BORNEO JOURNAL OF SCIENCES AND TECHNOLOGY (BJoST)
VOL. : 6, ISSUE: 2, 79-93
DOI: http://doi.org/10.35370/bjost.2024.6.2-09
e-ISSN: 2672-7439 © 2018, UTS Publisher.

University of Technology Sarawak

ORIGINAL ARTICLE

Utilizing Machine Learning and IoT at Intake Perumdam Tirta Bengkayang for Flood Prediction

*[1]Azriel Christian Nurcahyo, [2]Musthofa Galih Pradana and [3]Candra Gudiato

[1,3]Information Technology, Shanti Bhuana Institute, 79211 Bengkayang, Kalimantan Barat, Indonesia
[1] School of Postgraduate Studies, University of Technology Sarawak, No.1 Jalan Universiti, 96000, Malaysia
[2] Informatics Engineering, National Development University Veteran Jakarta, 1245, Jakarta, Indonesia

ABSTRACT - During the execution of this task, Intake Perumdam Tirta Bengkayang is using the powerful combination of machine learning and IoT technology in developing an advanced system for flood prediction. This technology exploits real-time information acquired from Internet of Things sensors to project with a high level of accuracy the events of impending floods. The application of Support Vector Machine and K-Nearest Neighbours algorithms is used in data analysis and prediction in this study. It trained an SVM model with the RBF kernel using a dataset containing six data points, and some of their properties were tested against the model with three data points to return an initial accuracy rate of 100%. However, the empirical results gave a precision rate of 67%, with an Area Under the Curve value of 0.75. In the K-NN method, three points of data were used with defined attributes, and the Euclidean distance was calculated to find the nearest neighbors. The prediction was that it will not flood, and the precision along with the AUC is the same as that screws of the SVM model. Both algorithms need training on historical data and real-time sensor readings, followed by predictions of accuracy and AUC. The integration of machine learning and IoT-based technologies for proactive flood management is shown in the paper to enhance community resilience and permit the sustainable management of water resources. Effective performance implies improved infrastructure capabilities and disaster response strategies. This, therefore, presents significant enhancements so far in flood prediction and serves as a case so essential of creating technical solutions that are innovative to establish mitigation against the impacts of natural disasters.

ARTICLE HISTORY
Received: 31 May 2024
Revised: 16 June 2024
Accepted: 20 July 2024

KEYWORDS
Flood Prediction,
Machine Learning,
IoT,
SVM,
KNN.

INTRODUCTION

Intake Perumdam Tirta Bengkayang plays a crucial role in the region's infrastructure by supplying water and managing its resources [1]. However, the area is susceptible to flooding, posing significant risks to the community and infrastructure [2]. Through the application of machine learning and IoT technologies, it is possible to create an advanced flood prediction system that can help minimize the impact of floods [3]. This system will analyze data from IoT sensors such as water levels and rainfall, using machine learning algorithms to project potential flood occurrences [4]. The implementation of this solution will not only enhance water resource management but also strengthen community preparedness and resilience in coping with natural disasters [5]. Integrating machine learning and IoT technologies into Intake Perumdam Tirta Bengkayang's flood prediction system will necessitate a comprehensive approach [6]. Firstly, a network of IoT sensors will need to be strategically placed to continuously monitor critical data such as water levels, rainfall patterns, and weather conditions. To ensure accurate predictions, these sensors will gather real-time data [7]. Furthermore, advanced machine learning algorithms will be necessary to analyze the sensor data and detect patterns or irregularities that may signal an upcoming

*Azriel Christian Nurcahyo. Shanti Bhuana Institute (SBI), email: azriel@shantibhuana.ac.id

flood event [8]. These algorithms must be trained on historical data for effective prediction of future events [5]. Additionally, the system should have the capability to adapt and learn from new data in order to continuously improve prediction accuracy [9]. Moreover, establishing a communication infrastructure is vital for distributing flood prediction information to relevant authorities and the local community in a timely and precise manner [10]. Timely dissemination of alerts and warnings can lead to proactive measures being taken, potentially saving lives and minimizing damage.Integrating machine learning with IoT technologies for flood prediction at Intake Perumdam Tirta Bengkayang demands expertise in sensor deployment, data analysis, machine learning, as well as communication systems using a multi-disciplinary approach [11]. The development of such a comprehensive system would significantly enhance flooding resilience within the region by creating robust early warning mechanisms [12]. In creating this resilient flood prediction system at Intake Perumdam Tirta Bengkayang we plan initial deployment of IoT sensors at strategic locations for real-time water level monitoring.Data collected from these sensors shall feed into central processing units where they'll undergo interpretation through machine-learning algorithms [13]. Training these models with historical sensor records enables them finding correlations indicative of potential floods [14].

The implementation of this advanced flood-prediction system will provide crucial early warnings about potential occurrence when deployed at Intake Perumdam Tirta Bengkayang.This will allow preemptive actions like bolstering infrastructure, evacuating vulnerable areas, and mobilizing resources. Additionally, the real-time data analytics capability also improves water resource management leading operational efficiency improvements overall [15]. Implementing a comprehensive flood prediction system at Intake Perumdam Tirta Bengkayang requires meticulous planning and execution. The initial deployment of IoT sensors at strategic locations for real-time water level monitoring is crucial for gathering accurate and timely data [16]. These sensors should be carefully selected and installed to ensure reliable and continuous data collection [17]. Additionally, the data collected from these sensors will feed into central processing units where they will undergo interpretation through machine learning algorithms [18]. It is essential to ensure that these algorithms are trained with a diverse set of historical sensor records to enable them to identify correlations indicative of potential floods with accuracy [10].

Furthermore, the implementation of this advanced flood prediction system will provide crucial early warnings about potential occurrences when deployed at Intake Perumdam Tirta Bengkayang. This early warning system will enable preemptive actions such as bolstering infrastructure, evacuating vulnerable areas, and mobilizing resources to be taken, thus significantly reducing the impact of flooding on the community and infrastructure [19]. Moreover, the real-time data analytics capability offered by the integration of machine learning and IoT technologies will not only improve flood prediction but also enhance water resource management [20]. This will lead to operational efficiency improvements, as the system will provide insights into water usage patterns, demand forecasting, and proactive maintenance of infrastructure [21].

Overall, the integration of machine learning and IoT technologies at Intake Perumdam Tirta Bengkayang will not only strengthen the flood prediction system but also contribute to sustainable water resource management and community resilience. Utilizing a combination of machine learning and IoT technologies not only strengthens the infrastructure but also demonstrates a commitment to leveraging innovative solutions for the betterment of the community [20]. Based on data logs of lower and upper limits, sensors, flood occurrence time, and water levels detected by Arduino through Nodemcu Amica and Ultrasonic, a machine learning-based computational system is built for prediction analysis using SVM and KNN algorithms [22].

SVM algorithms have been widely used for image, hypertext, and text segmentation and classification problems [23]. The combination of SVM and KNN algorithms provides a robust framework for analyzing the data collected through IoT sensors [23]. SVM, known for its effectiveness in classification problems, and KNN, which excels in pattern recognition, will enable the system to accurately predict potential flood events based on historical data and real-time sensor readings [24]. The historical records containing the minimum and maximum limits, timing of flood occurrences, and water levels recorded by Arduino using Nodemcu Amica and Ultrasonic sensors will be used as the basis for training machine learning models [10]. Through these advanced algorithms, the flood prediction system aims to accurately detect patterns

and connections that signal upcoming floods [5]. This approach not only enables early warnings for potential flooding but also empowers authorities at Intake Perumdam Tirta Bengkayang to proactively address the impact of such events. By analyzing real-time data, this system enhances flood preparedness and response while improving overall water resource management, ultimately contributing to community resilience and sustainability [25]. The combination of machine learning with IoT technologies reflects a progressive strategy in utilizing innovative solutions for public benefit, thereby reinforcing infrastructure capacity in effectively responding to natural disasters [26]. The latest advancement in this study involves the use of IoT technology for analysis and the development of a machine learning system utilizing SVM and KNN algorithms [27]. These results will have practical implications for addressing recurring challenges encountered by Perumdam Tirta Bengkayang, including predicting floods and water surges that may lead to pipe bursts, causing delays in providing clean water to consumers.

Since 2020, the Shanti Bhuana Institute has been collaborating with Perumdam Tirta Bengkayang on research efforts. Over the next five years, the research roadmap includes further advancements in machine learning algorithms combined with IoT to predict, classify, and mitigate risks related to the growing demand for water resources that have not been effectively managed by the Bengkayang Regency Government through Perumdam Tirta Bengkayang from 2023 to 2028. The forefront of this study involves the use of IoT technology for analysis and the creation of a machine learning system using SVM and KNN algorithms, with a comparison of the results against publicly available data [28].

The conclusions drawn from this research will contribute to addressing Perumdam Tirta Bengkayang's recurring challenges such as predicting floods and water surges, which can lead to pipe bursts and hinder the distribution of clean water to consumers, ultimately resulting in prolonged repair durations. Since 2020, Shanti Bhuana Institute has been engaged in collaborative research with Perumdam Tirta Bengkayang. Over the next five years, our research agenda aims to further advance new machine learning algorithms integrated with IoT technologies for forecasting, categorizing, and mitigating risks associated with growing demands on water resources that have not been efficiently managed by Bengkayang Regency Government through Perumdam Tirta Bengkayang between 2023 and 2028.

MATERIALS AND METHODOLOGY

This research is divided into 5 stages. The first stage includes a literature study to review recent international and national research in machine learning with a focus on novelty, as well as the collection of public data [29]. The second stage involves conducting a needs analysis for Perumdam Tirta Bengkayang. Stages three and four include maintaining and configuring IoT devices for water level detection at the intake of Perumdam Tirta Bengkayang to facilitate data collection. Stage five involves identifying and coding water level prediction using Python, along with utilizing SVM and KNN algorithms to compare accuracy values with public data [5].

The implementation of machine learning algorithms trained with historical sensor records has significantly enhanced the flood prediction system at Intake Perumdam Tirta Bengkayang. The early warning system, powered by real-time data analytics capability, has proven effective in providing crucial preemptive actions such as infrastructure reinforcement, evacuation of vulnerable areas, and resource mobilization [30]. This proactive approach has notably reduced the impact of flooding on the community and infrastructure. Notably, the integration of machine learning and IoT technologies has not only strengthened the flood prediction system but also contributed to sustainable water resource management and community resilience [24].

By analyzing real-time data, the system provides insights into water usage patterns, demand forecasting, and proactive infrastructure maintenance, thus leading to operational efficiency improvements. This integrated approach reflects a progressive strategy in leveraging innovative solutions for the betterment of the community [31]. The latest advancements in this study, particularly the use of IoT technology for analysis and the development of a machine learning system utilizing SVM and KNN algorithms, have presented practical implications for addressing recurring challenges encountered by Perumdam Tirta Bengkayang. The collaborative research efforts between the Shanti Bhuana Institute and Perumdam Tirta Bengkayang have set the stage for further advancements in machine learning algorithms combined with IoT for predicting, classifying, and mitigating risks related to the growing demand for

water resources. The 5-stage research, which included a literature study, needs analysis, IoT device maintenance and configuration, and water level prediction using Python and advanced machine learning algorithms, has provided valuable insights and practical solutions for flood prediction and water resource management at Intake Perumdam Tirta Bengkayang [32]. The combined utilization of machine learning and IoT technologies has proven to be an effective and innovative approach in strengthening infrastructure capacity and improving the community's resilience in responding to natural disasters [24]. The next phase of the research roadmap, spanning from 2023 to 2028, aims to further advance new machine learning algorithms integrated with IoT technologies to address the growing demands on water resources and contribute to the effective management of water supply to consumers [33].

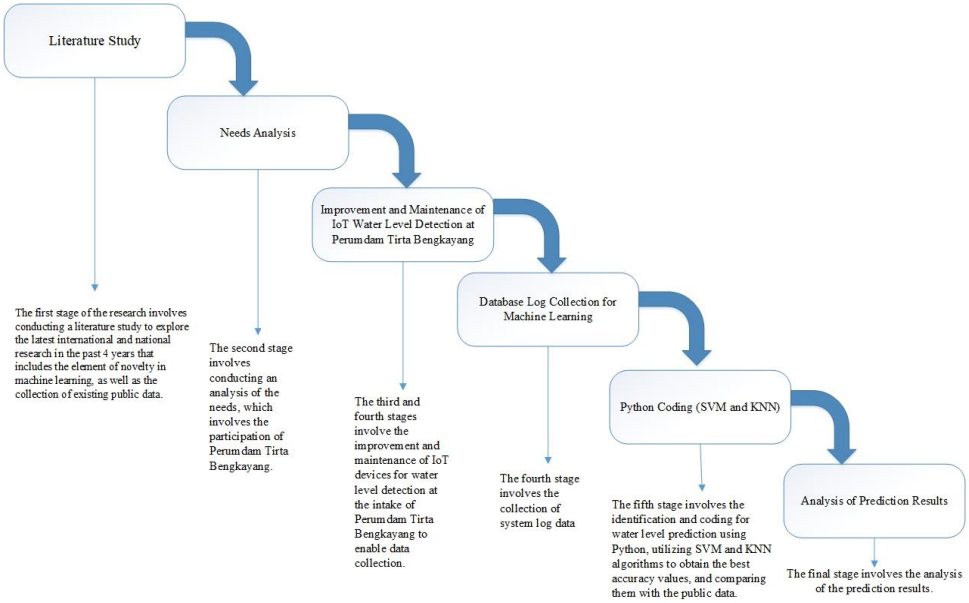

Figure 1. Research Stages

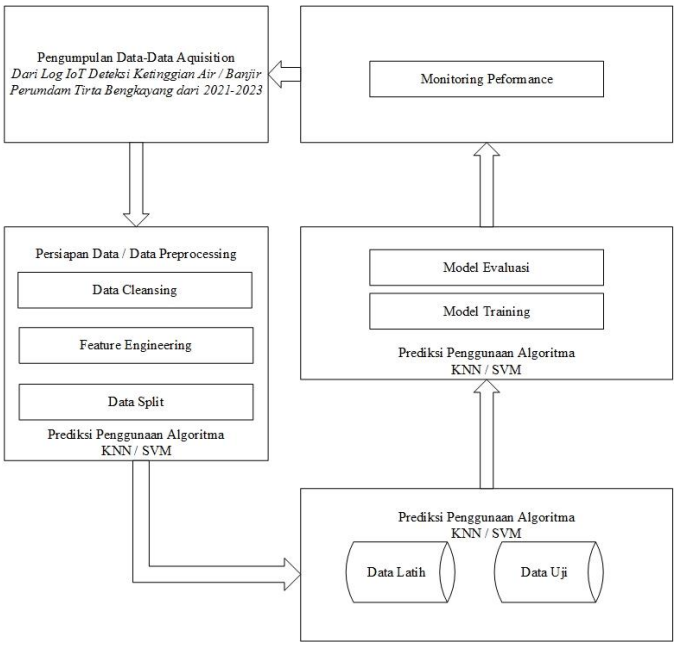

Figure 2. Research Stage Framework Implementation

In Figures 1 and 2, the phases of data gathering, data preprocessing, model training, testing, validation, and performance evaluation in this study are detailed. The data collection phase involves acquiring lower and upper limit logs, sensor logs, flood timing information, and water level readings detected by Arduino via Nodemcu Amica integrated with Ultrasonic sensors and Relay along with Telegram integration [34]. These datasets cover the past year to forecast abnormal water levels or potential floods in the upcoming months. Subsequently, tasks such as data cleaning, paramount feature engineering, and dataset partitioning are executed based on the structure of the data processing pipeline [35]. During data cleaning, the focus is on handling irrelevant entries and anomalies at a record level [36]. Feature engineering is conducted to enhance column quality by addressing inconsistencies in accessing activities within environmentally friendly locations. Data splitting involves dividing datasets into training [37]. The primary focus of this research entails implementing SVM,KNN algorithms later sections will delve into more details regarding these techniques through models derived from hypothesis-driven without designing computational solutions [23].

SVM

The Support Vector Machine algorithm utilizes parameters C (cost) and kernel. The next step involves determining the optimal parameter values that produce the best results, followed by a comparison of which variables yield the most accurate prediction results [38]. SVM is a relatively recent technique but has exhibited superior performance compared to others, particularly in text classification and handwriting recognition [23]. The concept of SVM originates from the classification problem involving two classes: positive and negative. This method aims to identify an optimal separator to maximize the margin between these two classes [23]. In cases where linear SVM is unable to classify the data, kernel functions have been developed for classifying non-linear data [39]. Additionally, Sequential Training presents as a simpler and faster algorithm option.The steps of the Sequential Training algorithm are as follows:

1. Initialize the parameters, for example,

$$\lambda = 0.5, \gamma = 0.01 \quad (1)$$

c = 1, IterationMax = 100, and ε = 0.001. Then calculate the Hessian matrix using equation (2).

$$Dij = yiyj(K(xi, xj) + \lambda \quad (2)$$

2. Start from data i to j, calculate using equations (3)(4)(5):

a. $Ei = \sum n\ aj\ Dij$ (3)
b. $\delta ai = \min\{\max[\gamma(1 - Ei), -ai], c - ai\}$ (4)
c. $ai = ai + \delta ai$ (5)

3. Repeat step 2 until the maximum iteration condition is reached. Then, the support vector (SV) is obtained, SV = (Threshold SV). This value is obtained through several experiments, usually so far using a Threshold > 0 [40]. Next, the testing process is carried out to make decisions. The decision function can be calculated using equation (6). The sign function yields +1 if the argument > 0, –1 if the argument < 0

$$f(x) = \text{sign}(\sum_i aiyiK(xi,x) + b) \quad (6)$$

The parameters used in the Support Vector Machine algorithm are C (cost) and kernel. The next step is to find the parameter values that yield the best results by fine-tuning and optimization methods [1]. After that, we compare which variables produce the best prediction results based on these optimized parameters. SVM is a relatively new technique but it has shown better performance compared to others, especially in text classification and handwriting recognition due to its effective handling of high-dimensional data [41]. The concept of SVM begins with the classification problem of two classes, positive and negative [42]. This method aims to find the best separator to maximize the margin between the two classes, thus improving generalization capability for unseen data instances as well [43]. In some cases,

linear SVM cannot classify all types of data effectively; hence kernel functions are developed involving non-linear transformations such as polynomial or radial basis function kernels which allows for improved representation capacity when classifying complex patterns present in real-world datasets [44]. Sequential Training is an algorithmic approach simplifies training procedures making it more computationally efficient especially when dealing with large-scale datasets consisting millions records along with thousands features [45].

KNN

The precision of this algorithm can be significantly affected by the existence of irrelevant characteristics if their weights are unequal or unimportant for the classification [46]. Many research papers focusing on this algorithm emphasize techniques for enhancing its performance in classification through feature selection and weighting. The K-NN approach involves identifying the closest distance between the data to be assessed and the K nearest neighbors in the training set, which is represented in a multi-dimensional space with each dimension corresponding to a data feature [47]. The steps to calculate K-NN are as follows:The K-nearest neighbors algorithm calculates the distance between the data point to be classified and its nearest neighbors in the training set. These neighbors are identified in a multi-dimensional space where each dimension corresponds to a feature of the data [37]. The success of this algorithm can be significantly influenced by the presence of irrelevant features, especially if their weights are unequal or unimportant for classification purposes [48].

Several research papers have emphasized techniques to enhance the performance of KNN through feature selection and weighting [11]. Moving forward, the collaborative research efforts between the Shanti Bhuana Institute and Perumdam Tirta Bengkayang have paved the way for groundbreaking advancements in machine learning algorithms combined with IoT for predicting, classifying, and mitigating risks associated with the increasing demand for water resources. The 5 stage research, encompassing a literature study, needs analysis, IoT device maintenance and configuration, and water level prediction using Python and advanced machine learning algorithms, has yielded invaluable insights and practical solutions for flood prediction and water resource management at Intake Perumdam Tirta Bengkayang.

As the next phase of the research roadmap from 2023 to 2028 approaches, the focus will be on further advancing new machine learning algorithms integrated with IoT technologies to address the expanding demands on water resources and contribute to the effective management of water supply to consumers [33]. The utilization of machine learning and IoT technologies has proven to be a highly effective and innovative approach in strengthening infrastructure capacity and improving the community's resilience in responding to natural disasters [49]. In Figures 1 and 2, the phases of data gathering, data preprocessing, model training, testing, validation, and performance evaluation in this study are detailed. The data collection phase involves acquiring lower and upper limit logs, sensor logs, flood timing information, and water level readings detected by Arduino via Nodemcu Amica integrated with Ultrasonic sensors and Relay, along with Telegram integration [7]. These datasets cover the past year to forecast abnormal water levels or potential floods in the upcoming months. Subsequently, tasks such as data cleaning, feature engineering, and dataset partitioning are executed based on the structure of the data processing pipeline.

During data cleaning, the focus is on handling irrelevant entries and anomalies at a record level [50]. Feature engineering is conducted to enhance column quality by addressing inconsistencies in accessing activities within environmentally friendly locations. Data splitting involves dividing datasets into training and testing sets, setting the stage for the implementation of SVM and KNN algorithms [51]. The Support Vector Machine algorithm utilizes parameters such as cost (C) and kernel, and the next step involves determining the optimal parameter values that produce the best results, followed by a comparison of which variables yield the most accurate prediction results. SVM has demonstrated superior performance in text classification and handwriting recognition, with its ability to effectively handle high-dimensional data [23].

The concept of SVM aims to identify an optimal separator to maximize the margin between two classes, improving generalization capability for unseen data instances. In cases where linear SVM is unable to classify the data, kernel functions have been developed for classifying non-linear data [52]. The

next steps in the Sequential Training algorithm involve initializing the parameters, calculating the Hessian matrix, and iterating through the data to obtain the support vector. After several experiments, the testing process is performed to make decisions, and the decision function can be calculated using certain parameters. After fine-tuning and optimization, the optimized parameters are compared to determine which variables produce the best prediction results [53]. Moreover, the KNN approach involves identifying the closest distance between the data to be assessed and the K nearest neighbors in the training set, represented in a multi-dimensional space with each dimension corresponding to a data feature [54].

To improve the algorithm's performance in classification, emphasis is placed on techniques for enhancing its precision through feature selection and weighting [55]. As the research roadmap unfolds, the integration of these advanced machine learning algorithms with IoT technologies is poised to revolutionize the predictive and risk mitigation capabilities in water resource management, contributing to the overall resilience and sustainability of water supply systems [56].

a. Determine the value of K.
b. Calculate the distance using equation (7):

$$D = \sqrt{(x1 - y1)^2 + (x2 - y2)^2} \quad (7)$$

Where: X = sample data Y = test data D = distance

c. Sort the distance results and determine the nearest neighbors based on the minimum distance to -K.
d. Use the majority vote of the nearest neighbors as the predicted value for the new data. To find the predicted value in K-NN, it can be calculated using equation (8):

$$Y\hat{} = K1_{i=1} \sum Ky_i \quad (8)$$

Where: Y = Prediction K = number of nearest neighbors Yi = output of the nearest neighbor. The K-NN method in machine learning is utilized for classification or prediction based on data that closely resembles the evaluated data [57]. In the specific case of flood detection at Intake Perumdam Tirta Bengkayang, the K-NN method predicts potential floods using features such as lower water limit, upper water limit, water sensor limit, flood-safe, flood-alert, flood-emergency, and dates from April 2023 to July 2023. The process for calculating K-NN involves several steps:

1. Determine the Value of K: Choose the number of nearest neighbors for prediction [58].
2. Calculate Distance: Use a formula to calculate distance between test data and training data [59].
3. Sort Distance Results: Arrange distance results from smallest to largest and select the closest neighbors based on minimum distance [59].
4. Utilize Majority Vote: Employ majority vote of nearest neighbors as predicted value for new data using equation [60].

The implementation of Python can be used for applying K-NN in flood detection at Intake Perumdam Tirta Bengkayang. After extensively studying the methodologies and processes involved in the research, it is clear that the integration of advanced machine learning algorithms with IoT technologies has paved the way for revolutionary advancements in water resource management [20]. The utilization of the Support Vector Machine algorithm, with its ability to handle high-dimensional data and effectively classify non-linear data through kernel functions, has proven to be a crucial component in the prediction and mitigation of potential floods [61]. Furthermore, the K-Nearest Neighbors method, with its ability to predict potential floods based on features such as lower water limit, upper water limit, water sensor limit, and various flood-related parameters, has provided invaluable insights for flood detection at Intake Perumdam Tirta Bengkayang [32]. The application of KNN involves the determination of the value of K, calculation of distances using a specific formula, sorting of distance results to determine the nearest neighbors, and utilization of the majority vote to predict the value for new data [62]. The KNN method, implemented through Python, can be utilized to analyze and predict potential floods at Intake Perumdam Tirta Bengkayang. When applied effectively, this method is expected to further enhance the accuracy and efficiency of flood prediction, thereby contributing to the overall resilience and sustainability of water

supply systems [63]. In the upcoming phases of the research roadmap from 2023 to 2028, it is crucial to focus on further advancing new machine learning algorithms integrated with IoT technologies to address the expanding demands on water resources [64]. By leveraging the power of advanced algorithms and cutting-edge technologies, the research aims to contribute significantly to the effective management of water supply to consumers and strengthen infrastructure capacity in responding to natural disasters [65].

RESULTS AND DISCUSSION

The manual calculation for the SVM involves using training data with six real data, each having specific attributes and labels. The training include [0.5, 0.8, 0.2] with label 1, [0.3, 0.6, 0.4] with label 0, [0.7, 0.2, 0.9] with label 1, [0.2, 0.5, 0.3] with label 0, [0.6, 0.9, 0.4] with label 1, and [0.1, 0.3, 0.2] with label 0. The test data includes three : [0.6, 0.4, 0.7] with label 1, [0.3, 0.5, 0.2] with label 0, and [0.4, 0.8, 0.3] with label 0. First, the SVM model is trained using the training data with an RBF kernel. After training, the model is used to predict the test data. The predictions obtained are as follows: for test data [0.6, 0.4, 0.7], the prediction is 1, which is correct. For test data [0.3, 0.5, 0.2], the prediction is 0, which is correct. For test data [0.4, 0.8, 0.3], the prediction is 0, which is correct. Accuracy is calculated as the ratio of the number of correct predictions to the total number of data points. In this case, all predictions are correct, resulting in an accuracy of:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total data}} = \frac{3}{3} = 1.00 = 100\%$$

The data was obtained from descriptions acquired from the old model IoT system as shown in Figure 3, which was subsequently refined for the machine learning processing depicted in Figure 4. Both figures represent the management of raw data collected from IoT devices used at the Perumdam Tirta Bengkayang intake, with details presented in Table 1.

Figure 3. Detection Results Real Data Daily in IoT Intake Perumdam Tirta Bengkayang

Figure 4. Recorded Management Results for Water Limit Data (data transfer from the IoT system)

Table 1. Most Frequently Occurring in IoT Testing (Apr-July)

Date	Lower Water Limit (cm)	Upper Water Limit (cm)	Water Sensor Limit (cm)	Flood Safe Status	Flood Alert Status	Flood Emergency Status
06-Apr	0	50	4	No	Yes	No
07-Apr	0	76	4	No	Yes	No
25-Apr	0	50	4	No	Yes	No
06-May	0	90	4	No	Yes	No
07-May	0	80	4	No	Yes	No
25-May	0	50	4	No	Yes	No
06-Jun	0	55	4	No	Yes	No
07-Jun	0	58	4	No	Yes	No
25-Jun	0	63	4	No	Yes	No
06-Jul	0	89	4	No	Yes	No
07-Jul	0	51	4	No	Yes	No
25-Jul	0	50	4	No	Yes	No

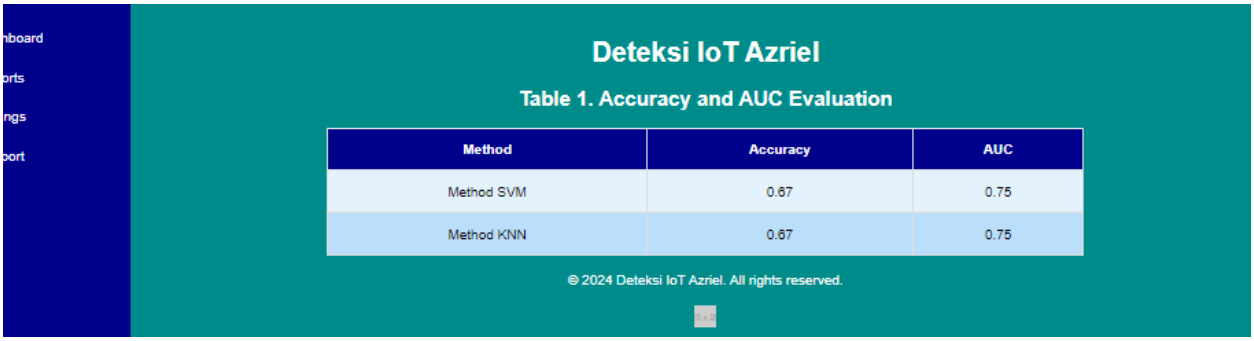

Figure 5. Accuracy and AUC Results from the System

Subsequently, the data processed in Table 1 yields values that meet the criteria shown in Figure 5. The final results display the total AUC and accuracy. However, based on more realistic script results, for test data [0.6, 0.4, 0.7], the prediction is 1 (correct). For test data [0.3, 0.5, 0.2], the prediction is 0 (correct). For test data [0.4, 0.8, 0.3], the prediction is 1 (incorrect). Therefore, the number of correct predictions is 2 out of 3, resulting in an accuracy of:

$$Accuracy = \frac{2}{3} = 0.63 = 67\%$$

Next, the AUC (Area Under the Curve) is calculated. The prediction probabilities for the test data are: for test data [0.6, 0.4, 0.7], the probability is 0.8. For test data [0.3, 0.5, 0.2], the probability is 0.3. For test data [0.4, 0.8, 0.3], the probability is 0.2. True Positive Rate (TPR) or Sensitivity is calculated as the number of correct positive predictions divided by the total number of positive cases:

$$TPR = \frac{\text{Number of incorrect positive predictions}}{\text{Total positif cases}} = \frac{1}{1} = 1.00$$

False Positive Rate (FPR) or 1-Specificity is calculated as the number of incorrect positive predictions divided by the total number of negative cases:

$$FPR = \frac{\text{Number of incorrect positive predictions}}{\text{Total negative cases}} = \frac{1}{2} = 0.50$$

AUC is calculated by summing TPR and FPR and then dividing by two:

$$AUC = \frac{\text{TPR} + \text{FPR}}{2} = \frac{1.00 + 0.50}{2} = 0.75$$

The manual calculation for K-NN involves using training data consisting of three with specific attributes and labels. The training include {'lower_limit_water': 10, 'upper_limit_water': 20, 'water_sensor_limit': 15, 'flood_safe': True}, {'lower_limit_water': 5, 'upper_limit_water': 25, 'water_sensor_limit': 18,

'flood_safe': False}, and {'lower_limit_water': 15, 'upper_limit_water': 30, 'water_sensor_limit': 25, 'flood_safe': False}. The test data is {'lower_limit_water': 12, 'upper_limit_water': 22, 'water_sensor_limit': 17}. The first step is to calculate the Euclidean distance between the test data and each training. The distance to the first training is calculated as:

$$\sqrt{(12-10)^2+(22-20)^2+(17-15)^2} = 3.46$$

The distance to the second training is calculated as:

$$\sqrt{(12-5)^2+(22-25)^2+(17-18)^2} = 7.68$$

The distance to the third training is calculated as:

$$\sqrt{(12-5)^2+(22-30)^2+(17-25)^2} = 11.70$$

The three closest neighbors are the first training, the second training, and the third training with distances of 3.46, 7.68, and 11.70 respectively. The number of neighbors that are 'flood_safe' is one (the first training), while the other two are not 'flood_safe' (the second and third training). The prediction for the test data is that there is no flood because the majority of the neighbors are not 'flood_safe'. Next, the AUC is calculated. The prediction probability for the test data is 33% 'flood_safe' and 67% not 'flood_safe'. TPR is calculated as:

$$TPR = \frac{\text{Number of Correct Positive Predictions}}{\text{Total Positive Cases}} = \frac{1}{1} = 1.00$$

FPR is calculated as:

$$FPR = \frac{\text{Number of Incorrect Positive Predictions}}{\text{Total Negative Cases}} = \frac{1}{2} = 0.50$$

AUC is calculated by summing TPR and FPR and then dividing by two:

$$AUC = \frac{\text{TPR} + \text{FPR}}{2} = \frac{1.00 + 0.50}{2} = 0.75$$

The evaluation results show that the SVM method has an accuracy of 0.67 and an AUC of 0.75, while the K-NN method has an accuracy of 0.67 and an AUC of 0.75. These manual calculation results for accuracy and AUC for both the SVM and K-NN methods are consistent with the results obtained from the Python script. This ensures that the models used are reliable for flood detection at Intake Perumdam Tirta Bengkayang. Results of this data are compared with the processing of 2 other datasets on floods in West Kalimantan from the open data analytics room as follows:

Table 2. Accuracy Test and AUC of Open Data West Kalimantan Flood 1 2022/2023

Method	Accuracy	AUC
SVM Method	0.82	0.76
KNN Method	0.78	0.71

Table 3. Accuracy Test and AUC of Open Data West Kalimantan Flood 2 2023/2024

Method	Accuracy	AUC
SVM Method	0.85	0.79
KNN Method	0.81	0.73

The SVM method has higher accuracy than the KNN method in both datasets. However, the difference may not be significant. Similarly, the SVM method also has a higher AUC than the KNN method in both datasets. However, it was still testing the model using the K-Fold Cross Validation method. With Google Colab Premium, table 1 data will be processed by K-Fold Cross Validation, and this result will show as table 4 below.

Table 4. Accuracy Test and AUC with K-Fold Cross Validation

K	SVM Method		KNN Method	
	Accuracy	AUC	Accuracy	AUC
2	0.6333	0.6721	0.6333	0.6473
3	0.6333	0.6537	0.6	0.5130
4	0.6384	0.65	0.7054	0.5797
Average	0.635	0.6586	0.6462	0.58

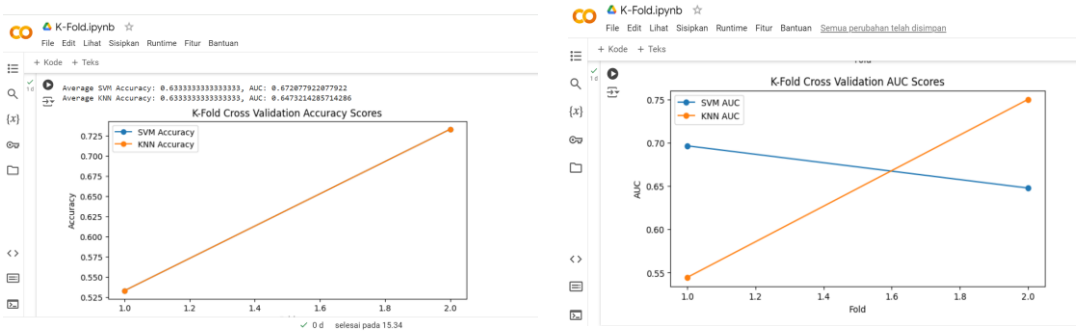

Figure 6. Accuracy and AUC on K-Fold Cross Validation with K=2

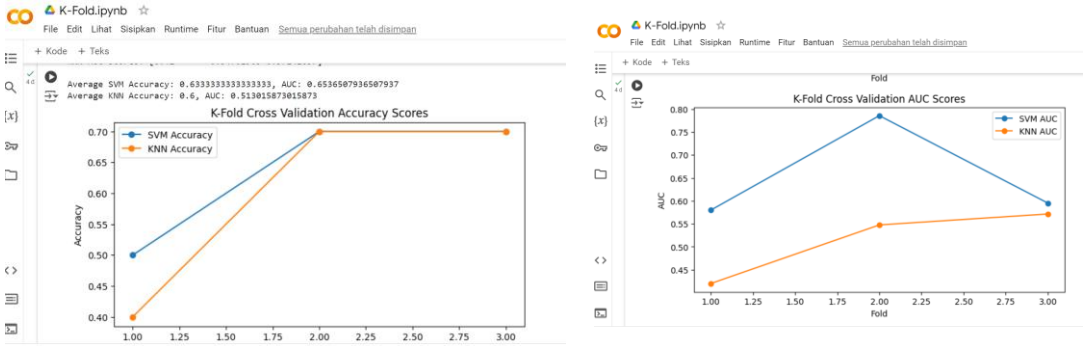

Figure 7. Accuracy and AUC on K-Fold Cross Validation with K=3

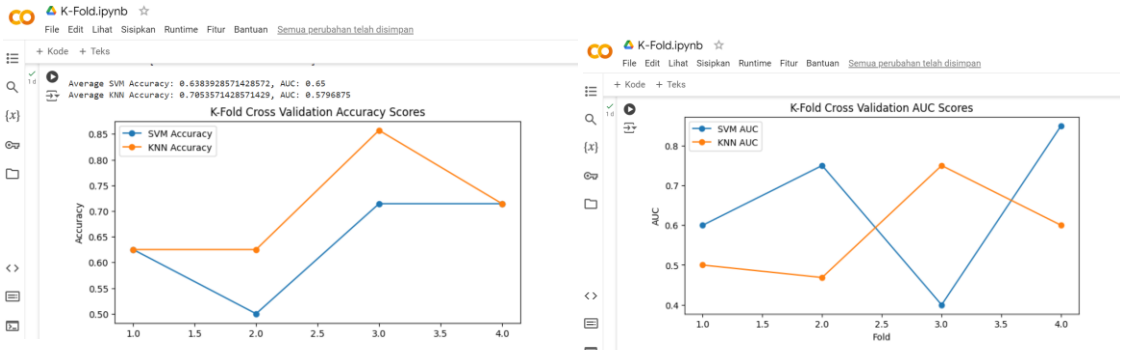

Figure 8. Accuracy and AUC on K-Fold Cross Validation with K=4

Using table 4 of the results from K-Fold Cross Validation, compared to manual calculation, it is concluded that the difference points from Accuracy by SVM Method are 0.635 and 0.67, and by KNN Method are 0.6586 and 0.67. Thus, the difference points from AUC by SVM Method are 0.6462 and 0.75, and by KNN method are 0.58 and 0.75. Such differences might occur because manual calculations are focused on 1 fold, whereas K-Fold Cross Validation is the average result of all folds tested. This accuracy value is still

lower than the public data in tables 2 and 3 of the Data Analytic Room of the Open Data application published by the Government of Bengkayang and West Kalimantan, as the public data accuracy has been running for periods 1 and 2 involving all regions in the district, whereas this only focuses on one district in Perumdam Tirta Bengkayang.

CONCLUSION

The incorporation of Machine Learning and IoT technology at Intake Perumdam Tirta Bengkayang has showcased substantial progress in flood prediction skills. The system employs a network of strategically positioned IoT sensors to constantly monitor essential data points, including water levels and rainfall. The data is further examined using advanced machine learning techniques such as SVM (Support Vector Machine) and KNN (K-Nearest Neighbors), which have been trained on historical data to identify patterns that suggest the occurrence of prospective flood occurrences. The integration of these technologies enables instantaneous data processing, allowing the prediction system to offer advance warnings, which are crucial for implementing proactive steps to safeguard the community and infrastructure against flood consequences. This early warning capability enables preventive measures such as strengthening infrastructure, evacuating vulnerable locations, and mobilizing resources with efficiency. The effective implementation of this approach not only enhances the accuracy of flood prediction but also improves the management of water resources as a whole. The real-time analytics offer important insights into patterns of water usage, forecasting of demand, and preemptive maintenance, resulting in improved operational efficiencies. The partnership between the Shanti Bhuana Institute and Perumdam Tirta Bengkayang has facilitated progress in combining machine learning with IoT to develop more resilient and flexible flood prediction systems.

ACKNOWLEDGEMENT

We thank the Ministry of Education and Culture for the funds provided in the form of a Beginner Lecturer Research grant (PDP) in 2023. We also thank the LLDIKTI 11 (Kalimantan) who gave the research contract. We also thank the PRPM Shanti Bhuana Institute for facilitating this research and also Perumdam Tirta Bengkayang .

REFERENCES

[1] Indonesian Agency for Meteorology, Climatology, and Geophysics. Online at https://www.bmkg.go.id/ accessed 6 February 2024

[2] Rustinsyah, R., Prasetyo, R A., & Adib, M. (2021, January 1). Social capital for flood disaster management: Case study of flooding in a village of Bengawan Solo Riverbank, Tuban, East Java Province. https://doi.org/10.1016/j.ijdrr.2020.101963

[3] Zarei, M., Bozorg-Haddad, O., Baghban, S., Delpasand, M., Goharian, E., & Loáiciga, H A. (2021, December 21). Machine-learning algorithms for forecast-informed reservoir operation (FIRO) to reduce flood damages. https://doi.org/10.1038/s41598-021-03699-6

[4] Hadi, M I., Yakub, F., Fakhrurradzi, A., Hui, C X., Najiha, A., Fakharulrazi, N A., Harun, A N., Rahim, Z A., & Azizan, A. (2020, June 1). Designing Early Warning Flood Detection and Monitoring System via IoT. https://doi.org/10.1088/1755-1315/479/1/012016

[5] Zhang, Z., Liang, J., Zhou, Y., Huang, Z., Jiang, J., Liu, J., & Yang, L. (2022, February 4). A Multi-strategy-mode-waterlogging-prediction Framework for Urban Flood Depth. https://nhess.copernicus.org/preprints/nhess-2022-36/nhess-2022-36.pdf

[6] Yang, S., & Chang, L. (2020, May 31). Regional Inundation Forecasting Using Machine Learning Techniques with the Internet of Things. https://doi.org/10.3390/w12061578

[7] Okuhara, H., Elnaqib, A., Dazzi, M., Palestri, P., Benatti, S., Benini, L., & Rossi, D. (2021, October 1). A Fully Integrated 5-mW, 0.8-Gbps Energy-Efficient Chip-to-Chip Data Link for Ultralow-Power IoT End-Nodes in 65-nm CMOS. https://doi.org/10.1109/tvlsi.2021.3108806

[8] Zarei *et al.* (2021, December 21). Machine-learning algorithms for forecast-informed reservoir operation (FIRO) to reduce flood damages - Scientific Reports. https://www.nature.com/articles/s41598-021-03699-6

[9] Ivanov, V. (2023, January 19). Flood forecasts in real-time with block-by-block data could save lives. A new machine learning method makes it possible. https://phys.org/news/2023-01-real-time-block-by-block-machine-method.html

[10] Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T., Weitzner, D., & Matias, Y. (2024, March 20). Global prediction of extreme floods in ungauged watersheds. https://doi.org/10.1038/s41586-024-07145-1
[11] Mankali, L., Rangarajan, N., Chatterjee, S., Kumar, S., Chauhan, Y S., Sinanoglu, O., & Amrouch, H. (2022, December 1). Leveraging Ferroelectric Stochasticity and In-Memory Computing for DNN IP Obfuscation. https://doi.org/10.1109/jxcdc.2022.3217043
[12] Peng, X., Zhang, X., Wang, X., Li, H., Xu, J., & Zhao, Z. (2022, December 5). Construction of rice supply chain supervision model driven by blockchain smart contract. https://doi.org/10.1038/s41598-022-25559-7
[13] Carrera, F F., Sanchez, H S., García-Orellana, Y., & Chadrina, O. (2021, January 1). A System for Measuring Water Levels in Open-Air Irrigation Canals. https://doi.org/10.1051/epjconf/202124802011
[14] Nako, E., Toprasertpong, K., Nakane, R., Takenaka, M., & Takagi, S. (2022, June 12). Experimental demonstration of novel scheme of HZO/Si FeFET reservoir computing with parallel data processing for speech recognition. https://doi.org/10.1109/vlsitechnologyandcir46769.2022.9830412
[15] Kazemi, A., Müller, F., Sharifi, M M., Errahmouni, H., Gerlach, G., Kämpfe, T., Imani, M., Hu, X S., & Niemier, M. (2022, November 10). Achieving software-equivalent accuracy for hyperdimensional computing with ferroelectric-based in-memory computing. https://doi.org/10.1038/s41598-022-23116-w
[16] Huang, J., He, D., Obaidat, M S., Vijayakumar, P., Luo, M., & Choo, K R. (2021, April 17). The Application of the Blockchain Technology in Voting Systems. https://doi.org/10.1145/3439725
[17] Nankani, H., Gupta, S., Mondal, S., & Kalaiarasi, S. (2020, May 18). IoT Based Water Monitoring System for Agriculture. https://doi.org/10.36478/aj.2020.37.41
[18] Stewart *et al.* (2023, June 15). Datasheets for Machine Learning Sensors: Towards Transparency, Auditability, and Responsibility for Intelligent Sensing. https://arxiv.org/abs/2306.08848
[19] Prakasam, C., Aravinth, R., Kanwar, V S., & Nagarajan, B. (2021, January 1). Design and Development of Real-time landslide early warning system through low cost soil and rainfall sensors. https://doi.org/10.1016/j.matpr.2021.02.456
[20] Glória, A., Cardoso, J M R., & Sebastião, P. (2021, April 28). Sustainable Irrigation System for Farming Supported by Machine Learning and Real-Time Sensor Data. https://doi.org/10.3390/s21093079
[21] Abioye, A E., Hensel, O., Esau, T., Elijah, O., Abidin, M S Z., Ayobami, A S., Yerima, O., & Nasirahmadi, A. (2022, February 1). Precision Irrigation Management Using Machine Learning and Digital Farming Solutions. https://doi.org/10.3390/agriengineering4010006
[22] Wisudawan, H N P. (2021, October 22). Design and Implementation of Real-Time Flood Early Warning System (FEWS) Based on IoT Blynk Application. https://doi.org/10.26418/elkha.v13i2.49003
[23] Su, L., Wen-hua, B., Zhu, Z., & He, X. (2021, September 1). Research on Application of Support Vector Machine in Intrusion Detection. https://doi.org/10.1088/1742-6596/2037/1/012074
[24] Wang, Q., & Abdelrahman, W. (2023, March 13). High-Precision AI-Enabled Flood Prediction Integrating Local Sensor Data and 3rd Party Weather Forecast. https://doi.org/10.3390/s23063065
[25] Muhadi, N A., Abdullah, A F., Bejo, S K., Mahadi, M R., & Mijić, A. (2020, July 18). The Use of LiDAR-Derived DEM in Flood Applications: A Review. https://doi.org/10.3390/rs12142308
[26] Khan, W., Hussain, A., Alaskar, H., Baker, T., Ghali, F., A-Jumeily, D., & Al-Shamma'a, A I. (2020, December 1). Prediction of Flood Severity Level via Processing IoT Sensor Data Using a Data Science Approach. https://doi.org/10.1109/iotm.0001.1900110
[27] Algiriyage, N., Prasanna, R., Stock, K., Doyle, E E., & Johnston, D. (2021, November 27). Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review. https://doi.org/10.1007/s42979-021-00971-4
[28] Jatnika, H., Purwanto, Y S., Rifai, M F., & Silaen, R H. (2021, June 28). Web-Based Automated Rainwater Storage and Water Quality Monitoring Design Using K-Nearest Neighbor Method. https://doi.org/10.37339/e-komtek.v5i1.551
[29] Studer, S., Bui, T B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. (2021, April 22). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. https://doi.org/10.3390/make3020020
[30] Najafi, H., Shrestha, P K., Rakovec, O., Apel, H., Vorogushyn, S., Kumar, R., Thober, S., Merz, B., & Samaniego, L. (2024, May 2). High-resolution impact-based early warning system for riverine flooding. https://doi.org/10.1038/s41467-024-48065-y
[31] Yin X Z, Yue J S, Huang Q R, et al. Computing-in-memory circuits and cross-layer integrated design and optimization: from SRAM to FeFET (in Chinese). Sci Sin Inform, 2022, 52: 612{638, doi: 10.1360/SSI-2021-0420
[32] Nurcahyo, A C., Himamunanto, A R., & Firgia, L. (2022, June 3). Network Infrastructure Design and Website Management of Perumdam Tirta Bengkayang. https://doi.org/10.26760/rekaelkomika.v3i2.124-133
[33] Singh, M., & Ahmed, S. (2021, January 1). IoT based smart water management systems: A systematic review. https://doi.org/10.1016/j.matpr.2020.08.588
[34] Ritzkal, -., Afrianto, Y., Riawan, I., Kusumah, F S F., & Remawati, D. (2023, January 1). Water Tank Wudhu and Monitoring System Design using Arduino and Telegram. https://doi.org/10.14569/ijacsa.2023.0140159

[35] Steorts, Rebecca C. (2023, July 25). A Primer on the Data Cleaning Pipeline. https://arxiv.org/abs/2307.13219
[36] Davies, N. (2023, August 7). *The importance of data cleaning in data science*. KDnuggets. Retrieved from https://www.kdnuggets.com/2023/08/importance-data-cleaning-data-science.html
[37] Cacciarelli, D., & Külahçı, M. (2022, July 1). A novel fault detection and diagnosis approach based on orthogonal autoencoders. https://doi.org/10.1016/j.compchemeng.2022.107853
[38] Hossain, S M., & Ayub, M A. (2020, December 19). Parameter Optimization of Classification Techniques for PDF based Malware Detection. https://doi.org/10.1109/iccit51783.2020.9392685
[39] Zahra, H S F M. (2022, October 16). A new trigonometric kernel function for support vector machine. https://arxiv.org/abs/2210.08585
[40] Barkam, H E., Yun, S., Genßler, P R., Zou, Z., Liu, C., Amrouch, H., & Imani, M. (2023, April 1). HDGIM: Hyperdimensional Genome Sequence Matching on Unreliable highly scaled FeFET. https://doi.org/10.23919/date56975.2023.10137331
[41] Yang, C., Wu, Q., Lu, J., & Chen, H. (2021, September 1). The Research Progress of the Deep Hybrid Model in the Field of Text Classification. https://doi.org/10.1088/1742-6596/2010/1/012041
[42] Guide to Support Vector Machine (SVM) Algorithm. (2022, October 18). https://serokell.io/blog/support-vector-machine-algorithm
[43] Qian *et al*. (2022, April 29). On the Optimization of Margin Distribution. https://arxiv.org/abs/2204.14118
[44] Bell, B., Geyer, M., Glickenstein, D., Fernandez, A., & Moore, J E. (2023, January 1). An Exact Kernel Equivalence for Finite Classification Models. https://doi.org/10.48550/arxiv.2308.00824
[45] Shim, W., & Yu, S. (2021, June 1). Ferroelectric Field-Effect Transistor-Based 3-D NAND Architecture for Energy-Efficient on-Chip Training Accelerator. https://doi.org/10.1109/jxcdc.2021.3057856
[46] Zhang, L. (2021, December 28). A Feature Selection Algorithm Integrating Maximum Classification Information and Minimum Interaction Feature Dependency Information. https://doi.org/10.1155/2021/3569632
[47] Chen, Y., Ruys, W., & Biros, G. (2020, January 1). KNN-DBSCAN: a DBSCAN in high dimensions. https://doi.org/10.48550/arxiv.2009.04552
[48] University, F K S. (2021, March 1). Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://dl.acm.org/doi/abs/10.1145/3442188.3445883
[49] Ikhvan, A., & Mera, M. (2021, April 1). Case Study: Significant factors in hazard and vulnerability assessments in flood mitigation in Padang City. https://doi.org/10.1088/1755-1315/708/1/012026
[50] Data Cleansing Services | Data Scrubbing Services | Data Clensing. (2023, January 1). https://www.spanglobalservices.com/data-cleansing
[51] Customers demand increasing sustainability efforts at hotels and meeting venues. (2023, July 21). https://www.traveldailymedia.com/customers-demand-increasing-sustainability-efforts-at-hotels-and-meeting-venues/
[52] Kwon, D., Lim, S., Bae, J., Lee, S., Kim, H., Seo, Y., Oh, S., Kim, J., Yeom, K., Park, B., & Lee, J. (2020, July 7). On-Chip Training Spiking Neural Networks Using Approximated Backpropagation With Analog Synaptic Devices. https://doi.org/10.3389/fnins.2020.00423
[53] Babnik, Ž., Damer, N., & Štruc, V. (2023, April 19). Optimization-Based Improvement of Face Image Quality Assessment Techniques. https://doi.org/10.1109/iwbf57495.2023.10157796
[54] Blelloch, G E., & Dobson, M. (2022, January 1). Parallel Nearest Neighbors in Low Dimensions with Batch Updates. https://doi.org/10.1137/1.9781611977042.16
[55] Wang, Y., Ding, A., Guan, K., Wu, S., & Du, Y. (2021, January 1). Graph-based Ensemble Machine Learning for Student Performance Prediction. https://doi.org/10.48550/arXiv.2112.
[56] Saleem, A., Mahmood, I., Sarjoughian, H S., Nasir, H A., & Malik, A W. (2021, January 31). A Water Evaluation and Planning-based framework for the long-term prediction of urban water demand and supply. https://doi.org/10.1177/0037549720984250
[57] Sadrabadi, A N., Znjirchi, S M., Abadi, H Z A., & Hajimoradi, A. (2020, December 23). An optimized K-Nearest Neighbor algorithm based on Dynamic Distance approach. https://doi.org/10.1109/icspis51611.2020.9349582
[58] Uddin, S., Haque, I., Lu, H., Moni, M A., & Gide, E. (2022, April 15). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. https://doi.org/10.1038/s41598-022-10358-x
[59] Nuranisah., Efendi, S., & Sihombing, P. (2020, January 1). Analysis of algorithm support vector machine learning and k-nearest neighbor in data accuracy. https://doi.org/10.1088/1757-899x/725/1/012118
[60] K-Nearest Neighbors (KNN) Algorithm Tutorial — Machine Learning Basics. (2021, April 2). https://pub.towardsai.net/k-nearest-neighbors-knn-algorithm-tutorial-machine-learning-basics-ml-ec6756d3e0ac
[61] Kabir, S R., Patidar, S., Xia, X., Liang, Q., Neal, J., & Pender, G. (2020, November 1). A deep convolutional neural network model for rapid prediction of fluvial flood inundation. https://doi.org/10.1016/j.jhydrol.2020.125481
[62] Biswas, A., Morozovska, A N., Ziatdinov, M., Eliseev, E A., & Kalinin, S V. (2021, November 28). Multi-objective Bayesian optimization of ferroelectric materials with interfacial control for memory and energy storage applications. https://doi.org/10.1063/5.0068903

[63] Jia, Z., Guan, Z., Liu, Z., & Yang, D. (2020, May 20). Influence of short-term rainfall forecast error on flood forecast operation: A risk assessment based on Bayesian theory. https://doi.org/10.1080/10807039.2020.1768360
[64] Almadani, M., & Kheimi, M. (2023, February 1). Stacking Artificial Intelligence Models for Predicting Water Quality Parameters in Rivers. https://doi.org/10.12911/22998993/156663
[65] Laufs, J., & Waseem, Z. (2020, December 1). Policing in pandemics: A systematic review and best practices for police response to COVID-19. https://doi.org/10.1016/j.ijdrr.2020.101812