Water Quality Classification and Spread Velocity Estimation using Deep Learning

Avishek Bhowmick[1], Dr. S.K. Jagtap[2], K.A. Pujari[3], Atharva Pol[4], Om Pimpalgaonkar[5]

Department of E&TC Engineering, SKNCOE, SPPU, Pune
[1]avmik77@gmail.com
[2]skjagtap.skncoe@sinhgad.edu
[3]kapujari.skncoe@sinhgad.edu
[4]atharvakpol@gmail.com
[5]pimpalgaonkarom21@gmail.com

Abstract—
Water pollution has always posed in a great threat to human health and aquatic life in the water bodies. Such pollution is certainly a major source of diseases like typhoid, diarrhea, jaundice and sometimes can also be fatal especially in the developing nations. So it is necessary to determine the probable contaminants that could be present in water. Therefore, this work focuses on the classification of water contamination based on color which would predict the probable contaminants so that it would be helpful to know what type of harmful contaminants could be present. For the classification of the contaminated water, a total of 850 images with only 6 colours viz. blue, black, brown, red, green and white were considered since the system had to classify into 6 different colour classes. This work mainly focused on the use of machine learning techniques so as to provide an efficient way to carry out the above stated problem. After succinct training with a dataset representing various colours of contaminated water , the deep learning algorithm (CNN) achieved an accuracy around 88%. Along with the water contamination, it was also equally important to address the impurities in water. So a unique method to calculate the spread velocity of liquid impurity in water using basic image processing technique was developed.

Keywords*— Water Classification, Contamination, Spread Velocity, Convolution Neural Networks, Image Processing.*

I. INTRODUCTION

Water is one of the basic needs for life. However, pollution and environmental contaminants have a negative impact on the water quality thereby making it unsafe to use. Today, the world is facing severe issue of water pollution due to various reasons such as discharge of domestic and industrial effluent wastes, leakage from water tanks, marine dumping, radioactive waste and atmospheric deposition. These events are a threat to the aquatic life as well as humans and therefore a solution is needed to determine the water quality by using machine learning techniques. So, to provide an effective solution to the above stated problem, the system proposed a method to classify the water contamination using neural networks. In machine learning, neural networks, use artificial intelligence to unravel and simplify extremely complex relationships. They form the basis of most methods of deep learning, a subset of machine learning that holds multiple sequential layers through which data is run through in order to perform classification and analysis. Convolutional neural networks (CNNs) represent another subset of neural networks and deep learning. CNNs perform super-vised learning, which takes input as an image, assigns importance to various objects in the image and be able to differentiate one from the other. CNNs are complex organizations of nodes known as neurons that form connections as they are trained on data. The role of CNNs is basically to reduce the images into a form that is easy to process, without any loss of features that are critical for obtaining a good prediction. Thus, the system develops the structure of the model, including various layers that perform different functions and also contribute to the model's performance in different ways. This work is mainly focused on classification of the contaminated images of water bodies based on colour and then detecting the probable contaminants which could be present in water. A total of 850 images with only

6 colours viz. blue, black, brown, red, green and white were considered as the image dataset. It could be relevant in field applications of geotechnical engineering such as surveys, monitoring and studies in maintaining a good ecosystem for the marine life and would also be applicable in dams where the water collected is mainly rainwater which could contain chemicals due to pollution in air. This work also estimated the spread velocity of the liquid impurities (ink) by using image processing. The system used a unique prospect by applying image subtraction and the resulting image was the average spread velocity of the liquid. Spread was calculated for various liquid impurity samples to determine the nature of the spread velocity with respect to time.

II. Literature Survey

K. Horak et al work dealt with a water quality assessment using image processing methods. The method for measurement of the water quality used two well-known biological organisms viz. Daphnia magna and Lemna minor. In this design, these two organisms were continuously scanned in separated vessels by two cameras and acquired images were then processed autonomously. Methods of a colour-space transformation and a motion analysis in an image processing stage were employed. At the end, an indicator of a relative water quality was computed on a basis of extracted features from the acquired images. In this study, the mentioned biological organism's response to the toxicity/contamination of the water. The response as in the change in colour of the Lemna minor and the change in the movement of the Daphnia magna was measured using image processing methods and conclusive results were generated. This study was designed as an autonomous camera-based inspection of the water quality [1]. U. Ahmed et al has used supervised machine learning algorithm to estimate Water Quality Index (WQI) for finding general quality of water and the water quality class (WQC), which was determined on the basis of the WQI. The method used by the author gave four input parameters viz. temperature, turbidity, pH and total dissolved solids. Gradient boosting machine learning algorithm, with learning rate of 0.1 and polynomial regression with a degree of 2 was used by the authors [2]. H. Mohammed et al developed adaptive neuro fuzzy inference system (ANFIS) models for detecting the safety condition of water in pipe networks when concentrations of water quality variables in the pipes exceeded their maximum thresholds. The detection done by the author was based on time series data composed of pH, turbidity, colour and bacteria count measured at the effluent of a drinking water utility and nine different locations of sensors in the distribution network in the city of Ålesund, Norway [3]. A. Sweidan et al proposed an automatic classification approach for assessing water quality based on fish liver histopathology. As fish liver is a good bioindicator for detecting water chemical pollution, the proposed approach utilized fish liver microscopic images in order to detect water pollution. The proposed approach consisted of three phases; namely pre-processing, feature extraction, and classification phases. Also, it was implemented using Principal Components Analysis (PCA) along with Support Vector Machines (SVMs) algorithms for feature extraction and water quality degree classification [4]. M. Ali Ghorbani et al transformed images using image analytics (needed for conversion of unstructured big data into readable machine data) for extracting information to form a modelling dataset and constructed predictive models by learning inherent correlation between observed SSC values and their image analytics. The capability for monitoring SSC was based on photometric features of Red, Green, and Blue (RGB), where modern cheap high resolution cameras were capable of capturing subtle changes in tone and color, both expressed in bits. RGB imaging tracked down the changes in the color of river water through simple high-resolution images. The correlation between SSC and color variations of RGB-based high-resolution images was explored for the prediction of SSC. In the research done by authors, image analytics comprise 8 input variables and comprised mean, mean intensity, entropy and standard deviation as well as target values in terms of measured SSC. For any correlation in image analytics, the author used following techniques : GLM and DRF [5]. V. Kilic et al presented a mobile platform for colorimetric water quality detection based on the use of a built-in camera for capturing a single-use reference image. They developed an app for processing the image for training and creating reference models. They also cited to have achieved approximately 100% accuracy. The project finds its application at the time when the Department of Water Resources each month estimates the water contamination of various water bodies such as rivers, lakes, ponds, streams, dams, etc. and the pollutants which causes this contamination. These pollutants may be natural or man made. Collecting the samples, analysing the samples takes huge amount of time. The project aims to target this by predicting the probable contaminants by using colour classification. The only work left is to physically test in the lab to confirm the contaminants predicted by the system is correct or not [6]. S. Nivesh et al have made a comparison between multiple linear regression and artificial neural networks (ANNs). Based on study, the performance results which are based on root mean square error (RMSE), correlation coefficient (r) and coefficient of efficiency (CE), can predict sediment load more efficiently than traditional models like multiple linear regression [7].

III. Methodology

A. Water Quality Classification

The overall methodology for the water quality classification involves six steps which are represented in the form of block diagram in fig.1.

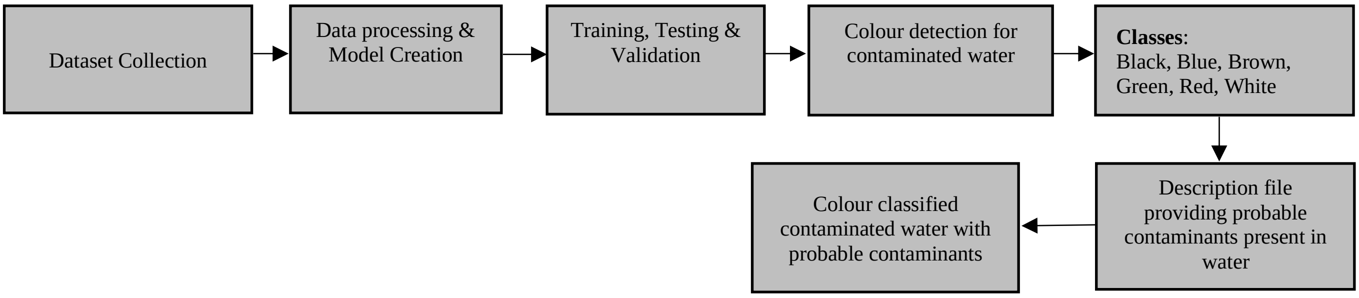

Dataset Collection
Data processing & Model Creation
Training, Testing & Validation
Colour detection for contaminated water
Classes: Black, Blue, Brown, Green, Red, White
Description file providing probable contaminants present in water
Colour classified contaminated water with probable contaminants

Fig.1 Block diagram representing water quality classification

Dataset Collection: For the classification of water contamination, images representing 6 classes of colour were prepared.

Data processing and model creation: Data (images) are stored into respective colour classes. Each image is rescaled and stored into separate directories. A sequential model has been implemented from the machine learning library.

Training, Testing and Validation: In this block, the pre-processed data was passed to the model which trained the model and simultaneously tested and validated it. This provides respective accuracies viz. training, testing and validation.

Colour Detection: The model trained and defined to detect the colour of the water, which performs prediction on the test dataset. It provides the colour of water and from the description file provided, it indicates the possibility of the probable contaminants present.

B. Spread Velocity Estimation

The process for determining the average spread velocity of liquid impurity is represented in fig.2 with the help of block diagram.

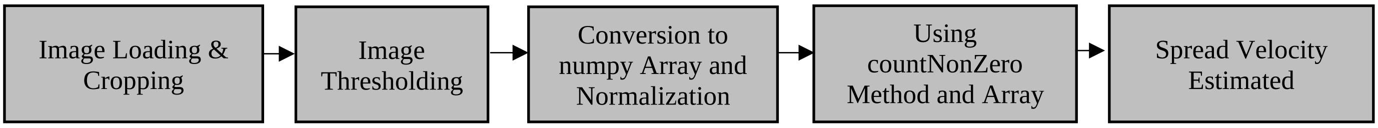


Image Loading & Cropping
Image Thresholding
Conversion to numpy Array and Normalization
Using countNonZero Method and Array
Spread Velocity Estimated

Fig.2 Block diagram representing spread velocity estimation

This work further involves the determination of velocity of the impurities present in the water. Basic principle of image thresholding and mathematical operations on array are used to determine the spread velocity of any liquid impurity having color not similar to the solvent liquid. Images with time interval of 't' seconds are loaded initially, which are further cropped to same dimension i.e. width and height. After cropping, both images are converted into grayscale and then by applying basic thresholding they are converted into binary images. These binary images are further converted into numpy array and normalized. By using the 'countNonZero' method from OpenCV library the number of non-zero pixels are counted i.e. number of white pixels. These pixels are subtracted from the number of total pixels in the image i.e. image size. This gives the total number of black pixels in each image, which are further subtracted with each other to achieve the spread in terms of pixels. Finally by using the general mathematical formula of velocity i.e. final state – initial state / time interval the system attains the spread in terms of pixels per unit time.

IV. Experimentation

This work dealt with the classification of the contaminated water images. The images were classified into 6 colour classes viz. Blue, Black, Brown, Red, Green, White. The system was trained which predicted the colour of contamination and based on the colour. A description file was provided so as to indicate the probable contaminants present in water. Fig.3 indicates datasets comprising 850 images representing 6 colour classes.

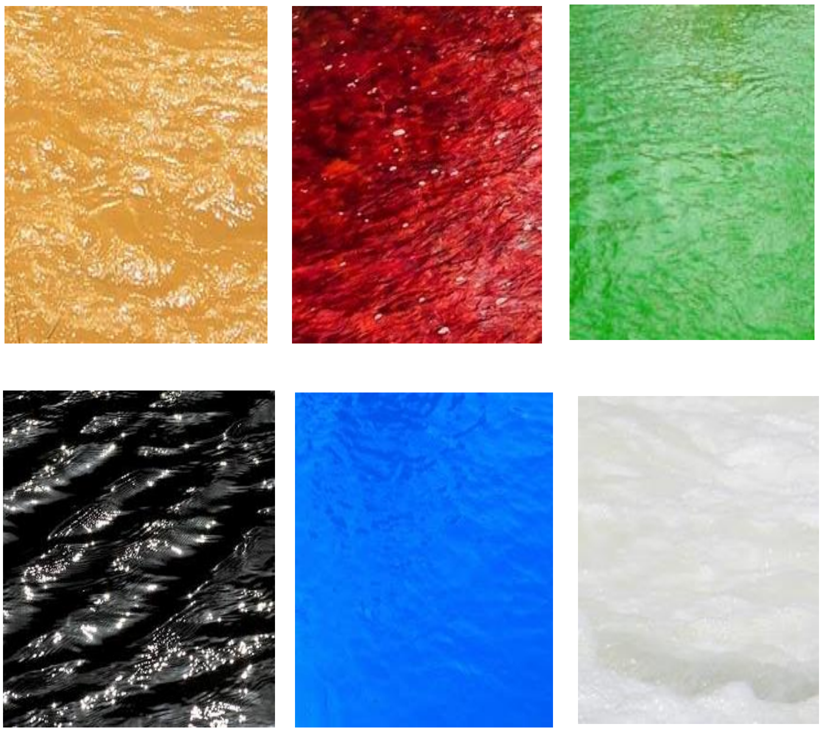
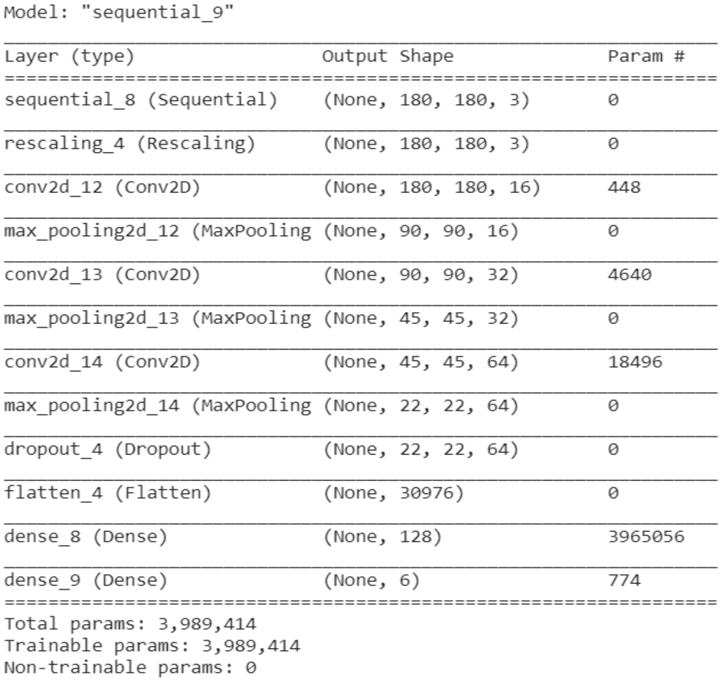


Fig.3 Images representing 6 classes of contamination

The implementation of machine learning algorithm involves importing libraries, raw data, defining the dataset, training and testing the dataset generation, label generation, auto tuning and data augmentation. Table.1 indicates the model which is implemented using different layers such as conv_2D to extract the features of the image, Maxpooling_2D which calculates the maximum or the largest value in each patch of each feature map and helps to reduce overfitting, reducing the number of parameters to learn, Dropout layer is used so that the neural network doesn't overfit i.e. it gives extra zeroes to some nodes which actually tells the model to not actually learn the way the data is fed. Basically it dropouts certain neurons by learning the process , flattening and dense layer. The next step involves the compilation of sequential model so that the system can model the neural network and can have an object with the sequential which will be called as the model. After training the model by considering epochs with 20, validation accuracy was found to be 88.27% and thus was able to predict the colour of water with possible contaminants.



```
Model: "sequential_9"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
sequential_8 (Sequential)    (None, 180, 180, 3)       0         
_________________________________________________________________
rescaling_4 (Rescaling)      (None, 180, 180, 3)       0         
_________________________________________________________________
conv2d_12 (Conv2D)           (None, 180, 180, 16)      448       
_________________________________________________________________
max_pooling2d_12 (MaxPooling (None, 90, 90, 16)        0         
_________________________________________________________________
conv2d_13 (Conv2D)           (None, 90, 90, 32)        4640      
_________________________________________________________________
max_pooling2d_13 (MaxPooling (None, 45, 45, 32)        0         
_________________________________________________________________
conv2d_14 (Conv2D)           (None, 45, 45, 64)        18496     
_________________________________________________________________
max_pooling2d_14 (MaxPooling (None, 22, 22, 64)        0         
_________________________________________________________________
dropout_4 (Dropout)          (None, 22, 22, 64)        0         
_________________________________________________________________
flatten_4 (Flatten)          (None, 30976)             0         
_________________________________________________________________
dense_8 (Dense)              (None, 128)               3965056   
_________________________________________________________________
dense_9 (Dense)              (None, 6)                 774       
=================================================================
Total params: 3,989,414
Trainable params: 3,989,414
Non-trainable params: 0
```

Table.1 Model Summary

The entire classification and determination of water quality was achieved using Convolutional Neural Networks (CNN). Some images were produced manually whereas some were readily available as dataset. These images were then collectively trained and validated for further process using machine learning approach. Fig.4 indicates real world dataset of 89 images which was prepared to analyse whether the accuracy achieved was satisfactory or not. The images were collected at different locations across Pune city at various time. The locations include Sinhagad road canal, Vadgaon canal, Khadakwasla Dam backwaters and lake, Mulla Mutha river. When deployed, the system was able to classify 66 images out of the total 89 images, based on the 6 colours passed as the parameters along with the probable contaminants. The model attained an accuracy of around 74.15% and loss of 0.2584.

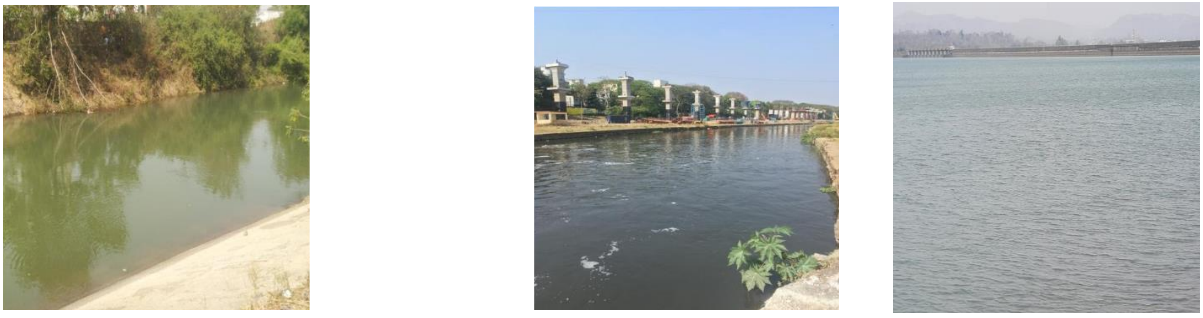
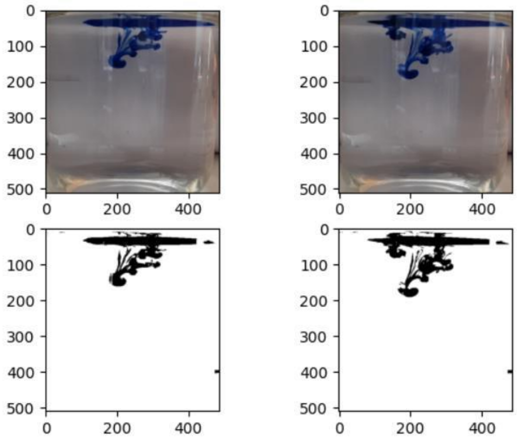


Fig.4 Real World Images

The models were trained and tested on a PC running on Windows 10 operating system with dedicated 4GB NVIDIA GeForce 940MX GPU. Tensorflow and Keras library were used in order to accelerate the training of the model. Google Colab was used for water classification process and for finding velocity of liquid impurity.

In the next objective, the average spread velocity of liquid impurity (ink) was determined by using a small image dataset of ink diffusion in the water using image processing techniques. Fig.5(a) and Fig.5(b) represents the RGB images which were converted into gray scale images for calculating the spread velocity of ink along with its diffusion. These images were originally captured from video representing the complete spread and diffusion process of the ink in water at regular intervals of time. First, two images at a specific time interval were taken as an input. They were resized so as to bring them into same dimension. Further, the operation of image subtraction was performed which provided the net area spread. Determination of velocity of impurities was carried along with the images acquired after capturing them from a video representing the spread and diffusion of ink.



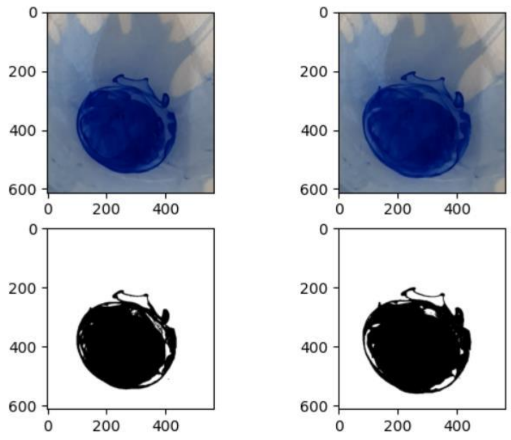


Fig. 5(a) Fig. 5(b)
(Side view indicating the spread)
(Top view indicating the spread)

Ink was considered as a liquid impurity in this work because it was quite easier to observe the diffusion and spread process from any point of view. It was much easier to collect some images from the video at regular intervals of time.

V. RESULTS & DISCUSSION

The model was able to predict the probable contaminants present in the water. The system achieved an accuracy of 88.27% and loss of 0.3644. When tested on real world images, accuracy of 74.15% attained was quite satisfactory with loss of 0.2584. The calculation of the spread velocity of ink as mentioned in this work, was successful and hence it can be concluded based on the achieved results that velocity is not constant. It keeps on changing at different time intervals. In both the cases i.e. for vertical as well as surface spread, the velocity was found to be different as per their respective time interval.

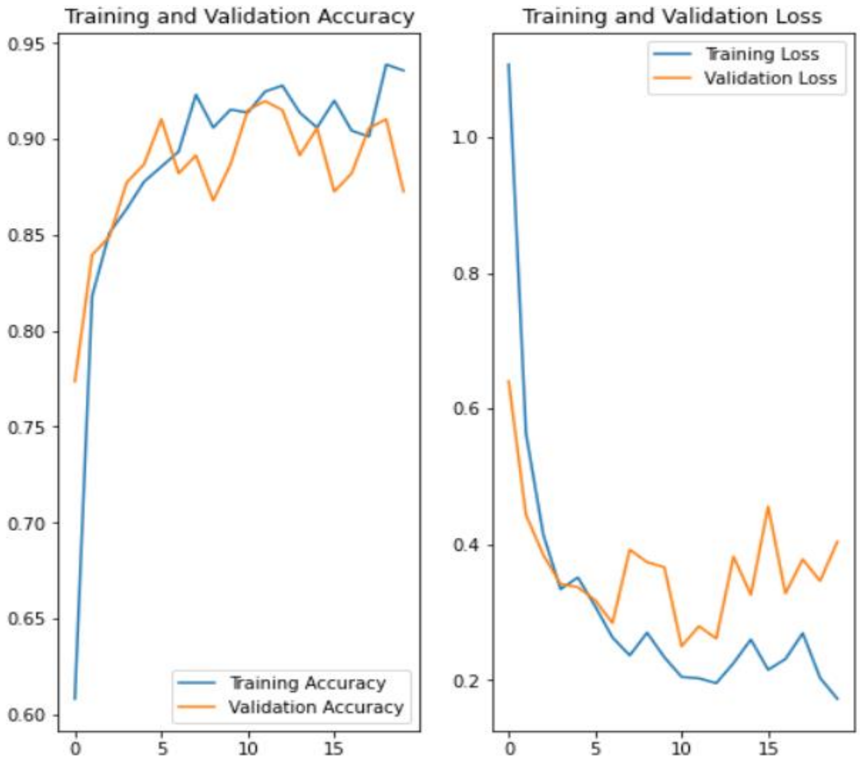
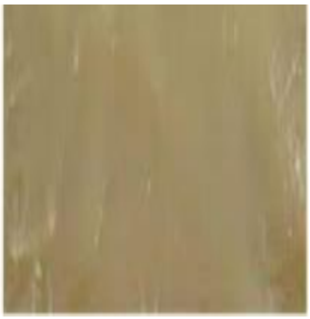
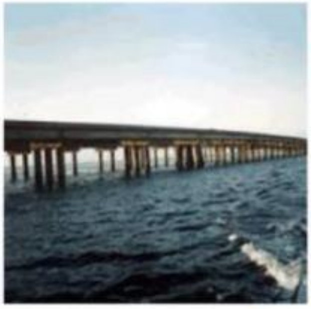


Fig.6 Graphs representing Training Accuracy vs Validation Accuracy and Training Loss vs Validation Loss for classification of contaminated water

Fig.6 represents the variation in the training accuracy vs validation accuracy and training loss vs validation loss involved during the entire model execution as per the epochs.





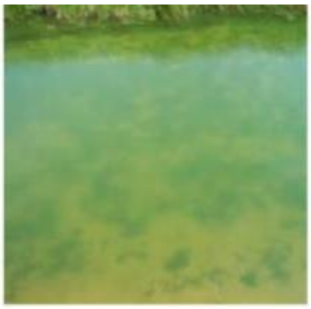
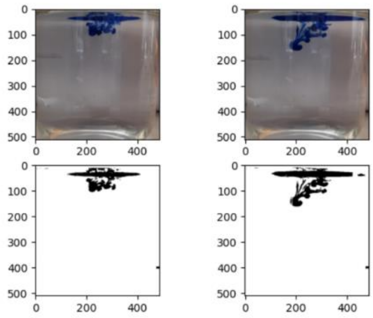


Fig.7 Brown water Fig.8 Black water Fig.9 Green water

The above fig.7 is most likely representing a brown water image with a 99.08 percent confidence. It might contain traces of mud, rust, tannins, suspended sediments. Similarly, fig.8 is representing a black water image with a 96.53 percent confidence. It might contain traces of water softeners, sulphides, manganese, cement, clay. Also, fig.9 represents a green water image with a 100.00 percent confidence with probable traces of algae, phytoplankton, chlorine. It is known that water contamination has become a serious issue in today's world. Various methods have been proposed earlier to solve this problem. In this work, machine learning approach as well as image processing techniques have been applied to address this issue. Going through this work many challenges regarding the dataset collection and accuracy were encountered. To overcome this challenge, dataset was prepared in the lab for water quality classification. The system has also tracked the spread velocity of liquid impurity by taking ink as an assay and calculated the same at different time intervals by using image processing. Fig.10(a) and fig.10(b) indicate the vertical spread of the ink and similarly fig. 11(a) and fig.11(b) indicate the surface spread of the ink.



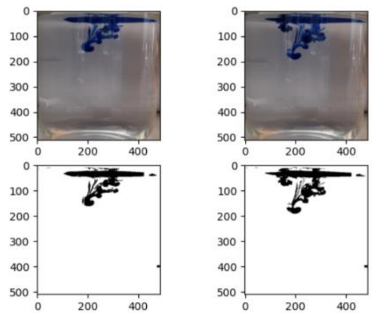
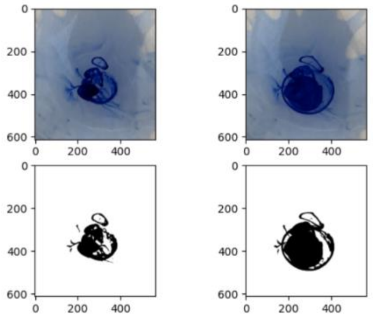


Fig.10(a) Fig.10(b)
Vertical Spread from 0 to 2 seconds
Vertical Spread from 2 to 4 seconds



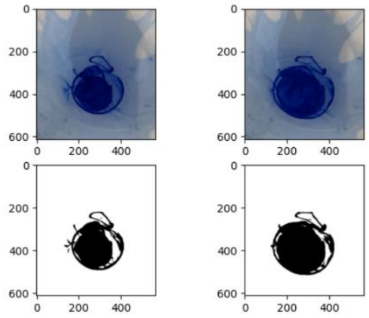


Fig.11(a)
Fig.11(b)
Vertical Spread from 0 to 2 seconds
Vertical Spread from 2 to 4 seconds

In fig.10(a), the average spread velocity from 0 to 2 seconds came out to be 2345 pixels/sec and in fig.10(b), the average spread velocity from 2 to 4 seconds was 1442.5 pixels/sec. Similarly for fig.11(a) and fig.11(b) which represented the surface spread of ink from 0 to 2 seconds and 2 to 4 seconds is 9226.5 pixels/sec and 7936.5 pixels/sec respectively. From this, it was observed that there is a non-linear relationship in the spread velocity at definite time intervals i.e. the velocity was not constant.

VI. CONCLUSION

The model was able to predict the probable contaminants present in the water. The system achieved an accuracy of 88.27% and loss of 0.3644. On the basis of real world images, accuracy of 74.15% attained was quite satisfactory with loss of 0.2584.

The calculation of the spread velocity of ink as mentioned in this work, was successful and hence it can be concluded based on the achieved results that velocity is not constant. It keeps on changing at different time intervals. In both the cases i.e. for vertical as well as surface spread, the velocity was found to be different as per their respective time interval.

ACKNOWLEDGEMENT

This project consumed huge amount of work, research and dedication. Still, implementation would not have been possible without the support of many individuals and organization. Therefore, we would like to extend our sincere gratitude to all of them.

Authors are thankful to Dr. Selva Balan Sc.E, Central Water and Power Research Station (CWPRS), Khadakwasla, Pune for providing the concept of this project and necessary ideas concerning project's implementation.

REFERENCES

[1] K. Horak, J. Klecka, and M. Richter 'Water Quality Assessment by Image Processing', Oct 2015, [Online], Available: https://ieeexplore.ieee.org/document/7296329

[2] U. Ahmed, R. Mumtaz, H. Anwar , A. Shah , R. Irfan and J. García-Nieto 'Efficient Water Quality Prediction Using Supervised Machine Learning', MDPI

[3] H. Mohammed, I. Hameed, R. Seidu 'Machine Learning – Based Detection of Water Contamination in Water Distribution Systems', Researchgate

[4] A. Sweidan, N. Bendary, A. Hassanien, O. Hegazy, A. Mohamed, 'Machine Learning based Approach for Water Pollution Detection via Fish Liver Microscopic Images Analysis', IEEE

[5] M. Ali Ghorbani, R. Khatibi, V. P. Singh, E. Kahya, H. Ruskeepaa, M. Kaur Saggi, B. Sivakumar , S. Kim, F. Salmasi, M.H. Kashani, S. Samadianfard, M. Shahabi, R. Jani 'Continuous monitoring of suspended sediment concentrations using image analytics and deriving inherent correlations by machine learning', Scientific Reports

[6] V. Kilic , G. Alankus, N. Hozrum, Y. Mutlu 'Single-Image-Referenced Colorimetric Water Quality Detection Using a Smartphone', Researchgate

[7] S. Nivesh , P. Kumar 'Modelling river suspended sediment load using artificial neural network and multiple linear regression: Vamsadhara River Basin, India

[8] A. Gupta , E. Ruebush 'AquaSight: Automatic Water Impurity Detection Utilizing Convolutional Neural Networks', *Department of Computer Science TJHSST Alexandria, USA, Department of Computer Science University of Maryland Bethesda, USA,* July 2019, arXiv:1907.07573v1, [Online], Available: https://arxiv.org/abs/1907.07573

[9] X. Wang, F. Zhang , J. Ding, 'Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China', Scientific Reports

[10] S. Ouellet-Proulx , A. St-Hilaire , S.C. Courtenay , K. Haralampides 'Estimation of suspended sediment concentration in the Saint John River using rating curves and a machine learning approach'. Article in Hydrological Sciences Journal/Journal des Sciences Hydrologiques , May 2015

[11] I. Mohamed and I. Shah 'Suspended Sediment Concentration Modeling Using Conventional and Machine Learning Approaches in the Thames River, London Ontario

[12] H. Vu, University of New Orleans ‘A Machine Learning Assessment to Predict the Sediment Transport Rate Under Oscillating Sheet Flow Conditions

[13] F. Muharemi, D. Logofătu & F. Leon ‘Machine learning approaches for anomaly detection of water quality on a real-world data set’, Journal of Information and Telecommunication

[14] T. Acharya , A. Subedi , H. He and D. Ha Lee ‘Classification of Surface Water using Machine Learning Methods from Landsat Data in Nepal’, MDPI

[15] M. Tabatabaein, A. S. Jam, S.A. Hosseini ‘Suspended sediment load prediction using non-dominated sorting genetic algorithm II’, International Soil and Water Conservation Research

[16] K. Peterson , V. Sagan , P. Sidike, A.L. Cox, M. Martinez, ‘Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine’, MDPI