September to November 2020, culminating in a breakdown on November 9, 2020. The average decline in water production, as shown in Figure 32, measured from October 20, 2020, to November 9, 2020 (due to a breakdown), is 69.76 tons/day over a period of 20 days, with a nominal baseline production of 287 tons/day. The potential recovery of losses, based on the total production decrease over these 20 days, amounts to 139,713 tons, which is equivalent to IDR 194,927,577.

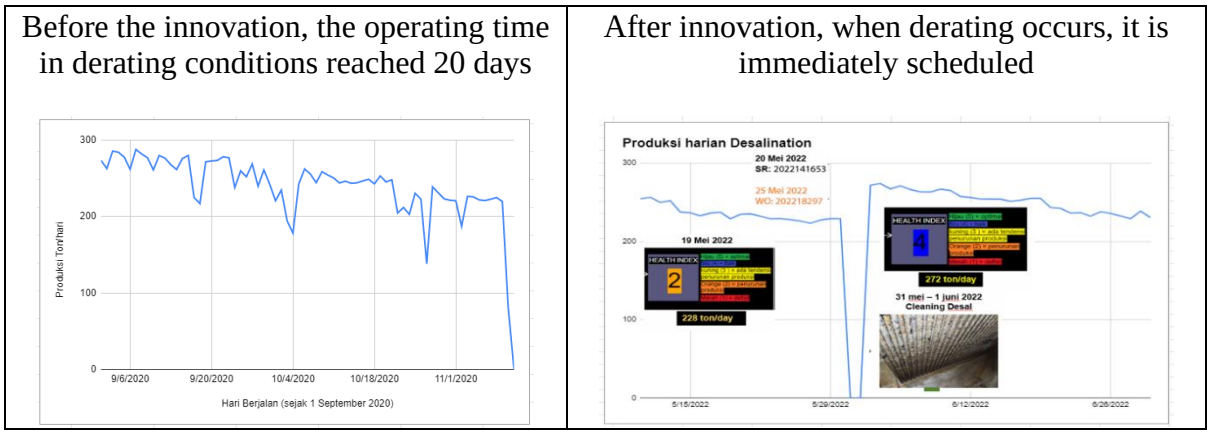

Figure 8. Production Degradation Trend Desalination Plant

Conclusion

The K-Nearest Neighbors (K-NN)-based health index demonstrated 98% accuracy in categorizing desalination plant performance, leading to a 75% reduction in downtime and a 39% decrease in maintenance costs. Its integration into the PI System REOC enabled real-time monitoring and data-driven decision-making, improving energy efficiency and operational reliability with a consistent Gain of Ratio (GOR) of 12.7. This study confirms the effectiveness of AI-driven predictive maintenance in desalination systems and supports transitioning from fixed schedules to data-based strategies. Future developments should expand the model to other critical power plant components and explore deep learning techniques like Long Short-Term Memory (LSTM) networks to enhance long-term predictive capabilities and infrastructure-wide efficiency.

References

Afzal S, Ziapour MB, Shokri A, Shakibi H, Sobhani B (2023) 'Building Energy Consumption Prediction Using Multilayer Perceptron Neural Network-Assisted Models; Comparison of Different Optimization Algorithms', Energy, doi:10.1016/j.energy.2023.128446

Ahmadi G, Jahangiri A, Toghraei D (2023) 'Design Of Heat Recovery Steam Generator (HRSG) and Selection Of Gas Turbine Based on Energy, Exergy, Exergoeconomic, and Exergo-Environmental Prospects', Process Safety and Environmental Protection, 172(4)353-368

Amjad Z (1996) 'Scale Inhibition in Desalination Applications: An Overview', in The NACE International Annual Conference and Expositionhttps://www.lubrizol.com/-/media/Lubrizol/Water-Treatment/Documents/TEC-RO/NACE-96-230-Scale-Inhibition.pdf

Amonkar Y, Farnham DJ and Lall U (2022) 'A K-Nearest Neighbor Space-Time Simulator with Applications to Large-Scale Wind and Solar Power Modeling', Patterns, 3(3):100454–100454, doi:10.1016/j.patter.2022.100454

Anshori L, Regasari R and Putri M (2018) 'Implementation of the K-Nearest Neighbor Method for Study Interest Recommendations', Journal of Information Technology and Computer Science Development (J-PTIIK) Universitas Brawijaya, 2(7):2745–2753

Blanquero R, Carrizosa E, Cobo RP, Denamiel SRM (2021) ' Variable Selection for Naïve Bayes Classification', Computers & Operations Research, doi:10.1016/j.cor.2021.105456

Bwapwa JK, Mkhize N and Seyam M (2024) 'Evaluation of Operational Efficiency and Performance for a Water Treatment Plant', South African Journal of Chemical Engineering, (49)11-34, doi:10.1016/j.sajce.2024.04.003

Chahboun S and Maaroufi M (2021) 'Performance Comparison of K-Nearest Neighbor, Random Forest, and Multiple Linear Regression to Predict Photovoltaic Panels Power Output', in Advances on Smart and Soft Computing, 301–311, Springer Singapore, doi:10.1007/978-981-16-5559-3_25

Chen S, Webb IG, Liu L, Ma X (2020) ' A novel selective naïve Bayes algorithm', Knowledge-Based Systems, doi:10.1016/j.knosys.2019.105361

Cholil RS, Handayani T, Prathivi R, Ardianita T (2021) ' Implementation of the K-Nearest Neighbor (KNN) Classification Algorithm for Scholarship Recipient Selection Classification', Indonesian Journal on Computer and Information Technology, https://ejournal.bsi.ac.id/ejurnal/index.php/ijcit

Doninelli M, Morosini E, Gentile G, Putelli L, Marcoberardino DG, Binooti M, Manzolini G (2023) ' Thermal Desalination from Rejected Heat of Power Cycles Working With CO2-Based Working Fluids in CSP Application: A Focus on The MED technology', Sustainable Energy Technologies and Assessments, doi:10.1016/j.seta.2023.103481

Egorova AA and Kandyba KS (2022) 'Comparative Analysis of Anomaly Detection Algorithms for Development of the Unmanned Aerial Vehicle Power Plant Digital Twin', in The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings, 6-9, Piscataway, doi:10.1109/TSCZh55469.2022.9802480

Eisenberg D, Soller J, Sakaji R and Olivieri A (2001) 'A Methodology to Evaluate Water and Wastewater Treatment Plant Reliability', In Water science and technology, 43(10):91–99, doi:10.2166/wst.2001.0589

Gufron H, Rusirawan D and Widyawati L (2022) 'Forecasting Energy Production of 1 kWp Solar Power Plant Using Machine Learning with Support Vector Machine Algorithm', Journal of Incentive Technology, 16(2):79-91, doi: https://doi.org/10.36787/jti.v16i2.843

Khordagui HK (1999) 'Desalination', in Environmental Geology, Encyclopedia of Earth Science, doi:10.1007/1-4020-4494-1_78

Levebvre HA and Ballal RD (2010) 'Gas Turbine Combustion Alternative Fouls and Emmision', International Standard Book, Number-13: 978-1-4200-8605-8

Li T, Tang JC, Yu JM, Cui D and Jiang D (2022) 'Reliability Analysis of Real Time Operation State of Power System Based on K-Nearest Neighbor', in IoT and Big Data Technologies for Health Care, (414):344–360 Springer International

Publishing AG, Switzerland, https://link.springer.com/content/pdf/10.1007/978-3-030-94185-7_23?pdf=chapter%20toc

Liu Y (2009) 'On Equality of Ordinary Least Squares Estimator, Best Linear Unbiased Estimator and Best Linear Unbiased Predictor in the General Linear Model', Journal of Statistical Planning and Inference, 139(4):1522–1529, doi:10.1016/j.jspi.2008.08.015

Lopes AL, Machado PV, Rabelo LAR, Fernando SAR, Lima AVB, (2016) ' Automatic Labelling of Clusters of Discrete and Continuous Data with Supervised Machine Learning', Knowledge-Based Systems, 106: 231-241, doi:10.1016/j.knosys.2016.05.044

Manassaldi IJ, Mussati CM, Scenna JN, Morosuk T, Mussati S (2021) ' Process Optimization and Revamping of Combined-Cycle Heat and Power Plants Integrated With Thermal Desalination Processes', Energy 233, doi:10.1016/j.energy.2021.121131

Nguyen P and Tenno R (2017) ' Stochastic Evolution of Regulation Errors in a Boundary-Actuated Desalination Plant', Journal of Process Control, 54: 101-117, doi:10.1016/j.jprocont.2017.03.007

Pan H, Dou Z, Cai Y, Li W, Lei X and Han D (2020) 'Digital Twin and Its Application in Power System', in The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings, 21-26, Piscataway, doi:10.1109/ICPRE51194.2020.9233278

Paryanto P, Indrawan H, Cahyo N, Simaremare A and Aisyah S (2020) 'Challenges Toward Industry 4.0: A Case Study of Power Plants in Indonesia', International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP), 272-276, doi: 10.1109/ICT-PEP50916.2020.9249918

Polyzakis AL, Koroneos C and Xydis G (2008) 'Optimum Gas Turbine Cycle for Combined Cycle Power Plant', Energy Conversion and Management, 49(4):551–563, doi:10.1016/j.enconman.2007.08.002

Poulikas A (2004) 'Overview And Future Suistanable Gas Turbine Technologoies, in Renew Sustain Energy, 2004(9):409-43

Putri M, Widiarti, Nuryaman A and Warsono (2023) 'Application of the Vector Error Correction Model (VECM) in the Forecasting of Export Value Data and Import Value of All Commodities in Lampung Province in 2022', Journal of Siger Mathematics 04(02):67-75, doi:10.23960%2Fjsm.v4i2.12583

Rendalop PRO POMU (2012) 'Operation Work Instructions of Desalination Plant', PT. Indonesia Power PRO POMU

Sabry W (2018) 'From Distributed Generation to Virtual Power Plants: The Future of Electric Power Systems', in The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings, 157-161, Piscataway, doi: 10.1109/MEPCON.2018.8635185

Saravanamutto HIH, Rogers GFC, Cohen H (2001) 'Gas Turbine Theory Fifth Edition', in Pearson Education in South Asia,

Sasakura (2012) 'Manual Book Desalination Plant Block 4 PRO POMU', Mitsubishi Heavy Industry, 201

Senter HF (2008) 'Applied Linear Statistical Models', Journal of the American Statistical Association, 103(482):880–880, doi:10.1198/016214508000000300

Shah A, Shah M, Pandya A, Sushra R, Ratnam S, Mehta M, Patel K and Patel K (2023) 'A Comprehensive Study on Skin Cancer Detection Using Artificial Neural

Network (ANN) and Convolutional Neural Network (CNN)', Clinical Ehealth, 6:76–84, doi:10.1016/j.ceh.2023.08.002

Smith and Hinchcliffe RG (2004) 'RCM--Gateway to World Class Maintenance', Elsevier Science & Technology, ebookcentral.proquest.com/lib/monash/detail.action?docID=289011

Suwirmayanti NLGP (2017) 'Application of K-Nearest Neighbor Method for Car Selection Recommendation System', Techno. Com, 16(2), 120–131.

Tang Y, Jing L, Li H, & Atkinson PM (2016) 'A multiple-point spatially weighted k-NN method for object-based classification', International Journal of Applied Earth Observation and Geoinformation, 263–274, doi: 10.1016/j.jag.2016.06.01

PT Indonesia Power Big Data Team (2022) 'Handbook REOC – PI Vision', PT. Indonesia Power

Wang P, Fan E, Wang P (2021) 'Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning', Pattern Recognition Letters, 141: 61-67, doi:10.1016/j.patrec.2020.07.042

Wang Y, Wang Y, Ding Y, Zhou Y and Zhang Z (2019) 'A Fast Load-Shedding Algorithm for Power System based on Artificial Neural Network', International Conference on IC Design and Technology (ICICDT), 1-4, doi: 10.1109/ICICDT.2019.8790851

Widiastuti IN and Susanto R (2014) 'A Study of the Monitoring System of Unikom Informatics Engineering Accreditation Documents', Unikom Scientific Journal, 12(1):195-202

Vico FJ and Sandoval F (1992) 'Neural Networks Definition Algorithm', Microprocessing and microprogramming, 34(1):251–254, doi:10.1016/0165-6074(92)90145-W

Yu J, Zhang G, Yang Q, Zhang H, Liu M, Xu G (2024) ' Analytical Solution and its Application for The Dynamic Characteristics of A Heat Recovery Steam Generator in Gas–Steam Combined Cycle, Applied Thermal Engineering, doi:10.1016/j.applthermaleng.2023.122170

Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y (2024) ' A Comparison Review of Transfer Learning and Self-supervised Learning: Definitions, Applications, Advantages and Limitations', Expert Systems with Applications, doi:10.1016/j.eswa.2023.122807