A PROJECT REPORT

ON

WATER QUALITY ASSESSMENT & VELOCITY MEASUREMENT USING IMAGE PROCESSING AND ML TECHNIQUES

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING
IN
ELECTRONICS AND TELECOMMUNICATION

BY

Om Pimpalgaonkar	**Roll.No. BB037**
Avishek Bhowmick	**Roll.No. BB039**
Atharva Pol	**Roll.No. BB038**

UNDER THE GUIDANCE OF

Dr. S. K. JAGTAP

Sinhgad Institutes

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING
STES'S
SMT. KASHIBAI NAVALE COLLEGE OF ENGINEERING
VADGAON BK, OFF SINHGAD ROAD,
PUNE 411041

2020-21

CERTIFICATE

This is to certify that the project phase-I report entitles

"WATER QUALITY ASSESSMENT & VELOCITY MEASUREMENT USING IMAGE PROCESSING AND ML TECHNIQUES"

Submitted by

Om Pimpalgaonkar	**Roll No : BB037**
Avishek Bhowmick	**Roll No : BB039**
Atharva Pol	**Roll No : BB038**

is a bonafide work carried out by them under the supervision of **Dr. S. K. Jagtap** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of **Bachelor of Engineering** (Electronics and Telecommunication Engineering).

This project phase-I report has not been earlier submitted to any other institute or University for the award of any degree.

Prof. K.A. Pujari / Prof. A.S. Pandit
Co-Guide
Department of E&TC

Dr. S. K. Jagtap
Guide
Department of E&TC

Prof. P.G. Chilveri / Prof S.S. Jahagirdar
Project Coordinator
Department of E&TC

Dr. S.K. Jagtap
Head
Department of E&TC

Dr. A.V. Deshpande
Principal
S.K.N.C.O.E, Pune-41

Place: Pune
Date: 7th June,2021

ACKNOWLEDGEMENT

This project consumed huge amount of work, research and dedication. Still, implementation would not have been possible if we did not have the support of many individuals and organization. Therefore, we would like to extend our sincere gratitude to all of them.

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

We are grateful to **Dr. S.K. Jagtap, Prof. K.A. Pujari** for provision of expertise and exemplary guidance, support , monitoring and constant encouragement throughout this project.

Om Pimpalgaonkar
Avishek Bhowmick
Atharva Pol

ABSTRACT

Estimation of water contamination is of utmost importance so as to maintain a good water quality. It is equally important to have an efficient, low cost and accurate method for assessment of water quality. Thus, in order to inscribe the ubiquitous issue of augmenting water pollution, the system proposed a methodology using deep learning method, Convolution Neural Network (CNN), for water impurity detection. After compendious training with a dataset representing various levels of contamination, the deep learning algorithm achieved an accuracy around 90.38%. Furthermore, the system has estimated the average spread velocity for liquid impurities. When deployed, the machine learning algorithm would provide an efficient way for determining water contamination level, thereby cautioning the local authority and goad them into action**.** Further, for the classification of the contaminated water, a total of 850 images with only 6 colors viz. blue, black, brown, red, green and white were considered since the system had to classify into 6 different color classes. The entire classification and determination of water quality was achieved using Convolutional Neural Networks (CNN). Some images were developed manually whereas some were readily available as dataset. These images were then collectively trained and validated for further process using machine learning approach. 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.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

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 determine 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.

1.2 RELEVANCE

This project could be relevant in field applications of geotechnical engineering such as surveys, monitoring and studies in maintaining a good ecosystem for the marine life. It is also applicable in dams where the water collected is mainly rainwater which could contain chemicals due to pollution in air.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

[1] 'Water Quality Assessment by Image Processing', Karel Horak, Jan Klecka, and Miloslav Richter, Oct 2015.

Abstract: This study deals with a water quality assessment using image processing methods. The method for measurement of the water quality uses two well-known biological organisms viz. Daphnia magna and Lemna minor. In this design, these two organisms are continuously scanned in separated vessels by two cameras and acquired images are then processed autonomously. There are employed methods of a colour-space transformation and analysis and a motion analysis in an image processing stage. Finally, an indicator of a relative water quality is computed on a basis of extracted features from the acquired images.

Introduction: In this study the mentioned biological organism's response to the toxicity/contamination of the water. The response as in the change in color of the Lemna minor and the change in the movement of the Daphnia magna is measured using image processing methods and conclusive results are generated. The two biological indicators of water quality: Lemna minor and Daphnia magna. This study is designed as an autonomous camera-based inspection of the water quality.

Methodology: Using a mechanically designed hardware consisting of Industrial grade CCD cameras, few computing machines and glass tanks, this study is implemented:
Lemna minor: The images captured by the camera are in the RGB color space. This organism in a healthy state (in a non toxic environment) is of green color and thus for better analysis the captured images are transformed into YCbCr color space and by observing the change in their color conclusions are made. Daphnia: The images are captured in the RGB space and transformed into Gray level space. The movement of this organism in a healthy environment is more as compared to a toxic environment.

Tools: Any image processing platform.

Conclusion: In this study by observing the physical attributes of an organism the toxicity of the water is determined by the author.

[2] 'AquaSight: Automatic Water Impurity Detection Utilizing Convolutional Neural Networks', Ankit Gupta , Elliott Ruebush , Department of Computer Science TJHSST Alexandria, USA, Department of Computer Science University of Maryland Bethesda, USA, July 2019.

Abstract: The author proposes AquaSight, a novel mobile application that utilizes deep learning methods, specifically Convolutional Neural Networks, for automated water impurity detection. After training with a dataset of 105 images, the deep learning algorithm achieved a 96 percent accuracy and loss of 0.108. This AquaSight system provided an efficient way for individuals to secure an estimation of water quality, alerting local and national governments to take action and potentially saving millions of lives worldwide.

Introduction: People in environments without easy access to purified water would benefit from technology allowing them to determine to what extent potential drinking water appears contaminated. In order to provide this technology and quickly determine the level of contamination of water based on an image, the author proposes AquaSight, a deep learning approach to water quality analysis that utilizes the power of modern hardware and machine learning techniques(CNN).

Methodology:
Dataset: Produced by the author itself consisting of 105 water images classified into Clean water, Colored water, contaminated with Sand, Salt, Black pepper, Oil paint.

Tools:

Keras machine learning library on top of a Tensorflow backend

The model was trained and tested locally on a PC running Windows 10. The NVIDIA Cuda Developer Toolkit was used with a GTX 1070 GPU and the Tensorflow-GPU library in order to accelerate model training
Result: The model had an accuracy of 96 percent (101/105 images) and a loss of 0.1077 after evaluating it on our image dataset.

Conclusion: The model produced easily understandable results, and the prediction values allowed the author to analyze how confident the model was in its choice, providing a spectrum of contamination levels. Additional statistical measures such as F-beta, Precision, and Sensitivity all yielded uniformly strong results as well, reiterating the effectiveness of the model.

[3] 'Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China', Xiaoping Wang, Fei Zhang , Jianli Ding, Scientific Reports.

Introduction: Water quality is a general term which refers to various attributes of water such as color, density, composition, etc. In this study the author has referred to the WHO standards provided under the term of Water Quality Index to define water quality.

Methodology:
Hyperspectral data collection: The FieldSpec3 ASD Spectroradiometer device is an optical sensor that uses detectors other than photographic film to measure the distribution of radiation in a particular wavelength region to measure the radiant energy level (radiance and irradiance).

Fractional Derivative method: Fractional derivative methods have been widely used in certain fields because models described by the fractional derivative are more accurate and efficient than methods based on integer derivatives. The most frequently used definitions are the following: Grunwald - Letnikov (G-L), Riemann - Liouville (R-L), and Caputo41. As it is less complex than the others, the G-L definition was employed in this study.

Calculation of the Water Quality Index: The Water Quality Index (WQI) is an extracted and estimated index that reflects the composite effects of all water quality parameters.

Tools: MATLAB2014a

Conclusion: Comparisons of the predictive effects of the 22 water quality index estimation models calibrated by POSSVR show that the model based on RI, DI, and NDI values of the 1.6 order is much better than the others while better predicting the water quality index of the study area (R2 (0.92), RMSE=58.4, RPD (2.81) and a slope of curve fitting of 0.97).

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

Abstract : The main objective of this study was to use some of the hydro-meteorological variables to compare estimates of hourly suspended sediment concentration in a river using sediment rating curve(SRC) and a model tree (M5') with different combinations of predictors.

Introduction : Sediments are usually of 2 types accordingly to the size viz. Wash and Bed material. Wash material consist of sediments smaller than 63 micron (clay and slit) which is carried away with the current on the surface and in the suspended part of the stream, on the other hand the bed material consist of sediments greater than 63 microns which is transported partially through the suspension part and the bed loading process such as rolling, sliding and saltation. These sediments are measured together using turbidity monitoring thus these sediments are termed to suspended sediments altogether. Factors controlling the amount of the suspended sediment in a system are classified into 3 categories viz. Hydrological (discharge, amplitude and timing), meteorological(rain, wind, etc.) and physiographic factors such as land use and soil properties. These factors are taken into consideration to determine the suspended sediment concentration (SSC). Many machine learning approaches have been used in this subject viz. SRC, SVR, ANN, RT, MT

The objectives of the present paper are to:

Investigate which regularly monitored hydro-meteorological variables may be good predictors of suspended sediment concentrations in the river;

Compare SRC and the M5' algorithm to estimate hourly SCC based on the selected predictors;

Suggested a model to estimate suspended sediment using commonly monitored variables to compensate in part for the lack of a suspended sediment monitoring program in the river

Methodology:

Study area: The area of the river, mean discharge, mean current velocity, max and min velocity, average annual precipitation and surrounding profile is obtained.

Monitoring: An autonomous probe equipped with the turbidity sensors.

Calibration of turbidity meters: Because of many factors affecting light scattering in a natural environment such as particle size, color and shape, turbidity should not be used without a good knowledge of its relation with SSC at the monitoring site. Therefore, a calibration curve was built to relate turbidity values to SSC

SRC: It is a non-linear empirical relation between discharge and suspended sediment concentration.

M5' Model tree: It is the combination of binary classification of the dependent variables with multivariate linear models. A decision tree is built and performance indicators are built and compared.

Model performances: Best estimation was provided by M5' over SRC.

Conclusion: The building of correlograms demonstrated the importance of the relation between SSC and lagged water levels, and accumulated rain. The calibration of M5' confirmed this relationship through the selection of the predictor. Water level rise expressed through the predictor was also proven to be valuable to estimate SSC in the river.

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

Abstract: Water resources management, hydraulic designs, environmental conservation, reservoir operation, river navigation and hydro-electric power generation all require reliable information and data about suspended sediment concentration (SSC). To predict such data, direct sampling and sediment rating curves (SRC) are commonly used.

Introduction: Suspended sediment carried in a river has an impact on various aspects of river use (e.g. water quality, navigation, fisheries and aquatic habitat). It is a site-specific problem that depends on several factors (e.g. the catchment area, rainfall intensity, vegetation cover) and should be studied for every river, creek or channel.

Methodology:
SRC: SSC=aQ^b
Simple linear regression
Multiple linear regression
ANN
Parameters: River temperature, flow rate, conductivity

Testing dataset: The testing dataset was also randomly selected; 50 weekly data records of the total data were used for each input and output variable in testing the various models to determine the best model. Statistics for temperature, discharge, electrical conductivity and suspended sediment concentration were recorded.
Indicators: MAE(Mean absolute error), RMSE, NSE(Nash Sutcliffe efficiency)
Tools: ANFISEDIT toolbox in MATLAB R2016b.

Results: The machine learning models ANFIS and ANN performed better than the other conventional models, SRC, SLR and MLR, in estimating SSC. Adding additional input variables to the major input improved the various model estimates of SSC.ANN is a superior approach.

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

Abstract: The paper presents 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 utilizes fish liver microscopic images in order to detect water pollution. The proposed approach consists of three phases; namely pre-processing, feature extraction, and classification phases. Also, it is implemented using Principal Components Analysis (PCA) along with Support Vector Machines (SVMs) algorithms for feature extraction and water quality degree classification.

Introduction: Water quality refers to the chemical, physical and biological characteristics of water. It is a measure of the condition of water relative to the requirements of one or more biotic species and/or to any human need or purpose. It is most commonly used with reference to a set of criteria that can assess compliance. The most common criteria used to assess the quality of water related to the health of ecosystems. The aim of this article is to design and implement an easy and rapid system to monitor and assess the quality of water through studying and classifying occurrence of different morphological changes in the fish liver.

Methodology:

Preprocessing: Resizing the images to 284x259 pixels, using background subtraction techniques, conversion to HSV color space.

Feature extraction: PCA algorithm application.

Classification: Support Vector Machines algorithm for water quality degree classification.

PCA (Principal component analysis): Principal Component Analysis is a statistical procedure, which is a type of dimensional reduction that is achieved by projection to lower dimensional space using linear transformation.

Loading the data.

Subtracting the mean of the data from the original dataset.

Finding the covariance matrix of the dataset.

Finding the eigenvector(s) associated with the greatest eigenvalue(s).

Projecting the original dataset on the eigenvector(s)

Color and texture features: Texture is one of the important features to determine objects or regions of interest in the image, and describes the visual patterns that contain important information about the structural arrangements of the surface and its relationship to the surrounding environment.

SVM: Support Vector Machines algorithm is a group of supervised learning models used for classification and regression analysis of high dimensional datasets as well as associated learning algorithms that analyze data and recognize patterns

Dataset: Collected datasets contain colored JPEG images of 125 images as training dataset and 45 images as testing dataset, respectively. Training dataset is divided into 4 classes representing the different histopathological changes and their corresponding water quality degrees. Experimental results showed that the proposed classification approach has obtained water quality classification accuracy of 93.3%, using SVMs linear kernel function with 37 images per class for training.

Result: Excellent water quality: where fish hepatocytes are rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong Periodic Acid Schiff (PAS) stain of glycogen reactivity reflected by deep magenta coloration.

Good water quality: Where fish exposed to copper at pH 9 and rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong PAS reactivity reflected less than control.

Moderate water quality: where fish exposed to copper at pH 7 and rich in glycogen content, hence the cytoplasmic area of the liver cells displayed a strong PAS reactivity reflected by moderate magenta coloration.

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

Abstract: In this study, the authors 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.

Introduction: Multiple linear regression (MLR) is a statistics based technique that uses several independent variables to predict the outcome of a dependent variable. The paper deals with the development, performance evaluation and validation of ANNs, and regression models for predicting sediment load. MATLAB (R2009a) software was used to model suspended sediment load.

Methodology:
Artificial neural networks: A neural network provides the relationship between a set of inputs and the respective outputs without giving any information about the actual processes involved. In artificial neural network an adder is used for adding the input signals. The amplitude range of the output of a neuron is normalized and is taken to be in the interval 0 to 1 or alternatively –1 to +1
Multiple Linear Regression: The multiple linear regression analysis was performed on the same data set to estimate sediment concentration.
Three performance indicators were used root mean square error (RMSE), correlation coefficient (r) (measures the linear relationship between two variables and the value is always between +1 to -1) and coefficient of efficiency (CE) (It is commonly used to assess the performance of rainfall runoff models).

Tools/algorithms used: Six ANN models were tried for each case using Levenberg- Marquardt (used to solve non linear squares problem) (shown in fig. below) as a training function with sigmoid as an activation function, subjected to maximum 1000 iterations and were trained with the help of back propagation learning algorithm(finds the minimum value of the error function) .

MLR (Multiple Linear Regression) (provides relationship between one dependent variable and two or more independent variable) sediment model.

Conclusion:

From the study, the author concluded that,

1) The ANN-2 provided better result as compared with the MLR models for estimating suspended sediment load.

2) The ANN models (2,11 & 17) showed better result as compared to all MLR models

3) The MLR models fit poorly for the data set under study.

4) It can be concluded that Artificial Neural Network models are superior to regression models in predicting suspended sediment load in all respects.

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

Abstract: In this study, the author has utilized the available dataset of the SedFoam multidimensional two-phase model (A three-dimensional two-phase flow solver, is presented for sediment transport applications) They have used linear regression and gradient boosting algorithm (Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble (method that uses multiple algorithms so as to obtain a better performance than could be obtained from any of the constituent algorithms alone). of weak prediction models, typically decision trees) so as to analyze the lowest average mean squared error and searched for the best method for the simulation setup.

Introduction: The two-phase models of sediments and waters follow two schemes: Eulerian and Lagrangian. The Lagrangian approach focuses on the individual particle movement, while the Eulerian approach emphasizes the flow at a specific point in space as a function of time. The study utilizes two ML methods, artificial neural network and model trees. The author has used linear regression and gradient boosting estimators for constructing the models. Linear regression is a standard and fast estimator, gradient boosting is a nonlinear approach to estimate the outputs.

The author has split the dataset into even parts where each part was grouped by height and also split into three sections with various sizes and hence recorded the average mean squared error of the models for the accuracy analysis.
According to the author, partitioning the dataset by the domain height is essential since different height regions will have different initial concentration field.

Methodology: This includes Computational Domain Configuration and oscillatory sheet flow setup. This thesis used XGBoost (eXtreme Gradient Boost) library as an estimator for its being robust to noise, nonlinear, and a tree-based method(based on the gradient boosting algorithm) .According to the author ,the purpose of this paper is to apply the available dataset in the SedFoam model [28] on assessing the effectiveness of ML algorithms to predict sediment transport and of partitioning the dataset.

Dataset format: The dataset is from the 2D SedFoam model that runs with an initially flatbed. The dataset showed multiple agreements between the computed dataset with the measured one.

Tools/algorithms used: The assessment utilized linear regression and gradient boosting algorithm to analyze the lowest average mean squared error, artificial neural network and model trees, XGBoost (eXtreme Gradient Boost) (XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.

Results and observations: Result of MSE (Mean squared error) is always non- negative and the model is more accurate as MSE values get closer to zero. Statistical analysis includes mean square error. A cross-validation approach, namely the k-fold cross-validation method (technique which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance). , is utilized to assess several sediment transport parameters: the concentration, horizontal fluid velocity, vertical fluid velocity, horizontal sediment velocity, and vertical sediment velocity.

Conclusion: In this paper, author has analyzed various statistical machine learning methods that can be used in to solve bigger problems. The author used various models built on various partitions of the dataset based on the vertical height which improves the performance of the model. The results of experiments show the model to be consistent with the physical laws on how sediment transport occurs. The results show that using separate models for the points with different rates of sediment transport improves the model.

[9] 'Machine learning approaches for anomaly detection of water quality on a real-world data set', Fitore Muharemi, Doina Logofătu & Florin Leon,Journal of Information and Telecommunication.

Abstract: In this study, the following models are applied to water quality data: logistic regression (is a supervised learning classification algorithm used to predict the probability of a target variable), linear discriminant analysis (LDA) (used to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier), support vector machines (SVM) (mainly used for classification and regression problems), artificial neural network (ANN), deep neural network (DNN), recurrent neural network (RNN) and long short- term memory (LSTM) (is a form of RNN with a more complex cell architecture for more accurately maintaining the memory of important correlations(relationship between two variables)).The performance evaluation is conducted using F-score metric (In statistical analysis of binary classification, the F score is a measure of a test's accuracy). A simulation study is conducted to check the performance of each algorithm using F-score. The results show that all algorithms are vulnerable although SVM, ANN and logistic regressions tend to be a little less vulnerable, while DNN, RNN and LSTM are very vulnerable.

Introduction: The experiment was conducted in two parts. First, for the classification author used the statistical algorithm logistic regression. Second, the author extends the experiment by using machine learning techniques, ANN, DNN, RNN, LSTM, and LDA to compare if they can outperform the logistic regression for this specific data set.

Methodology: The methodology of solving the problem includes seven classification algorithms on the same classification task. When sufficiently representative training data were used, most algorithms perform reasonably well, but in the experiment even though author had a large data set, not every algorithm gave many promising results. The goal of this work was to find the most suitable algorithm for the problem under investigation. In this research, author used the F1 score, considered as one of the best performance metrics for classification algorithms, especially good for imbalanced data.

Tools/algorithms used: Boruta algorithm (The Boruta algorithm is a wrapper built around the random forest classification algorithm. It tries to capture all the important, interesting features we might have in our dataset with respect to an outcome variable. First, it duplicates the dataset, and shuffle the values in each column.), Logistic Regression, Support Vector Machines, Linear Discriminant Analysis, Neural Networks algorithm, Recurrent Neural Network, Deep Neural Network, and Long Short-Term Memory.

Results and observation:

The goal of this work was to find the best performing model for water quality data, and check whether machine learning models are more accurate than logistic regression.

Conclusion: In this work, author presented approaches for event detection on water quality time series data. This work represents a case study and its aim is to find the best model on anomaly detection on water quality systems. Time series are analyzed using statistical algorithms for a long time.

[10] 'Classification of Surface Water using Machine Learning Methods from Landsat Data in Nepal', Tri Dev Acharya 1, Anoj Subedi 2, Huang He 3 and Dong Ha Lee , MDPI.

Abstract: In this study, author tried to use satellite image from Landsat 8(American Satellite) to classify surface water in Nepal. Input of Landsat bands and ground truth from high resolution images available from Google Earth is used.

Introduction: For the identification of surface water using Landsat image, author used various techniques in previous studies, such as, using water index methods (methods are mostly used for surface water estimation which separates the water from the background based on a threshold value). single or combined, decision tree based classification and segmentation of scene in diverse area of Nepal.

Methodology: A total of 800 ground truths with 614 non-water and 186 water points within the scene were extracted as from high resolution images available form Google Earth. Feedforward neural networks are simpler than their counterpart, recurrent neural networks. They are called feedforward because information only travels forward in the network (no loops), first through the input nodes, then through the hidden nodes (if present), and finally through the output nodes.

Tools used: Random Forest, Recursive Partitioning (is a statistical method for multivariable analysis), Support Vector Machine, Neural Networks.

Results and Observations: The result of the overall accuracy from all the four methods shows that random forest performs with 1 as highest and SVM performs lowest with 0.926. Both RPART (Recursive Partitioning) and NN showed overall accuracy of 0.95. With same training points, there is a vast improvement against index methods single or combined. However, the segmentation accuracy is still higher i.e. 0.96 against machine learning methods except random forest.

Conclusion: In this study, application of four machine learning methods: RF, RPART, SVM and NN were used to derive the surface water map using a Landsat 8 image in Nepal. In addition, random forest has shown maximum overall accuracy 1 for the scene with given dataset.

[11] 'Suspended sediment load prediction using non-dominated sorting genetic algorithm II', Mahmoudreza Tabatabaein, Amin Salehpour Jam, Seyed Ahmad Hosseini, International Soil and Water Conservation Research.

Abstract: In this study ,in order to increase the efficiency of SRC model, a multi- objective optimization approach is proposed using the Non-dominated Sorting Genetic Algorithm II (NSGA-II) algorithm. In the first part of the study, the data was clustered and classified in 70% and 30% for use in calibration and evaluation of SRC models, respectively. In the second part of the study, two different groups of SRC model comprised of conventional SRC models and optimized models (single and multi-objective optimization algorithms) were extracted from calibration data set and their performance was evaluated. The comparative analysis of the results revealed that the optimal SRC model achieved through NSGA-II algorithm was superior to the SRC models.

Introduction: The standard SRC model is the power regression equation ,which is firstly calibrated by taking logarithm of variables corresponding to sediment and flow discharge and formulating a linear regression, and lastly ,the linear regression coefficients is computed with the error least square technique.

Tools used: Non-Dominated Sorting Genetic Algorithm II (NSGA-II) (The NSGA-II algorithm is recognized as an algorithm which is a well known, fast sorting and elite multi objective genetic algorithm and simultaneously optimizes each objective without being dominated by any other solution).

Conclusion: SRC model in which NSGA-II algorithm has been used showed better efficiency.

[12] 'Efficient Water Quality Prediction Using Supervised Machine Learning', Umair Ahmed 1, Rafia Mumtaz 1,* , Hirra Anwar 1, Asad A. Shah 1, Rabia Irfan and José García-Nieto, MDPI.

Abstract: In this research paper the author 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 is determined on the basis of the WQI. The method used by the author gives 4 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 is used by the author.

INTRODUCTION: The author has chosen an alternative and inexpensive method to predict the water quality in real time over the time consuming and expensive methods.

Source of DATA: The dataset collected from Pakistan Council of Research in Water Resources (PCRWR) contained 663 samples from 13 different sources of Rawal Water Lake collected throughout 2009 to 2012. It contained 51 samples from each source and the 12 parameters (Alkalinity, Appearance, calcium, chlorides, conductance, fecal coliforms ,hardness as CaCO3,nitrite(NO2-),pH ,Temp ,Total dissolved solids ,turbidity)

Methodology: Author obtained data from PCRWR and cleaned through box plot analysis and then normalized using q-value normalization for converting them into scale of 0-100 to calculate the WQI using six available parameters. Once the WQI was calculated, all original values were normalized using z-score, so they were on the same scale and further using classification algorithms, WQC was calculated.
Algorithms used: 8 regression and 10 classification algorithms:
Multiple Linear Regression, Polynomial Regression, RF, Gradient Boosting, SVM, Ridge Regression, Lasso Regression, Elastic Net Regression, Neural Networks/Multi-Layer Perceptron (MLP), Gaussian Naïve Bayes ,Logistic Regression, Stochastic gradient descent, K Nearest Neighbor, Decision Tree, Bagging Classifier.

Results: The author through study came to a conclusion that gradient boosting and polynomial regression performed better in predicting WQI, whereas MLP performed better in predicting WQC.

Future scope: In future works, author proposed, integrating the findings of the research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and Water communicate those readings using an Arduino microcontroller and ZigBee transceiver.

[13] 'Machine Learning – Based Detection of Water Contamination in Water Distribution Systems', Hadi Mohammed1, Ibrahim A. Hameed2, Razak Seidu1, Researchgate.

Abstract: This paper develops adaptive neuro fuzzy inference systems(ANFIS) models for detecting the safety condition of water in pipe networks when concentrations of water quality variables in the pipes exceed their maximum thresholds. The detection done by the author is based on time series data composed of pH, turbidity, color 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.

Introduction :

Objectives of Author's study:

1. Evaluate the effects of variations in the water quality parameters on the safety condition of water in the distribution network using Pearson's correlation analysis.
2. Use of adaptive neuro-fuzzy inference system (ANFIS) model to detect deviations of water quality from baseline values established in Norwegian water safety regulations.

Methodology:

1.Study Area and Data Set

The data used in this study consist of monthly measurements at the outlet of the Ålesund water treatment plant(WTP) and the various sampling locations in the distribution pipes. The data, which were taken from January 2013 to June 2017, were composed of water pH, turbidity (NTU), color (mg Pt/l) and counts of total bacteria (counts/ml) in the clean water. To obtain enough data for training and testing the models, we merged the data from all the locations to constitute 504 data samples. Descriptive Statistics of Data: Pearson correlation coefficients (r) Adaptive Neuro-Fuzzy Inference System (ANFIS): ANFIS optimizes the distribution of membership functions by using a back-propagation gradient descent algorithm either alone or combined with a least-squares method.

Results: Results were classified as : Descriptive statistics, ANFIS model results.
Accuracy : ANFIS models can correctly detect between 92%
and 96% of the safety condition of the water in the pipe network, with approximately 1% false alarm rate during the testing stage.

Conclusion: This paper presented the development of classification models for predicting the safety condition of water in distribution pipes. The models, based on ANFIS technique, were built using water quality variables measured from the effluent of the water treatment plant in Ålesund, Norway, as well as seven different locations across the pipe network.

[14] 'Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine', Kyle t. Peterson , Vasit Sagan , Paheding Sidike, Amanda L.Cox, Megan Martinez, MDPI.

Abstract: This study developed a method for quantifying SSC based on Landsat imagery and corresponding SSC data obtained from United States Geological Survey monitoring stations from 1982 to present.t uses feature fusion based on canonical correlation analysis to extract pertinent spectral information, and then trains a predictive reflectance–SSC model using a feed-forward neural network (FFNN), a cascade forward neural network (CFNN), and an extreme learning machine (ELM).

Introduction: Machine learning (ML) such as neural networks capture both linear and non-linear relationships. Neural network regression methods have been proven to outperform traditional techniques when applied to a wide range remote sensing studies.

Objectives of Author's study:

1. To utilize freely available Landsat multispectral data to produce a highly predictive reflectance-SSC modeling paradigm for large fluvial systems using ML techniques.

The author's study also intended to address issues related to modeling a wide range of SSC concentrations using a single ML algorithm.

Using ML methods , SSC maps were produced to analyze the results generated for each Landsat sensor.

Tools used : ELM,FFNN,CFNN.

Methodology:

1. Study Area- For creating predictive reluctance-SSC model, six of USGS sampling stations were selected. The criteria for author's selection of USGS stations were :

a) SSC must be directly measured and not estimated from turbidity.

b) channel width must be at least 100 m to resolve Landsat 30 m spatial resolution imagery.

c) the USGS station must contain at least one year of data collection to appropriately represent the site characteristics.

2. The predictive SSC model from Landsat reflectance data consists of several stages: a) Suspended Sediment Data - Taken from USGS(US GEOLOGICAL SURVEY) and NWIS(National Water Information System). The filtered images in the database were compared to the SSC database to select matching pairs for the regression modeling analysis

Suspended sediment modelling.

Feature Fusion

Regression Modeling

Quantitative Evaluation

Results: ELM generated the highest R2 and lowest RMSE values in all cases when applied to the testing data. Compared to RF and SVM, the ELM model performed slightly better in terms of R2 but displayed significantly lower RMSE. Results for the FFNN and CFNN also displayed significant R2 correlations but produced much higher RMSE values than RF, SVM, and ELM.

Conclusion: The results of the SSC modeling showed that ELM outperformed both FFNN and CFNN, as well as the popular ML techniques such as RF and SVM by significant margins when evaluated Remote Sens. 2018, 10, 1503 14 of 17 by R2 and RMSE. ELM models for all Landsat sensors generated R2 values above 0.9 and displayed noteworthy generalization abilities along with negligible overfitting.

[15] 'Single-Image-Referenced Colorimetric Water Quality Detection Using a Smartphone', Volkan Kilic, Gazihan Alankus, Nesrin Hozrum, Yusuf Mutlu, Researchgate.

Abstract/Main idea - In this paper , the author 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. Author developed an app for processing this image for training and creating reference models .Author also said to have achieved approx. 100% accuracy.

Introduction - Author demonstrated several smartphone colorimeter designs for water quality sensing. Apart from this author obtained color data from paper-based sensors, where color change is quantified by parameters in various color spaces, that is RGB, HSV, and L*a*b*.Author proposed a new methodology for simple colorimetric detection of water quality using smartphone-embedded color matching algorithms on JPEG images. Author claims that his study is distinct by using simple color matching algorithms to detect the concentration in real time with rapid response necessary for lab-on-a-chip systems. Author also gives citations of 'Salles et al' and 'Helfer et al' who used hierarchical clustering analysis and principal component analysis (PCA) and employed linear correlation respectively.

Methodology - Author prepared set up cardboard painted white from inside.

Four different colorimetric assays of nitrite (NO_2^-), phosphate (PO_4^{3-}), hexavalent chromium (Cr(VI)), and phenol were prepared according to the colorimetric standard methods . The author made an app to test the samples for single image reference . Users can train models for multiple sample types and later use these models to measure concentration levels in newly acquired photographs. To measure a solution of unknown concentration on a new image, the previously trained model that is appropriate for the solution to be tested first needs to be selected. Then capture the photo from mobile. The app then uses the trained model along with color matching algorithm to calculate the concentration level of the solution. The author faced the challenge to finalize the results so he used two different color matching algorithms with low computational complexity and compared them to select the best method to be run on a smartphone which were CC(correlation coefficient)method and CIELAB.

Algorithm - First, the cropped patch from the test assay was converted to L*a*b* color space as all reference patches were in that format. For ΔE* score calculation, ΔL*, Δa*, and Δb* were calculated by subtracting L*, a*, and b* channels of test and reference images, respectively. By averaging ΔE* scores of the pixels, final ΔE* scores between test and reference image were obtained. This process was repeated for all reference images. Therefore, one test image had several ΔE* scores, and the smallest one points to the most similar matching. Similar process was repeated for the CC method and correlation coefficients between test image and reference images were calculated, and the highest coefficient indicates the highest similarity.

Conclusion - In this paper , the author developed a smartphone based platform water quality detection . main features can be summarized as follows:
The user can create reusable experimental models for respective samples. Multiple regions of interest can be accurately selected in real. Rapid and reliable results are calculated using pretrained model.

[16] 'Continuous monitoring of suspended sediment concentrations using image analytics and deriving inherent correlations by machine learning', Mohammad Ali Ghorbani, Rahman Khatibi, Vijay P.Singh, Ercan Kahya, Heikki Ruskeepaa, Mandeep Kaur Saggi, Bellie Sivakumar , Sungwon Kim, Farzin Salmasi, Mahsa Hasanpour Kashani, Saeed Samadianfard, Mahmood Shahabi, Rasoul Jani, Scientific Reports.

Feasibility of proposed solution by authors: Authors used already available data sets. They transformed images using image analytics (needed for conversion of unstructured big data into readable machine data) for extracting information to form a modelling dataset. By constructing predictive models by learning inherent correlation between observed SSC values and their image analytics.

Abstract: The capability for monitoring SSC is based on photometric features of Red, Green, and Blue (RGB), where modern cheap high resolution cameras are capable of capturing subtle changes in tone and color, both expressed in bits. RGB imaging tracks 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 is explored for the prediction of SSC. In the research done by authors, image analytics comprise 8 input variables, and comprise: 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.

Algorithms: This paper used 2 models DRF and GLM to predict the SSC using image analytics in the H2O platform. GLMs connect multivariable inputs to outputs in their predictor mode for regression analysis and DRF connects multivariable inputs to outputs in their predictor mode, which generates a forest of regression trees, rather than a single regression tree. Author has used the H2O platform because it removes the requirement of the normalization of the error distribution by adding an explicit error term as a function of mean , non-normal errors, and a non-linear relation between the response and covariates

Results & observations: 166 observations, which were divided randomly into 111 training data points and 55 testing data points

Conclusion: This paper presented evidence for the proof-of-concept for a continuous monitoring capability of Suspended Sediment Concentration (SSC).It tested the transformation of high-resolution images of the fows carrying suspended sediments through a laboratory fume into image analytics to serve as input data into machine learning models. With this paper, the author offers evidence for existence of correlation between subsequent image analytics and SSC values and shows the correlation to be strong enough.

Further scope : The paper by author, however, identified some problems for predicting at higher SSC values that need to be solved. The authors attribute this to possible shortfalls in achieving steady state flows at higher concentration, which can be investigated by attention to the following aspect: (i) using larger fumes with higher capacities; (ii) standardizing the laboratory procedure to ensure that suspended matters are well mixed and steady state is ensured; (iii) testing the performance of different suspended matter.

2.2 SUMMARY

Paper	Title	Algorithm	Method	Data Set	Complexity
1	Water quality assessment using Image Processing.	Color-space transformation and motion analysis	Image processing	RGB images	Simple
2	AquaSight: Automatic Water Impurity Detection Utilizing Convolutional Neural Networks	Convolutional Neural Network	Machine Learning	105 samples	Moderate
3	Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China	Support Vector Regression	Remote Sensing, WQI, Hyper-spectral analysis	48 water samples classified into 5 categories	Complex
4	Estimation of suspended sediment concentration in the Saint John River using rating curves and a machine learning approach	Model Tree, Regression	Suspended Sediment Concentration, Rating Curves, Machine Learning	Comparison Study	Moderate
5	Suspended Sediment Concentration Modeling Using Conventional and Machine Learning Approaches in the Thames River, London Ontario	ANFIS, ANN	Comparison between ANFIS, ANN & SRC, Linear Regression	Training: 420 Testing: 50 data records	Complex

Table. 2.1

Paper	Title	Algorithm	Method	Data Set	Complexity
6	Machine Learning based Approach for Water Pollution Detection via Fish Liver Microscopic Images Analysis	Support Vector Machines	Principle Component Analysis	Training: 125 images Testing: 45 images	Moderate
7	Modelling river suspended sediment load using artificial neural network and multiple linear regression: Vamsadhara River Basin, India.	Multiple Linear Regression, ANN, Levenberg-Marquardt Algorithm	Root Mean Square error, Correlation Coefficient, Coefficient of efficiency	105 samples	Moderate
8	A Machine Learning Assessment to Predict the Sediment Transport Rate Under Oscillating Sheet Flow Conditions	Linear Regression & Gradient Boosting algorithm	Computational Domain Configuration, Eulerian & Lagrangian approach	Sedfoam multidimensional two-phase model	Complex
9	Machine learning approaches for anomaly detection of water quality on a real-world data set	Boruta Algorithm, Logistic regression, Support Vector Machines, Linear discriminant analysis, Neural networks algorithm, Recurrent neural network, deep-neural network, Long short term memory	Performance conducted using F-score metric	Comparison Study	Moderate
10	Classification of Surface Water using Machine Learning Methods from Landsat Data in Nepal	Random-Forest, Recursive partitioning, Support Vector Machines and Neural networks	Feed-forward neural networks	800 ground truths with 614 non-water and 186 water points available from Google Earth	Complex

Table. 2.2

Paper	Title	Algorithm	Method	Data Set	Complexity
11	Suspended sediment load prediction using non-dominated sorting genetic algorithm II	Non-Dominated Sorting Genetic Algorithm II	Evaluation of SRC model	Comparison Study	Complex
12	Efficient Water Quality Prediction Using Supervised Machine Learning	Multiple Linear Regression, Polynomial Regression, RF, Gradient Boosting, SVM, Ridge Regression, Lasso Regression, Elastic Net Regression, Neural Networks/Multi-Layer Perceptrons (MLP), Gaussian Naïve Bayes ,Logistic Regression, Stochastic gradient descent, K Nearest Neighbor, Decision Tree, Bagging Classifier.	Box plot analysis, q-value normalization, Water Quality Index	Collected from Pakistan Council of Research in Water Resources(PCRWR) contained 663 samples from 13 different sources of Rawal Water Lake collected throughout 2009 to 2012	Simple
13	Machine Learning – Based Detection of Water Contamination in Water Distribution Systems	Adaptive neuro fuzzy inference systems(ANFIS)	Data composed water pH, turbidity (NTU), color (mg Pt/l) and counts of total bacteria (counts/ml) in the clean water.	504 data samples	Complex
14	Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine	Extreme learning machine (ELM), Cascade forward neural network (CFNN), Feed-forward neural network (FFNN)	Using ML methods , SSC maps were produced to analyze the results generated for each Landsat sensor.	SSC data obtained from United States Geological Survey monitoring stations	Complex

Table. 2.3

15	Single-Image-Referenced Colorimetric Water Quality Detection Using a Smartphone	Color matching algorithms (CC and CIE)	Comparing correlation coefficient method and CIE method	Self prepared	Moderate
16	Continuous monitoring of suspended sediment concentrations using image analytics and deriving inherent correlations by machine learning	Generalized Linear machine (GLM) and Distributed random forest (DRF)	Comparing models by using datasets	166 points; 111 training and 55 testing points	Moderate

Table. 2.4

CHAPTER 3

DESIGN AND DRAWING

3.1 INTRODUCTION

This work mainly focuses on identification of water quality along with its classification and mixture composition where we emphasize the probable contaminants which can be present in the contaminated water. Further spread velocity of impurities is detected at regular intervals of time. The complete process of identifying the water quality and classification of water contamination is illustrated with the help of flowchart as shown in fig.3.1 below.

1) Dataset Collection: For predicting the water quality, a compact dataset of 209 contaminated and 50 non-contaminated images was prepared. For the classification of water contamination, 850 images representing 6 classes of color was prepared.
2) Data processing and model creation: Data (images) are stored into respective classes viz. contaminated, non-contaminated and other color classes. Each image is rescaled and stored into separate directories. A sequential model has been implemented from the Tensorflow and Keras machine learning library.
3) Training, Testing and Validation: After using the dataset for water quality determination, validation accuracy of 90.38% and loss of 0.2273 was achieved. After predicting whether the water is contaminated or not, the system further predicts the color of water contamination if found contaminated. After implementing the model the system achieved a validation accuracy of 88.27% and validation loss of 0.3644.
4) Color Detection: This model detected the color and hence a description file has been provided to indicate the possibility of the probable contaminants in water according to the respective color detected.
5) For the classification of contamination in water bodies, the model predicted the color of the contamination of the given assay. 6 classes of color contaminations viz. Black, Blue, Brown, Green, Red.

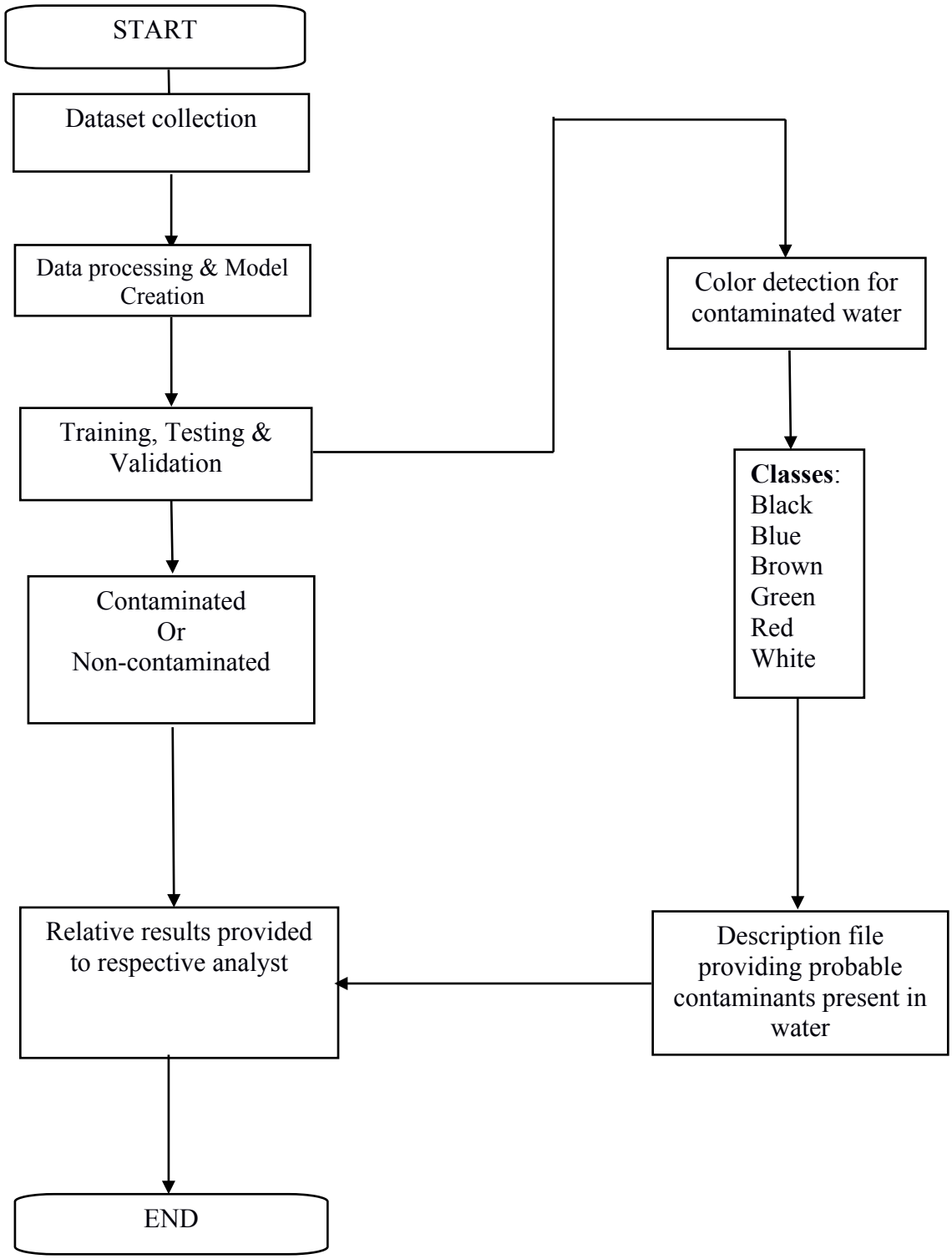

Fig.3.1. Flowchart representing water quality assessment

CHAPTER 4

IMPLEMENTATION

4.1 INTRODUCTION

The implementation begins with the involvement of three objectives which are identification of the quality of the water, proceeding with the classification of contaminated water and finally calculating the velocity of liquid impurity.

Firstly, the system predicts the water quality i.e. whether the water is contaminated or non-contaminated. So in order to train the model two classes viz. contaminated and non-contaminated were created and respective images were loaded into them accordingly. Datasets comprised 209 contaminated images and 50 non-contaminated images which were then trained and validated. The system involved Convolution Neural Network (CNN) using which the validation accuracy of 90.38% and a validation loss of 0.2273 was achieved. The system was able to predict that which image represented contaminated and non-contaminated water and thus provided the output as shown in the fig.4.1 and fig.4.2 below.

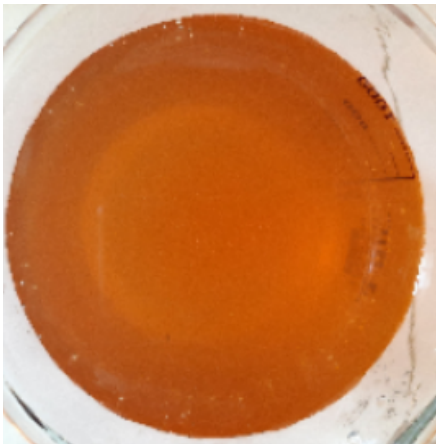

Fig.4.1
Contaminated water

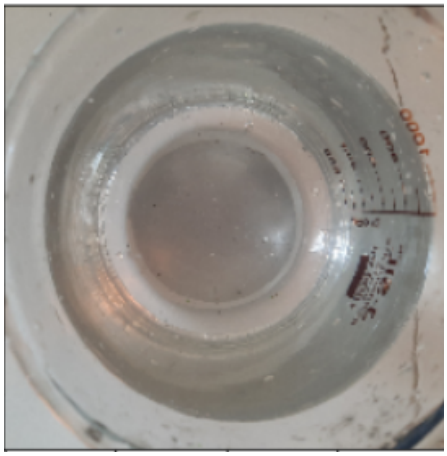

Fig.4.2
Non-contaminated water

The overall summary involved total parameters as 17,358,881 out of which 17,358,881 parameters were trainable thereby leaving no non-trainable parameters. 4 layers were used which are Convolution 2d layer, Maxpooling 2d layer, Flattening Layer and the Dense layer. This model summary is obtained using the model.summary() function used in Jupyter Notebook and is represented in the following fig.4.3 below

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 198, 198, 16)      448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 99, 99, 16)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 97, 97, 32)        4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 48, 48, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 46, 46, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 23, 23, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 33856)             0
_________________________________________________________________
dense (Dense)                (None, 512)               17334784
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 17,358,881
Trainable params: 17,358,881
Non-trainable params: 0
_________________________________________________________________
```

Fig.4.3. Model Summary representing water quality

In this model, the algorithm used was Convolutional Neural Network (CNN) which is a deep-learning algorithm which takes an input as image assigned learnable weights and biases to various objects in the image and able to segregate one from the other. The images were preprocessed because of which CNN will have the ability to learn filters and characteristics. The primary role of CNN is to reduce the image into a form which is much easier to process without missing on any of the features of the image to produce a highly efficient model.

Next, our project dealt with the classification of the images detected as contaminated water. The images were classified into 6 color classes viz. Blue, Black, Brown, Red, Green, White. The system having trained, predicted the color of contamination and based on the color, a description file was provided so as to indicate the probable contaminants present in water. Datasets comprised 850 images representing 6 color classes as shown in the fig.4.4 below

Fig.4.4. 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, defining the model by implementing different layers such as conv_2D, Maxpooling_2D, Dropout, flattening and dense layer. The next step involves the compilation of model and its summary as shown in fig.4.5 below. After training the model by considering epochs with 20, validation accuracy was found to be 88.27% and thus was able to predict the color 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
_________________________________________________________________
```

Fig.4.5. Model summary representing classification of contaminated water

This project further involves the determination of velocity of the impurities present in the water. In this objective, the average spread velocity of liquid impurity (ink) was determined by using a small image dataset of ink diffusion in the water where using image processing techniques, the RGB images 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 as shown in the fig.4.6 and fig.4.7 given below.

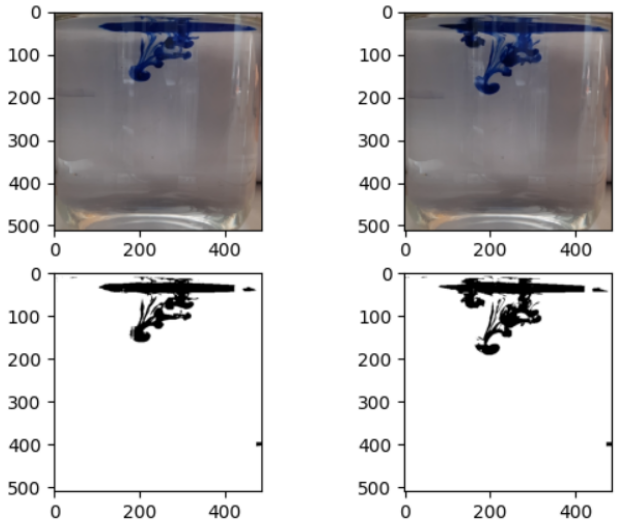

Fig.4.6
(Side view indicating the spread)

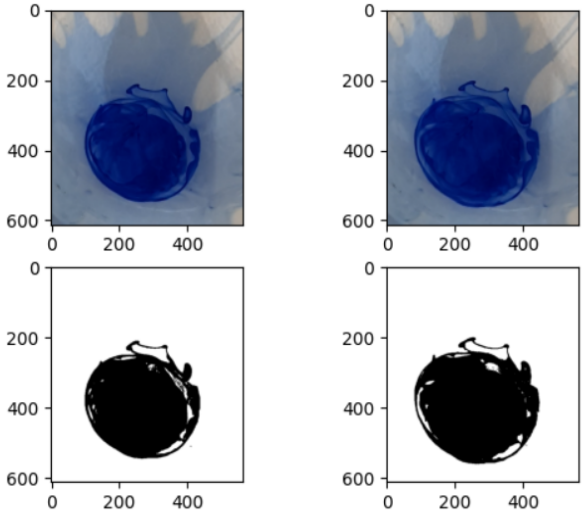

Fig.4.7
(Top view indicating the spread)

First, two images at a specific 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. Finally, the average spread velocity was calculated using the formula given below:

avg_spread_velocity_2D = spread_area_2D / time interval

CHAPTER 5

EXPERIMENTATION

5.1 INTRODUCTION

Beginning with the first objective i.e. to determine the quality of the water, a comprehensive dataset consisting of 209 contaminated images and 50 non-contaminated images was prepared. These contaminated images were prepared using various ingredients such as mud, ink, glitter, chilly flakes, detergent, coffee powder, brown powder, turmeric, crayon dust, various water colors such as blue, black, brown, red, green, pink and yellow. These ingredients were used in order to make the water contaminated.

All such ingredients were poured into a glass beaker which consisted water. After mixing these ingredients in the water thoroughly, images were captured of the same by adjusting with the lighting conditions accordingly. Different combinations were made for example, mud, along with some crayon dust, mud with small quantity of detergent to make the water look turbid etc. These images were captured using high resolution camera of 48MP. Thus, the image dataset proved to be well trained and hence was able to come up with a good accuracy after validation.

Further, for the classification of the contaminated water, a total of 850 images with only 6 colors viz. blue, black, brown, red, green and white were considered since the system had to classify into 6 different color classes. The entire classification and determination of water quality was achieved using Convolutional Neural Networks (CNN). Some images were developed manually whereas some were readily available as dataset. These images were then collectively trained and validated for further process using machine learning approach.

A real world dataset of 89 images was prepared to analyze 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. Some of the images are shown below in fig.5.1. When deployed,

the system was able to classify 66 images out of the total 89 images, based on the 6 colors passed as the parameters along with

Fig.5.1 Real world image data

the probable contaminants. The model attained an accuracy of around 74.15% and loss of 0.2584.

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. Ink was considered as a liquid impurity in this project 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. These images were then processed using image processing techniques where these RGB images were converted into gray scale images for simpler calculation as shown in fig.4.6 and fig.4.7 above. Image processing techniques such as resizing was performed so as to perform the operation of image subtraction.

5.2 TECHNICAL SPECIFICATION & SOFTWARE REQUIREMENTS

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 and Jupyter Notebook were used for water quality determination and classification process whereas for finding velocity of liquid impurity Python 3.6.12 and 3.7.4 was used as the programming language.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 RESULTS

SR.NO	OBJECTIVE	ACCURACY	LOSS
1	**Water Quality determination**	**90.38%**	**0.2273**
2	**Classification of contaminated water & mixture composition**	**88.27%** **74.15%** (based on real world dataset)	**0.3644** **0.2584** (based on real world dataset)

Table. 6.1

[2] Ankit Gupta et al performed an experiment using 105 images and achieved accuracy of 96% and loss of 0.107 by considering datasets involving contaminated water images which comprised sand, sand and salt, sand, salt and black pepper etc. Our model achieved an accuracy of 90.38% and loss of 0.2273. The model was well trained where contaminated images were successfully predicted as contaminated image which is same for non-contaminated image as well. However, some contaminated images were predicted incorrectly as non-contaminated image. The output of the model predicting contaminated and non contaminated images is shown in fig.6.1(a) and 6.1(b) below.

SR.NO	OBJECTIVE	TYPE OF SPREAD	TIME INTERVAL (sec)	SPREAD VELOCITY (pixel/sec)
3	**Calculation of velocity of liquid impurity (ink)**	**Vertical Spread**	**0 to 2**	**2345**
			2 to 4	**1442.5**
		Surface Spread	**0 to 2**	**9226.5**
			2 to 4	**7936.5**
			4 to 6	**11475.5**
			6 to 8	**6203**

Table. 6.2

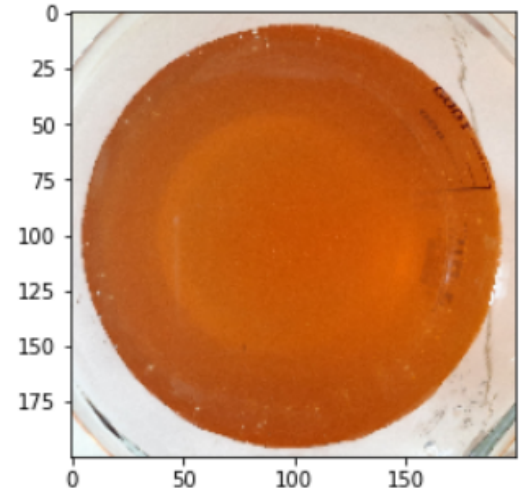

Fig.6.1.(a) Simulation result of identification of water quality

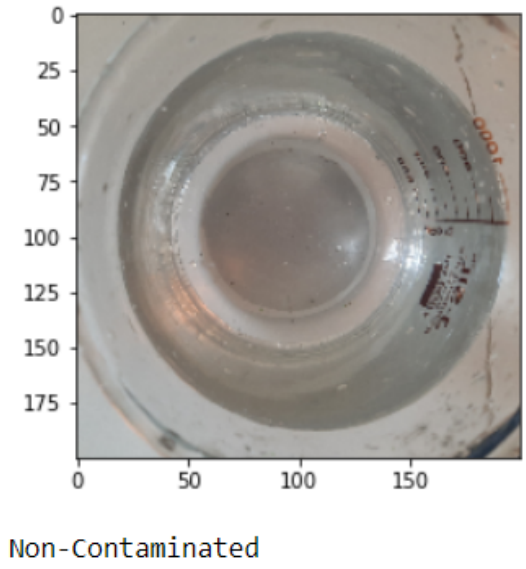

Fig.6.1.(b) Simulation result of identification of water quality

Fig.6.2. Graphs representing Training Accuracy vs Validation Accuracy and Training Loss vs Validation Loss for classification of contaminated water

This is most likely a Brown water image with a 99.08 percent confidence.
It might contain traces of : mud, rust, tannins, suspended sediments

This is most likely a Black water image with a 96.53 percent confidence.
It might contain traces of : water softeners, sulphides, manganese, clay, cement

This is most likely a Green water image with a 100.00 percent confidence.
It might contain traces of : algae, phytoplankton, chlorine

Fig.6.3.(a) Simulation results for classification of contaminated water

This is most likely a Black water image with a 99.70 percent confidence.
It might contain traces of : water softeners, sulphides, manganese, clay, cement

This is most likely a Green water image with a 65.80 percent confidence.
It might contain traces of : algae, phytoplankton, chlorine

Fig. 6.3.(b) Simulation results for classification of contaminated water

6.2 DISCUSSION

We know that water contamination has become a serious issue in today's world. Various methods have been proposed earlier to solve this problem. In this project, we have proposed machine learning approach as well as image processing techniques to address this issue. Going through this project we faced many challenges regarding the dataset collection and accuracy. To overcome this challenge, we have prepared our own dataset in the lab for water quality determination and classification. We have also tracked the spread velocity of liquid impurity by taking ink as an assay. We calculated the spread velocity at different time intervals by using image processing . From this, we observed that there is a non-linear relationship in the spread velocity at definite time intervals i.e. the velocity is not constant.

CHAPTER 7

CONCLUSION

From the above mentioned results and observations, it can be said that the model was well trained and therefore it can be concluded that CNN proved to be worthy in providing a good validation accuracy of 90.38% with loss of 0.2273 (Refer Table.6.1). The model was also able to differentiate that which images were representing contaminated and non-contaminated water.

Further 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 (Refer Table.6.1). 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 report, was successful and hence we can also conclude based on the achieved results (Refer Table.6.2) 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.

In this conclusion, this report represents a strong start to the ongoing project to develop an effective system for efficient and easy water quality assessment.

REFERENCES

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

[2] Ankit Gupta , Elliott 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

[3] Xiaoping Wang, Fei Zhang , Jianli Ding, 'Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China', Scientific Reports

[4] Sébastien Ouellet-Proulx , André St-Hilaire , Simon C. Courtenay , Katy 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

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

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

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

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

[9] Fitore Muharemi, Doina Logofătu & Florin Leon ‘Machine learning approaches for anomaly detection of water quality on a real-world data set’, Journal of Information and Telecommunication

[10] Tri Dev Acharya, Anoj Subedi, Huang He and Dong Ha Lee ‘Classification of Surface Water using Machine Learning Methods from Landsat Data in Nepal’, MDPI

[11] Mahmoudreza Tabatabaein, Amin Salehpour Jam, Seyed Ahmad Hosseini ‘Suspended sediment load prediction using non-dominated sorting genetic algorithm II’, International Soil and Water Conservation Research

[12] Umair Ahmed, Rafia Mumtaz, Hirra Anwar , Asad A. Shah , Rabia Irfan and José García-Nieto ‘Efficient Water Quality Prediction Using Supervised Machine Learning’, MDPI

[13] Hadi Mohammed, Ibrahim A. Hameed, Razak Seidu ‘Machine Learning – Based Detection of Water Contamination in Water Distribution Systems’, Researchgate

[14] Kyle t. Peterson , Vasit Sagan , Paheding Sidike, Amanda L.Cox, Megan Martinez, ‘Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine’ ,MDPI

[15] Volkan Kilic , Gazihan Alankus, Nesrin Hozrum, Yusuf Mutlu ‘Single-Image-Referenced Colorimetric Water Quality Detection Using a Smartphone’, Researchgate.

[16] Mohammad Ali Ghorbani, Rahman Khatibi, Vijay P.Singh, Ercan Kahya, Heikki Ruskeepaa, Mandeep Kaur Saggi, Bellie Sivakumar , Sungwon Kim, Farzin Salmasi, Mahsa Hasanpour Kashani, Saeed Samadianfard, Mahmood Shahabi, Rasoul Jani ‘Continuous monitoring of suspended sediment concentrations using image analytics and deriving inherent correlations by machine learning’, Scientific Reports.