Journal of Physics: Conference Series

Water quality prediction using CNN

Water quality prediction using CNN

Vijay Anand M[1], Chennareddy Sohitha[1], Galla Neha Saraswathi[1] and Lavanya GV[1]

[1]Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, Tamil Nadu, India.

E-mail: vijayanand@saveetha.ac.in

Abstract. The interaction of solar radiation with the water level concentration and the elements of the water cause the water to have its characteristic hue. The alteration of the color of the water is reflective of the alteration of the water's properties and the degree to which it is suitable for use. Due to disasters like floods, tsunami in the last few years and water pollution has been an increasing problem. In world the intake of contaminated water causes 40% of deaths. Drinking unclean water is not safe and in order to reduce the issue to a level of extent, prediction of water quality can be done before consuming. The process used in water plants is based on the parameters pH, turbidity, temperature, hardness etc., of water using filtration and the water quality prediction can also be done using IOT by including both hardware and software. This project mainly comprises the primary level of water prediction using machine learning. Based on the color and quality of water the system predicts whether the given water sample is suitable for drinking or any further use. Tensorflow, Keras and CNN are used to train the model to forecast the water quality prediction. This project is cost-effective and works efficiently and can be used as immediate and initial level of water quality check since image processing tool is used. This model of water quality prediction can be checked using mobile captured and Google earth images of water samples.

1. Introduction

Water, the most important resource for maintaining life, is essential for the continued existence of all living things, including humans. In order to maintain its viability, every living thing must have access to high-quality water. Certain levels of pollution are within the tolerance range of aquatic organisms [1]. When these restrictions are exceeded, the very existence of their life is put in jeopardy. The majority of the bodies of water, including rivers, lakes, and streams, all adhere to predetermined quality criteria. The rapid expansion of industry results in a decline in water quality, which has a substantial impact on the quality of the water that is used for drinking. The United Nations (UN) estimates that over 1.4 million people pass away every year as a direct result of illnesses that are caused by polluted water. As a result, evaluating and making forecasts about the water quality (WQ) are of the utmost significance [2]. There have been a number of different ways suggested for predicting WQ, including statistical methods, visual modelling, analysing algorithms, and predictive algorithms. The importance of accurately forecasting water quality is expanding with the rapidly changing globe and greater demands [3]. This is because successful water resource management requires accurate prediction of water quality. The use of machine learning, a subfield of artificial intelligence, in the forecasting and improvement of water quality has recently become more widespread. The challenge of picking the most effective machine learning algorithm from among the many available options is an essential one. An empirical investigation on machine learning algorithms has been carried out so that the most effective algorithm may be chosen. The Convolutional Neural Network, often known as CNN, is a kind of supervised algorithm used in deep learning. In this type of method, the model is trained using an image as its input. Tensor flow is brought in by CNN [4].

Tensor flow is a library written in Python that can be used for creating deep learning models as well as for doing rapid numerical calculations [5]. The concepts of machine learning and deep learning are brought together in Tensor flow. Deep neural networks, which are useful for accomplishing tasks such as image processing, may be implemented with the help of Tensor Flow, which is a compilation of a large number of distinct algorithms and models. Picture processing is the procedure of performing different operations on an image in order to extract some valuable information from it. These operations might include things like cropping, resizing, and rotating the image. It is the process of analysing and manipulating a digital picture, most often with the goal of improving the image's quality. Importing the picture, analysing and manipulating the image, and producing output, where the outcome might be a changed image or a report that is based on the analysis of the image, are the three primary stages that are involved in the process of image processing. The system makes a determination, based on the colour, as to whether or not the provided water sample may be used for drinking water or any other reasons [6].

2. Related works

In [7], a performance assessment of application of three different machine learning techniques, including deep neural networks (DNN), gradient boosting machines (GBM), and extreme gradient boosting (XG Boost), was carried out in order to evaluate the ground water indices. In the predictions of EWQI and WQI utilising these three models, it was found that the parameter with the greatest significance was electrical conductivity (EC), while the parameter with the least significance was pH.

The model that is presented in reference number [8] takes into account the following nine characteristics about the quality of the water: temperature, pH value, electrical conductivity, oxygen saturation, biological oxygen demand, suspended particles, nitrogen oxides, orthophosphates, and ammonium. Using data spanning the years 2013–2019 from five different places in the Vojvodina Province of Serbia, it is constructed in the Netica programme and then tested and confirmed using this data. In light of this, we can confidently propose it as a reliable instrument throughout the transition from analogue to digital water management.

In [9], it was determined via research that the quality of the water had a direct influence on both the ecosystem and public health. Water is put to use for a variety of purposes, including consumption, agricultural production, and industrial processes. Several studies concerning rivers found in different parts of the globe have been carried out, and a sub-discipline of engineering known as river engineering has been suggested. Therefore, in this study, the water quality components of the Tireh River, which is one of the main rivers of the Dez catchment (which is one of the major catchments in Iran), were predicted using a support vector machine, an article neural network, and a group method of data handling. [Citation needed]

According to the findings of [10], the quality of the water has a direct impact not only on human health but also on the natural environment. According to the findings, the variables that have the greatest impact on the orderly classification of water quality are nitrate, pH, conductivity, dissolved oxygen, total coliform count, and biochemical oxygen demand (BOD), with respective Variable Importance values of 74.78, 36.805, 81.494, 105.770, and 105.166 and 130.173.

The estimation of water quality was identified as one of the most critical difficulties that the globe has been confronted with in the most recent decades by [11]. In this research, we develop a model for predicting water quality that makes use of the method of principal component regression. The findings reveal that the principle component regression approach has an accuracy of prediction of 95% and that the Gradient Boosting Classifier method has an accuracy of classification of 100%. This shows believable performance in comparison to the state-of-the-art models.

The article [12] examines a number of different prediction models that were constructed with the use of machine learning and big data approaches, as well as the experimental outcomes of water prediction and assessment that these models produced. This article discusses a wide range of problems and difficulties, and it offers some potential answers to research-related problems.

A unique Long and Short-Term Memory (LSTM) model-based data-driven model was developed for the purpose of forecasting the water quality in urban reservoirs as part of this research. To deconstruct the water quality data, the Complete Ensemble Empirical Modal Decomposition with Adaptive Noise (CEEMDAN) methodology is used. This method addresses the issue of the high demand for basic data in the process of developing water quality hydrodynamic physical models. This model not only accurately predicts water quality over a variety of predetermined time periods, but it also solves the problem of low prediction accuracy caused by the high volatility and randomness of the observed data [13]. This problem was caused by the high volatility and randomness of the observed data.

Recently, a method known as deep learning has received more attention in the field of water quality modelling. Deep learning is a subfield of machine learning that makes use of neural network topologies that consist of an input layer, many hidden layers, and a single output layer. For the [14], a model for determining the quality of drinking water was developed by using a network of long-term and short-term memory. They came to the opinion that the recommended LSTM network has promise as a tool for projecting drinking-water quality, including pH, dissolved oxygen (DO), chemical oxygen demand (COD), and NH3-N. This was the result that they arrived to. The LSTM network was also used to predict other metrics about the quality of the water, such as the temperature of the water. Using a hybrid convolutional neural network (CNN)-LSTM model, Barzegar et al. predicted the DO and chlorophyll-a (Chl-a) levels in the Small Prespa Lake, which is located in Greece [15]. They made the discovery that the hybrid CNN-LSTM model performed superiorly than the separate machine learning models of CNN, LSTM, support-vector regression (SVR), and decision trees.

3. Materials and Methods

This paper mainly comprises the primary level of water prediction using machine learning. Based on the color and quality of water the system predicts whether the given water sample is suitable for drinking or any further use.

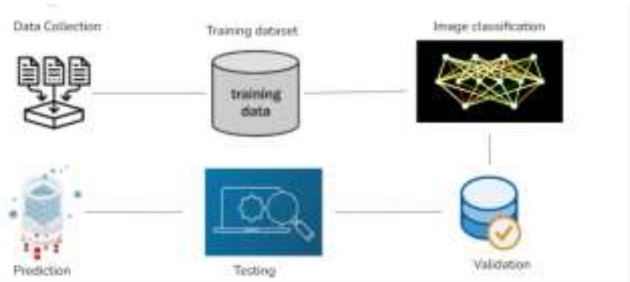

Figure 1. system design.

Tensorflow, Keras and CNN are used to train the model to forecast the water quality prediction. The following diagram represents the flow of the present study as shown in Figure 1.

3.1. Dataset

The dataset that was utilized for this research was taken from photographs found on Google Earth. The photographs were taken in a variety of river systems across the world; however, the vast majority of the shots is from a variety of states in India and was taken during the last few years. The dataset contains 200 samples and has color and appearance as its parameters.

3.2. Data Preprocessing
In the process of examining the data quality, processing the data plays a significant role. Before using the information in the research, the quality of the data is evaluated for accuracy. The data collection is split up into two different folders: one for photographs of clean water and another for images of polluted water. After finishing the initial steps in data preprocessing, since the size of each image is not same, all the images are rescaled into one size.

3.3. Classification of sample images
The process of prediction of water quality based on the Google earth images or captured images contain the following steps involved. They are:
- Dataset
- Training
- Validation
- Testing

The first part of the project is to collect the data i.e., water samples from different areas and from various rivers. Those images can be Google earth images and mobile captured one. The collected data is used to train the project model based on the water samples. Collection of data and training the dataset are done in data pre-processing part in the study. In order to get good accuracy, the dataset should be trained well. Validating the dataset proves the accuracy of the given water samples of rivers collected from different places which is categorized into two folders. Those two folders of validation dataset are compared with the training dataset which again contains two subfolders one is safe water images and the other one is unsafe water images.

3.4. Prediction of Water Quality
In this study we use CNN algorithm and keras library which is built on top of tensorflow. The final phase in the project or the study is testing. It tests the performance of the given dataset of water samples. It predicts whether the given image input water sample is safe or unsafe for further usage of the water.

3.5. Tensorflow
Tensorflow is a library in python which is used for fast numerical computing and to create deep learning models.

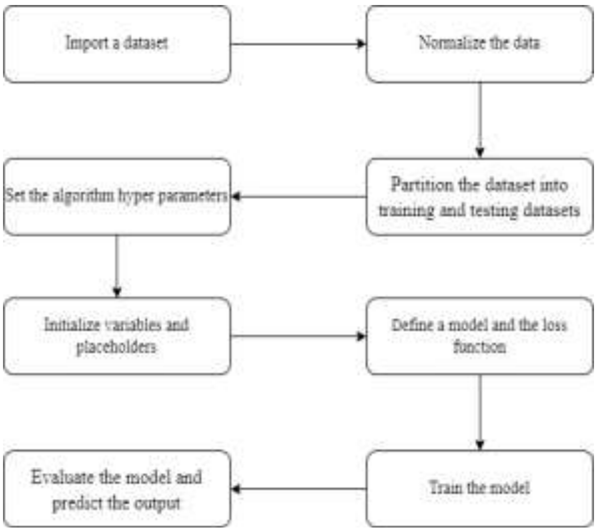

Figure 2. Tensorflow.

Journal of Physics: Conference Series **2484** (2023) 012051 doi:10.1088/1742-6596/2484/1/012051

Tensorflow is a combination of machine learning and deep learning. Figure 2 shows the tensorflow working flow.

3.6. Convolutional Neural Network (CNN)& Keras

In the field of deep learning, the Convolutional Neural Network is a kind of supervised algorithm that uses a picture as its primary form of data input. CNN has the capability of developing an internal representation of a two-dimensional picture. CNN focuses primarily on image processing, image recognition and image classification as its three primary areas of study.
Keras is a high-level neural network library that is built on top of Tensor Flow and is designed to operate on that platform. It also has applications in image processing and the training of models. Because it is a built-in library in Python, Keras is very user-friendly. The use of Keras is another component of deep learning. In this research, CNN and Keras are primarily responsible for the production of images, the processing of images, and the categorization of images.

4. Results and Discussion

The prediction testing is conducted in a specific environment (Google Colab). The software requirements that are used for development and execution are a system with i3 Processor and 4GB RAM. The below Table 1 represents the accuracy, water quality and no. of sample values of the present study and Figure 3, is plotted based on the values accuracy and no. of samples for the water quality prediction which is took while validating the system.

Table1. Performance analysis.

S.No	Accuracy	Water Quality	No. of Samples
1	>80	Clean	40
2	70-80	Good	50
3	60-70	Poor	35
4	50-60	Very Poor	40
5	<50	Unclean	30

The processing of data that has a grid-like structure, such as an image, is the area of expertise of a type of neural networks called Convolutional Neural Networks, which are sometimes abbreviated as CNNs or ConvNets for short. A binary representation of visual data is what we refer to as a digital picture. It has a number of pixels that are laid out in a grid-like pattern and has pixel values that indicate how bright and what colour each pixel should be. It also has a sequence of pixels.

No. of samples

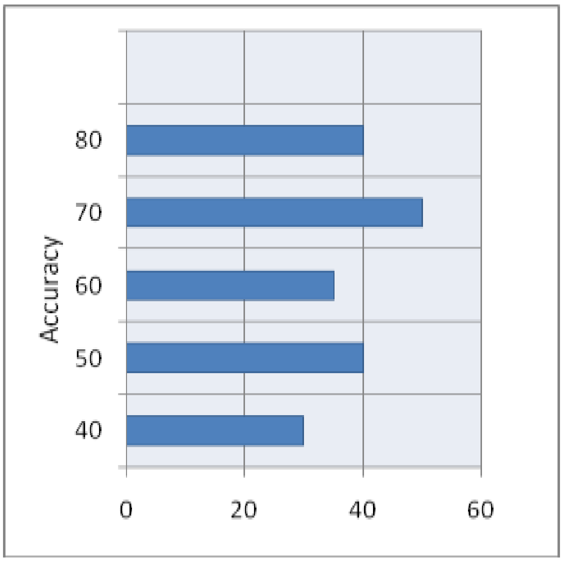

Figure 3. Performance analysis.

The parameters used in the other experimental algorithms are pH, DO, BOD, Nitrate, Conductivity etc., and water quality index is calculated using different algorithms to predict the quality of water. The above Table 2 and Figure 4 show the comparison between accuracy of different algorithms. Based on the references and comparisons the above graph states that accuracy level of CNN algorithm used in the present study has higher accuracy which is used to predict the water quality based on the images.

Table2. Comparison of algorithms.

Algorithm	**Accuracy**
KNN	65
CNN	80
SVM	75
Naive bayes	70 on the color and quality of water the system predicts whether the given water sample is suitable for

This project mainly comprises the primary level of water prediction using machine learning based drinking or any further use. Tensorflow, Keras and CNN are used to train the model to forecast the water quality prediction. This project is cost-effective and works efficiently and can be used as immediate and initial level of water quality check since image processing tool is used. This model of water quality prediction can be checked using mobile captured and Google earth images of water samples.

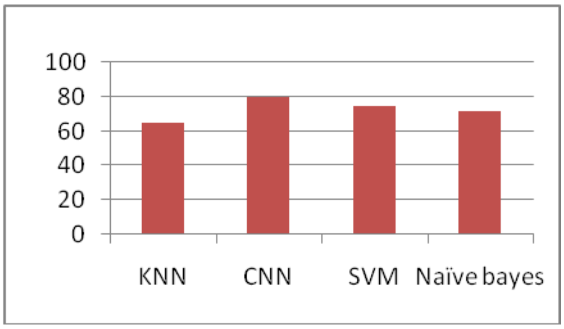

Figure 4. Various algorithms.

The dataset is divided into three subsets, one is for training another one is for validating and the other one is for testing. While tensorflow, keras and CNN is used for classification, image processing, data generation and prediction of water sample images.

5. Conclusion

The ability to accurately forecast the quality of the water is critical for the preservation of the ecosystem. This model is produced by using machine learning methods and algorithms in order to forecast the water quality based on the sample photos of water. This model is beneficial for the first testing phase in order to determine whether or not the water is appropriate for future usage. It is expected that the gathered water samples are accurate to a level of 85%. At this time, the study is only capable of making predictions about the quality of the water based on the information supplied by the user. It is possible to improve it further so that it can anticipate the water quality taking into account a variety of aspects and characteristics such as pH, turbidity, saturation, and so on. This model may be implemented as either a website or a mobile application, and its functionality will be determined by the GPS position. The primary emphasis of this research is placed on the foundational level of machine learning-based water prediction. The colour of the water as well as its overall quality are taken into consideration when the system determines whether or not a particular sample of water may be used further or consumed. Tensor flow, Keras, and CNN are used in the training of the model so that it can accurately predict water quality. This project's efficiency and cost-effectiveness are greatly enhanced by the use of an image processing application. In addition to this, it may be used as a simple and quick method for determining the quality of the water. This method for predicting the quality of water may be validated with the use of Google Earth and photographs of water samples obtained using mobile devices.

References

[1] Hemant Raheja, Arun Goel, Mahesh Pal,2022.Prediction of ground water indices using machine learning algorithms.*Water Practice and Technology*;**17(1):**336–351.

[2] Ilić M, Srdjević Z, Srdjević B, 2022.Water quality prediction based on Naïve Bayes algorithm. *Water Sci Technol* **8 (4):**1027–1039.

[3] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi;Abbas Parsaie**.**Water quality prediction using machine learning methods. *Water Quality Research Journal*;**53(1)**:3–13.

[4] Mehedi Hassan MD, Laboni Akter, Mushfiqur Rahman MD, Sadika Zaman, Khan MD Hasib Nusrat Jahan, Raisun Nasa Smrity, Jerin Farhana, Raihan, Swarnali Mollick M,2021. Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms.

[5] Saikat Islam Khan MD, Nazrul Islam, Jia Uddin, Sifatul Islam, Mostofa Kamal Nasir,2021. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach.

[6] Nair JP and Vijaya MS,2021. Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey. *2021 International Conference on Artificial*

Intelligence and Smart Systems (ICAIS):1747-1753,doi: 10.1109/ICAIS50930.2021.9395832.
[7] Yang Y, Xiong Q, WU C, Zou, Q, Yu Y, Yi, H, and Gao M, 2021. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. *Environmental Science and Pollution Research*;**28(39):**55129-55139.
[8] Chen Y, Song L, Liu Y,Yang L and Li D,2020.A review of the artificial neural network models for water quality prediction. *Applied Sciences*, ***10*(17):**5776.
[9] Barzegar R, Aalami MT and Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*;***34*(2):**415-433.
[10] Baek SS, Pyo J and Chun JA,2020. Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. *Water*, ***12*(12):**3399.
[11] Yan J, Liu J, Yu Y, and Xu H, 2021.Water quality prediction in the luan river based on 1-DRCNN and bigru hybrid neural network model. *Water*;**13(9):**1273.
[12] Jichang TU, Xueqin YANG, Chaobo CHEN, Song GAO, Jingcheng WANG and Chengm SUN, 2019.Water quality prediction model based on GRU hybrid network. *In 2019 Chinese Automation Congress (CAC):*1893-1898.
[13] Khan MSI, Islam N, Uddin J, Islam S, and Nasir MK,2022.Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. *Journal of King Saud University-Computer and Information Sciences*, ***34*(8):**4773-4781.
[14] Hussain deen A, Iqbal S and Ambegoda TD,2016. Multi-Label Prototype Based Interpretable Machine Learning For Melanoma Detection. *International Journal of Advances in Signal And Image Sciences*;***8*(1)**:40-53.
[15] Zhang L, Jiang Z, He S, Duan J, Wang P, and Zhou T, 2022.Study on water quality prediction of urban reservoir by coupled CEEMDAN decomposition and LSTM neural network model. *Water Resources Management*;***36*(10)**:3715-3735.