International Journal of Integrated Science and Technology (IJIST)
Vol.2, No.5, 2024: 507-515

Classification of Drinking Water Potability With Artificial Neural Network Algorithm

Indra Darmawan[1*], Muhammad Fatchan[2] Andri Firmansyah[3]
Universitas Pelita Bangsa

Corresponding Author: Indra Darmawan; indradarmawan606@gmail.com

A R T I C L E I N F O

Keywords: Water Potability, Artificial Neural Network, Machine Learning

Received : 20, March
Revised : 25, April
Accepted: 29, May

A B S T R A C T

Having safe water for consumption is essential for public health in every region. However, water quality is declining in some places, especially to meet human needs for drinking water. There are many efforts to maintain water potability, such as checking to see if there are bacteria or diseases in the water. This research classifies water potability using the Artificial Neural Network method, a technique in the field of machine learning. This research classifies water quality using a python library to analyze data and perform classification. Data is processed through stages such as data cleaning and data division into training and testing. In testing, the data is divided into 20% for testing and 80% for training. The results of the ANN algorithm show 70% accuracy. in conclusion, the ANN model has moderate performance in classifying the feasibility of drinking water. Model improvement is needed to improve accuracy and prediction, including the use of larger and more diverse datasets.

DOI: https://doi.org/10.59890/ijist.v2i5.1874

ISSN : 3026-4685

https://journal.multitechpublisher.com/index.php/ijist

INTRODUCTION

Having safe water for consumption is very important for public health in every area. However, water quality is deteriorating in several places, especially to meet human needs for drinking water. Poor water quality can cause many serious diseases, one of which is diarrhea, which causes approximately 800,000 deaths each year (Said et al., 2022)(Wati, 2020). The Food and Agriculture Organization (FAO) states that the lack of clean drinking water causes 3,800 children to die every day from diseases. Research conducted by the World Health Organization (WHO) shows that 2 billion people are affected by water scarcity every day in 40 countries, and the lack of access to clean drinking water results in the death of approximately 2 million newborns each year (Hardiana Said, 2022)(Generosa Lukhayu Pritalia, 2022).

For the coming years, the World Water Assessment Programme (WWAP), overseen by UNESCO, has planned for clean water conditions. 85% of clean water becomes waste under certain conditions, to meet their daily needs, each person uses one hundred liters of water every day(Riyantoko et al., 2021). There are many efforts to maintain water quality, such as conducting checks to determine if there are bacteria or diseases in the water, so that preventive measures can be taken if the water quality declines(Hartanti & Pradana, 2023). The quality of water can be explained through parameters that encompass the composition of the water itself. These parameters or variables can be utilized to predict water conditions using data classification techniques and machine learning(Mutoffar & Fadillah, 2022).

Machine learning is a part of AI that aims to create artificial intelligence. Machine learning also refers to a branch of computer science that focuses on the use of data and algorithms to understand how humans learn and gradually improve its accuracy(Faiza et al., 2022)(Muniroh & Agus Priatno, 2022). Classification is the process of finding patterns intended to estimate the class of unknown objects. The classification method is used in this writing(Sutisna & Yuniar, 2023). In classification, there are several manual classification methods based on computers that use machine learning, such as Support Vector Machine (SVM), Decision Tree, Naive Bayes, and Artificial Neural Network. These are some examples of machine learning classification techniques that can be used.

Each classification method mentioned above has its own advantages and disadvantages(Septhya et al., 2023). There have been several studies on drinking water classification, specifically the classification of drinking water quality using machine learning applications. The implementation of machine learning methods involves steps such as collecting water quality data, data processing, selecting relevant attributes, choosing machine learning algorithms, training the model using training data, and testing the model(Jesika et al., 2023). Therefore, the aim of this study is to classify water quality with machine learning that enhances the modeling process using Artificial Neural Network (ANN). This study uses Python libraries to analyze data during the classification process.

LITERATURE REVIEW

Neural Network

A system for information processing with performance characteristics similar to biological neural networks is called an artificial neural network (ANN), sometimes abbreviated as neural network (NN). An artificial neural network is a component of soft computing, which emphasizes the importance of using methods that prioritize accuracy and certainty. ANNs can be implemented very effectively due to their great potential, equivalent to the high-speed parallelism available. ANNs can be trained and used for natural language processing, image processing, and other tasks. They are generally more flexible and widely applicable with outstanding features such as adaptability, self-learning, fault tolerance, and the ability to perform nonlinear modeling without prior knowledge of the relationship between independent and dependent variables(Dastres & Soori, 2021).

There are various reasons why researchers from different domains are interested in artificial neural networks. Neural networks are typically used in signal processing and control theory in the field of electrical engineering. In computer science, neural networks have the potential to be efficiently implemented in hardware and by applications for robots. Many NN models have been used in classification, but only in the field of health, for the classification or diagnosis of diseases such as heart disease, various types of cancers, diabetes, and so on (Dhoriva Urwatul, 2023).

Klasifikasi

Classification methods are techniques or approaches applied to group or classify data according to patterns, attributes, or characteristics present in the relevant data. The goal of classification is to develop models that can identify and predict the class or label of newly gathered data based on previously collected information. These classification methods aim to streamline the machine learning process by using models or algorithms to learn from labeled training data, which consists of examples already classified. Subsequently, these models are used to classify data with unknown labels. Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, and Artificial Neural Network (ANN) are some commonly used methods in classification(Wibawa et al., 2021).

METHODS

This research is based on data obtained from Kaggle using the term "water potability" to predict whether water is suitable for drinking, utilizing a CSV file format processed using machine learning techniques. The algorithm employed is Artificial Neural Network. The dataset includes several parameters.

Table 1. Water Quality Parameter Dataset"

Parameter	Description

pH (power of Hydrogen)	Measuring the acidity or alkalinity of water.
Hardness	Measuring the content of calcium and magnesium minerals in water.
Solids	Measuring the amount of dissolved and suspended solids in water.
Chloramines	Compounds formed from the reaction of chlorine with ammonia in water.
Sulfate	Measuring the concentration of sulfate ions in water.
Conductivity	Measuring the water's ability to conduct electricity.
Organic Carbon	Measuring the amount of organic matter in water.
Trihalomethanes	Chemical compounds formed as byproducts of water disinfection with chlorine.
Turbidity	Measuring the clarity of water.
Potability	Indicators of water suitability for consumption and non-consumption.

"This research employs a method, which can be seen in Figure 1.

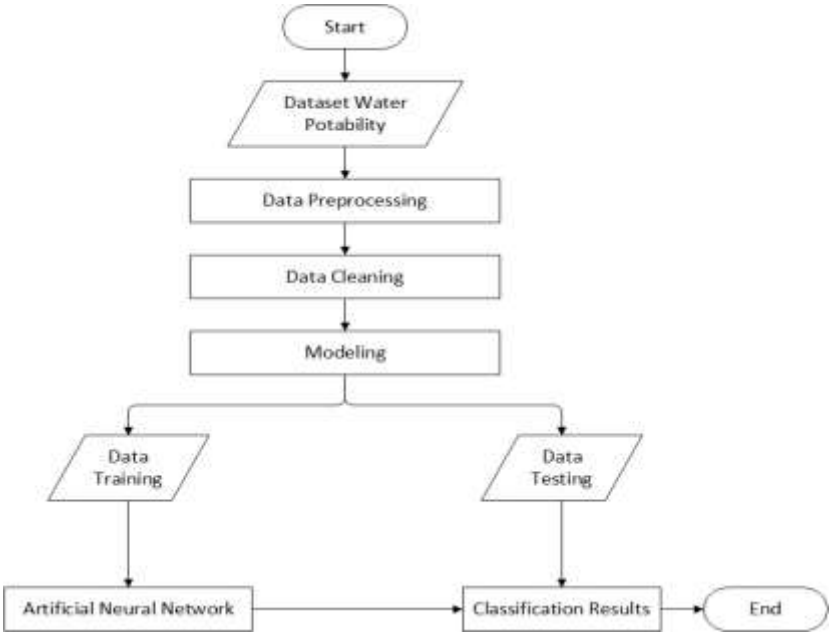

Figure 1. Research Methodology

In Figure 1, first we identify the dataset used, sourced from Kaggle under the name "water potability." The data cleaning stage involves initial data preprocessing tasks, which include selecting and handling missing data. During data cleaning, we check for missing values; if found, these missing values are replaced with the mean value.

The modeling stage entails applying neural network techniques to the training data to uncover patterns or models of knowledge. In this stage, we specifically examine the target variable, "potability." Next, we proceed with training and testing the data to evaluate the model's performance on both datasets. Artificial Neural Network algorithms are applied to derive insights from the training and testing data. For the testing phase, we vary the test size from 20% to 80% to assess the neural network algorithm's outcomes in terms of accuracy, precision, recall, and F1-score.

RESULTS AND DISCUSSION

"Below are the steps in classifying water quality datasets. This study employs a machine learning algorithm, specifically a neural network algorithm. The main objective of this research is to determine the accuracy of the model created using the neural network algorithm.

The dataset used in this program consists of 3276 rows and 10 columns (pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity), along with one labeling attribute (Potability). The data types in this dataset indicate that there are 2 types: float64 and int64.

Table 2. Data Type

Data	Parameter	Non-null count	Type
0	pH	2785 non-null	float64
1	Hardness	3276 non-null	float64
2	Solids	3276 non-null	float64
3	Chloramines	3276 non-null	float64
4	Sulfate	2495 non-null	float64
5	Conductivity	3276 non-null	float64
6	Organic Carbon	3276 non-null	float64
7	Trihalomethanes	3114 non-null	float64
8	Turbidity	3276 non-null	float64
9	Potability	3276 non-null	Int64

The next step is to identify missing values with the aim of determining the number of empty values in the data. In figure 4, it shows that there are missing values in pH, Sulfate, and Trihalomethanes, with 491 missing values for pH, 781 for Sulfate, and 162 for Trihalomethanes.

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

Figure 4. Viewing Missing Value

Next, due to the presence of missing values, the next step in data cleaning is to fill these missing values with the mean. Once the data has been cleaned, it is free from any missing values.

Here are the results of categorizing water quality analyzed through Exploratory Data Analysis (EDA).

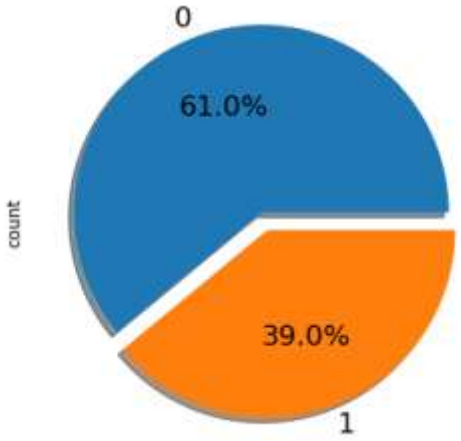

Figure 5. Pie chart of the water potability

From a total of 3276 data, it shows that 61% or 1998 data indicate that water is non-potable, while 39% or 1278 data indicate that water is potability. The water potability classification process involves modeling and data splitting, with 20% of the data reserved for testing and 80% for training. Table 3 shows the classification results of the Neural Network Algorithm.

Table 3. Classification Results

Algoritma	Class	Precision	Recall	F1 Score	Accuracy
ANN	0	0.71	0.87	0.78	0.70
	1	0.65	0.41	0.50	

The table above summarizes key evaluation metrics resulting from classification, including precision, recall, and f1-score. The evaluation results indicate that the ANN model performs well in classification tasks. For the "not potable" class (class 0), the model shows strong performance with a precision of

71% and a recall of 87%, resulting in a relatively high f1-score of 0.78. This indicates that the model is quite reliable in identifying non-potable water. In contrast, for the "potable" class (class 1), the model has a precision of 65% and a lower recall of 41%, yielding an f1-score of 0.50. This suggests that the model struggles to identify potable water accurately. The overall accuracy of the model is 70%, indicating moderate performance overall in classifying water potability.

CONCLUSIONS AND RECOMMENDATIONS

In this study, several conclusions and implementations are drawn from the research findings. However, the model shows lower performance in identifying drinkable water. The overall accuracy of the model is 70%, indicating moderate performance in classifying water potability. The model is more effective in identifying samples of water that are not potable, as evidenced by significant differences in recall and f1-score between classes. The average macro and weighted precision, recall, and f1-score indicate that the model performs balanced overall, despite differing performance between classes.

Based on these research findings, there are several recommendations for further implementation and development. Improving the model is necessary through adjusting parameters and artificial neural network architectures to enhance accuracy and predictions. Utilizing larger and more diverse datasets is crucial to improving model performance. Additional training data and the use of data augmentation techniques to expand existing datasets can help the model learn better and reduce biases. Incorporating additional technologies such as integration with Internet of Things (IoT) devices for real-time water quality monitoring and implementing other machine learning algorithms can complement and validate the results of artificial neural networks.

FURTHER STUDY

This research has several limitations that need to be considered. The size of the dataset used may not be sufficiently large or representative. An imbalanced dataset between samples of potable and non-potable water can also affect the model's performance. The artificial neural network model employed may not be entirely optimal due to suboptimal parameter selection and architecture. Although neural networks are capable of capturing complex patterns, they are also susceptible to overfitting if not properly configured. Moreover, the findings of this study may not directly generalize to all types of water across various geographic locations with differing characteristics, as a model trained on one type of dataset may not perform well on another dataset with a different data distribution.

ACKNOWLEDGMENT

The author wishes to express heartfelt gratitude to colleagues at Universitas Pelita Bangsa who have provided valuable advice and input for this research. Thanks are due to the supervising professors and fellow students who consistently offered support and assistance throughout the research process. Once again, thank you to everyone who has helped and supported me in completing this paper. May the findings of this research contribute positively to the development of drinking water quality monitoring technology.

REFERENCES

Dastres, R., & Soori, M. (2021). Artificial Neural Network Systems. *International Journal of Imaging and Robotics (IJIR)*, *2021*(2), 13–25. https://hal.science/hal-03349542

Dhoriva Urwatul. (2023). *PENERAPAN NEURAL NETWORK UNTUK KLASIFIKASI DAN PERAMALAN TIME SERIES*. https://www.uny.ac.id/id/fokus-kita/prof-dr-dhoriva-urwatul-wustqa-ms_penerapan-neural-network-untuk-klasifikasi-dan

Faiza, I. M., Gunawan, G., & Andriani, W. (2022). Tinjauan Pustaka Sistematis: Penerapan Metode Machine Learning untuk Deteksi Bencana Banjir. *Jurnal Minfo Polgan*, *11*(2), 59–63. https://doi.org/10.33395/jmp.v11i2.11657

Generosa Lukhayu Pritalia. (2022). Analisis Komparatif Algoritme Machine Learning dan Penanganan Imbalanced Data pada Klasifikasi Kualitas Air Layak Minum. *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, *2*(1), 43–55. https://doi.org/10.24002/konstelasi.v2i1.5630

Hardiana Said, N. H. M. H. N. I. (2022). Sistem Prediksi Kualitas Air Yang Dapat Dikonsumsi Dengan Menerapkan Algoritma K-Nearest Neighbor. *Seminar Nasional Mahasiswa Ilmu Komputer Dan Aplikasinya (SENAMIKA)*, *April*, 2962–6129.

Hartanti, D., & Pradana, A. I. (2023). Komparasi Algoritma Machine Learning dalam Identifikasi Kualitas Air. *SMARTICS Journal*, *9*(1), 1–6. https://doi.org/10.21067/smartics.v9i1.8113

Jesika, S., Ramadhani, S., Putri, Y. P., Iskandar, J. W., Medan, P. V, Tuan, S., & Serdang, D. (2023). Implementasi Model Machine Learning dalam Mengklasifikasi Kualitas Air. *Jurnal Ilmiah Dan Karya Mahasiswa*, *1*(6), 382–396. https://doi.org/10.54066/jikma.v1i6.1162

Muniroh, N., & Agus Priatno, E. (2022). PENERAPAN ALGORITMA K-NN PADA MACHINE LEARNING UNTUK KLASIFIKASI KUALITAS AIR BUDIDAYA AKUAPONIK BERBASIS IoT. *Jurnal Teknologi Dan Bisnis*, *4*(2), 73–86. https://doi.org/10.37087/jtb.v4i2.87

Mutoffar, M. M., & Fadillah, A. (2022). Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *4*(2), 138–146. https://doi.org/10.53580/naratif.v4i2.160

Riyantoko, P. A., Fahrudin, T. M., Hindrayani, K. M., Data, S., & Timur, J. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. *Seminar Nasional Sains Data*, *2*(Senada), 12–18.

Said, H., Matondang, N. H., & Irmanda, H. N. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kualitas Air Yang Dapat Dikonsumsi. *Techno.Com*, *21*(2), 256–267. https://doi.org/10.33633/tc.v21i2.5901

Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 15–19. https://doi.org/10.57152/malcom.v3i1.591

Sutisna, & Yuniar, N. M. (2023). Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes. *Jurnal Sains Dan Teknologi*, *5*(1), 243–246. https://doi.org/10.55338/saintek.v5i1.1383

Wati, A. (2020). Implementasi Artificial Neural Network Dalam Memprediksi Nilai Air Bersih Yang Disalurkan Di Provinsi Indonesia. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, *7*(3), 182–189. http://prosiding.seminar-id.com/index.php

Wibawa, A. P., Purnama, M. G. A., Akbar, M. F., & Dwiyanto, F. A. (2021). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134.