2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT) | 978-1-6654-7224-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/CCiCT56684.2022.00054

Analyzing the Potability of Water using Machine Learning Algorithm

Aakarsh Arora
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
aakarshverma330@gmail.com

Vinayak Singh
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
vinayaksooryavanshi@gmail.com

Mahendra Kumar Gourisaria
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
mkgourisaria2010@gmail.com

Ajay Kumar Jena
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar 751024, Odisha
ajay.bbs.in@gmail.com

***Abstract*—Human life cannot exist without water, as most individuals are taught at a young age. According to the World Health Organization (WHO), clean water is not readily available for one out of every six people on the planet, which is around 1 billion people. The potability of water is affected by many factors like pH, Hardness, and various other chemicals present in the water. Contamination in drinking water can cause severe diseases like Hepatitis A, Cholera, Dysentery, Typhoid, and Polio. The paper primarily focuses on the binary classification of water potability using various standard machine learning algorithms like K-Nearest Neighbors (KNN), XG-Boost, Decision Tree, Support Vector Machine (SVM), Logistic Regression, Gaussian Naïve Bayes, Random Forest, LightGBM, Bernoulli Naïve Bayes and Gradient Boost on selected features from the dataset by using correlation matrix and in terms of performance, Decision Tree outperformed all other machine learning algorithms by achieving an accuracy of 0.9358, F1 score of 0.9374 and AUC-ROC score of 0.9220 for classification of water potability.**

Keywords—Machine Learning, contamination, WHO, Random Forest

I. Introduction

Water accounts for 71% of the earth's volume, out of which, only 0.3% of the water is available to us. From this, a high percentage is unfitted for drinking. Whether the water available to us is potable or not is a serious concern as it puts our health in jeopardy. Contaminated water makes us prone to severe infections and could be even lethal to us. Of the total population of the world, around 13% of them lack access to clean drinking water, thus, making them prone to severe diseases like Typhoid, Cholera, Diarrhoea, etc. Contamination caused by infectious micro-organisms is the largest cause of waterborne disease worldwide. Besides the mortality rate of half a million which is caused by Diarrhoea alone, it was also found that the pollutants present in drinking water become major factors for cancers like Esophageal Cancer (EC) in digestive systems [1]. The 5-year survival rate for EC is estimated to range from 15% to 25% worldwide [2]. Detecting contaminants and making improvements in drinking-water supply, sanitation, hygiene, and water resource management may avoid nearly 10% of the entire global rate of disease.

UV-absorbance spectrophotometry is one of the technologies for detecting contaminated substances in water, in which the absorbed spectra are compared to a collection of normal and unclean water footprints. However, it is considered a difficult and tedious process due to so many variables affecting this comparison like outside light, electronic noise, or the presence of other contaminants [3]. Another way could be of using Deterministic or Statistical models. The Deterministic model accounts for the physical and chemical processes of data collected and then we simplify differential equations to find an appropriate solution for the model. In the Statistical model, we meticulously select the techniques to analyze the model and validate suppositions and data. Deterministic models solution may entail suppositions and a lot of practical experience is required before reaching an outcome. The Statistical models, the majority of them are quite complex and some factors of water quality can pose challenges to the statistical model, like change-point detection, seasonality, the existence of outliers, etc.

Another approach for measuring water quality is the Water Quality Index (WQI), a non-dimensional index calculated from specified water quality indicators. The variables used to calculate contaminant detection include Total Suspended Solids (TSS), Biological Oxygen Demand (BOD), pH, Chemical Oxygen Demand (COD), Temperature, Ammoniacal Nitrogen (AN) and Dissolved Oxygen (DO) [4]. However, calculating WQI is difficult since sub-indices are obtained from WQI equations. Even though there are other approaches to calculate WQI like the methods of Florida Stream Water Quality Index (FWQI), Interim National Water Quality Standards for Malaysia (INWQS), the United States National Sanitation Foundation Water Quality Index (NSFWQI), or Interim National Water Quality Standards for Malaysia (INWQS), it has some disadvantages like it needs a significant amount of time, the method is too complicated and the calculations are too long. Recently, Artificial Intelligence (AI) has been recognized as a viable substitute strategy to model complex nonlinear systems.

978-1-6654-7224-1/22/$31.00 ©2022 IEEE
DOI 10.1109/CCiCT56684.2022.00054

Nowadays, with the growth of technology, many fields are relying on AI for problem-solving. Machine learning and deep learning, which are sub-branch of AI, are playing an important role in medical and other fields like malaria detection [5], Text recognition [6], COVID-19 detection [7], Biomedical Data analysis [8], Retinal Disease detection [9], Prediction of Liver Disease [10] and also in agriculture fields like Maize leaf disease detection [11], etc.

In this paper, we have executed various standard machine learning algorithms like K-Nearest Neighbours (KNN), Logistic Regression, XG-Boost, Decision Tree, Support Vector Machine (SVM), Gaussian Naïve Bayes, Random Forest, Gradient Boost, Bernoulli Naïve Bayes, and LightGBM on the binary classification problem based on water potability whether the water is drinkable or not and for selecting the best classifier algorithm performance metric parameter such as Accuracy, AUC, Recall, Precision, Specificity, F1-score, and MCC were considered. The remainder of the paper is divided into the following sections: II. Related Works, discusses the work of other research literature on the same topic. III. Data Preparation shows how we reduced the data and used the targeted features to reach the highest level of accuracy. IV. Technology Used lists all of the machine learning methods that were used for binary classification of water quality. The accuracy and other metrics for several machine learning models are briefly discussed in the Results and Evaluation section. VI. Discussion provides a comparative study and justification for the vest model. VII. Conclusion and Future Work section consists of final results along with several alternative machine learning techniques that can be used in future work.

II. Related Work

Artificial Intelligence, specifically in the fields of Machine Learning and Deep Learning, is attracting more and more scientists and researchers to find the solutions to modern-day problems. They can build high-efficiency models in various fields on the problems like Spermatozoan Assessment [12] in which the major factor to diagnose infertility was Sperm Morphology Analysis (SMA), or Chronic Kidney Failure [13] using Random Forest classifier.

Many researchers are contributing to the domain of earth sciences, so water potability is one of the major issues which is getting a major contribution from many other researchers. Chang et al. (2001) [14] proposed a Counterpropagation Fuzzy Neural Network (CFNN) that can emphasize forecasting streamflow and they achieved a very less Mean Absolute Error (MAE) of 3.71 (average of 1990 – 1996).

In another study by Chang et al. (2005) [15], a Genetic Algorithm (GA) and Fuzzy Rule Base (FRB) methods were used to extract knowledge from the dataset, and then Adaptive Network Fuzzy Interference System (ANFIS) was applied to predict optimal operations on the water reservoir. ANFIS was implemented for building a prediction model for water reservoir management and minimizing the damage [16]. The ANFIS model, which relied on the human decision as an input variable, proved to be a better performer with correlation coefficients larger than 0.99.

Zhang et al. (2014) [17] took the approach of Multivariate Empirical Mode Detection (MEMD) using Mahalanobis Distance. MEMD is an algorithm to analyze non-stationary and nonlinear signals, for single time-frequency water quality detection. This method outperformed the classical Empirical Mode Detection (EMD), where only one variation in input could be dealt with by the EMD and certain signals could be missed. Twala et al. (2019) [18] conducted a survey and on comparing the works of other researchers on such anomaly detection, proposed Hybrid Deep Learning and Extreme Machine Learning methods to be performed on Water Quality Anomaly Detection (WQAD) data, which could give high and reliable results.

Muharemi et al. (2019) [19] used the Support Vector Machine (SVM) model for water quality assessment, and it outperformed other models with an F1- score of 0.9891. However, Logistic Regression was also used by Muharemi et al. (2018) [20] earlier, to determine the best model with an F1-score of 0.58412 on water quality anomaly detection-, where they incorporated interaction terms, based on domain knowledge.

Oliker et al. (2015) [21] developed a two-step classification model focused on Minimum Volume Ellipsoid (MVE) classifier using unsupervised learning for outlier identification and sequence exploration, i.e. multivariate and multi-dimension analysis, to detect contamination events in water distribution systems. The evaluation of the models was based on the Accuracy and Detection Ratio, where the maximum accuracy scored was 0.95 and the maximum detection ratio of 0.91 for events influencing a single parameter. The unsupervised technique comes at a cost since it inhibits the autonomic calibration of the model. Because there aren't any well-known examples of events, the model is unable to do verification testing and parameter adjustment. As a result, trial and error were used to establish all of the model parameters.

A study by Hou et al. (2013) [22] executed the Dempster–Shafer (DS) evidence theory to develop an algorithm for identifying contamination occurrences. Water contamination events with simulated strengths greater than 1.2 were detected using the DS fusion approach. Riyantoko et al. (2021) [23] implemented Q-Lattices methods for the classification of water quality and achieved an accuracy of 68%.

Lu et al. (2020) [24] proposed a hybrid decision-tree based model, comprising of XG-Boost and Random Forest models, introducing a new technique - Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). The model evaluation was based on Mean Absolute Percentage Error (MAPE), in which both the models performed very well with an average MAPE of 3.90% and 3.71% respectively.

Although all the recent works for detecting water potability were adequate, they lacked the broad set of metrics needed to predict with all of the ML algorithms. Most of them had completely different datasets, missing the important factors like pH and Hardness. Few of them were theoretical models which were fixed by results from practical experimentation. Others had so many input variables, making the model time-consuming, costly and complex. Water is one of the most crucial parts of our lives. So a high accuracy model is required to binary classify contamination in water. As mentioned before, the field of

Artificial Intelligence has now become a very important part of our lives. People's interests in this field are growing every day, as it helps us to bypass the primitive and conventional methods to find the solutions. Robust models have been made like Tuberculosis Detection [25], Corn Leaves Disease detection [26], Arrhythmia Detection [27], Breast cancer detection [28], and Skin cancer prediction [29].

TABLE I. TABULAR REPRESENTATION OF RELATED WORK

Ref. No	Authors Name and Year	Advantages	Disadvantages
[16]	Chang *et al.* (2005)	Combined the methods of GA, FRB, and ANFIS for the prediction of water anomaly. Achieved a correlation coefficient larger than 0.99.	The use of ANFIS can cause a problem with large datasets or inputs with higher dimensions. Lack of dataset preprocessing
[17]	Zhang *et al.* (2014)	The integrated approach of MEMD and Mahalanobis distance. Outperformed all classic EMD models. The method was found suitable for multi-dimensional indicator fusions.	The dataset used relied only on three parameters. Dataset was very small Models were not well connected with the dataset. Cannot be considered for the real-world scenario as water potability is dependent on many other factors.
[21]	Oliker *et al.* (2015)	Proposed a two-step classification model Implemented Minimum Volume Ellipsoid (MVE) model and achieved the accuracy of 0.95 and detection ratio of 0.91.	The model was not performing well during validation testing Dataset was small Hyperparameters of models were based on the hit and trial method.
[22]	Hou *et al.* (2013)	Performed Dempster–Shafer (DS) fusion evidence theory to develop and achieved stimulated strength greater than 1.2.	The dataset used was small The input parameters were very less.
[23]	Riyantoko *et al.* (2021)	Used FEYN and Q-lattice methods for classification of water quality.	The accuracy achieved is not high.

III. DATA PREPERATION

The Data Preparation here will give an overview of where the dataset was collected and the proceeding techniques and methodologies applied to make the data fit for learning. Dataset features had missing values and no minor information that could impact the results. So, for better performance and analysis, we have tuned the dataset up to a point to yield better accuracy and outcomes.

A. Dataset Used

The dataset was posted by A. Kadiwal (2021) [30] which was collected via the UCI machine learning repository. Dataset is having records of 3276 water samples. The dataset contains a total of 10 features, out of which 9 features were Chloramines, Trihalomethanes, Solids, pH, Sulphate, Conductivity, Organic carbon, Hardness, and Turbidity and the last feature was the target variable called Potability.

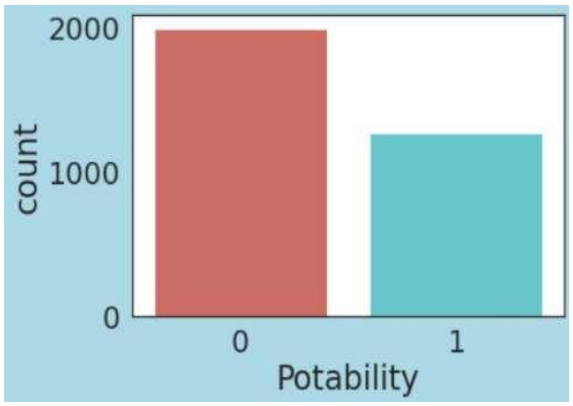

Fig. 1. Target Variable Evaluation

B. Data Exploration

Data exploration is one of the most crucial components of constructing a decent machine learning model. This section is critical for detecting missing values and mistakes, analyzing and extracting the dataset and attributes, and determining the most connected characteristics so that our algorithms can establish a solid relationship with the dataset.

Before employing the machine learning method, the dataset must be processed for excellent results. The dataset in our situation contained missing values and a large difference in the data range.

C. Missing Values, Correlation Matrix

After exploration of the dataset, it was found that there were 3 features in total containing missing values. Features like “pH”, “Sulphate” and “Trihalomethanes” had missing values count as 491, 781, and 162 respectively. The missing values were filled by the mean of all the data present in the respective features.

There were a few outliers that were present in the dataset. The correlation matrix was used to aid in the subsequent construction of algorithms. A correlation matrix is a table that displays the correlation coefficients for several variables. It's a very useful tool for encapsulating enormous datasets and seeing patterns in them.

D. Splitting Dataset, Hardware and Software used

The dataset was split into 2 parts, with 80 percent of the data utilized for training and the remaining 20 percent for testing, with the random state set to 42. Python 3.7 and TensorFlow were used to implement and evaluate all of the machine learning algorithms and implementations on a Google Colaboratory notebook. The workstation is equipped with an Intel i7 9th generation processor and 8GB of RAM.

Fig. 2. Correlation Matrix

IV. Technology Used

The dataset was trained and tested using widely used machine learning algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (NB), Support Vector Machine (SVM), Extreme Boost (XG-BOOST), Decision Tree (DT), Random Forest (RF), Gradient Boost (GB), Light Gradient Boosting Machine (LightGBM), and Bernoulli Naïve Bayes (NB). Machine learning is playing a major role in solving various day to days applications such as the classification of Diabetes Mellitus disease [31], phishing detection tools [32], and surface crack detection [33].

These algorithms are useful in optimizing the criteria of performance using the experience. These machine learning algorithms are used for developing predictive data models. Table II shows the advantages and disadvantages related to all the machine learning algorithms used.

TABLE II. Classifiers with Advantages and Disadvantages

Classifier	Advantage	Disadvantage
Logistic Regression (LR)	The rate of classification is fast. Easily implementable with efficient training.	Cannot be used for complex datasets as it is not a powerful algorithm. Used only for categorical classification.
Decision Tree (DT)	Requires less effort for data preparation. Do not get affected by non-existent values.	A small change in the dataset leads to large instability causing overfitting. Takes more complexity and time.
Support Vector Machine (SVM) (Linear)	Predicts with a fast rate. It is relatively memory efficient	Does not perform well when the dataset has more noise. Scaling is a necessity.
Random Forest (RF)	Reduces overfitting and improves accuracy. Handles non-linear parameters.	Creates more trees, and is more complex. Requires long training period.
Light Gradient Boosting Machine (LGBM)	Faster training speed. Lower memory usage.	Narrow user base. Easily overfit small dataset.
Gradient Boost (GB)	Highly flexible. Provides predictive accuracy.	Can cause overfitting. Time and memory are exhaustive.
K-Nearest Neighbors (KNN)	New data can be added easily. Easy to implement.	Does not work well with a large dataset. Sensitive to noisy, missing data and outliers.
Gaussian Naïve Bayes (Gaussian NB)	Works well with a large dataset. Less training time.	Does not work well with a small dataset. Have low performance.
Bernoulli Naïve Bayes (BNB)	More accuracy in small datasets. Ability to make real-time predictions.	Only works well with binary form features.

V. Implementation and Results

After the exploration of the data, we implemented all the machine learning techniques mentioned above. Before training the model, we have scaled the data, setting the random state to 42, and later splitting it into an 8:2 ratio of training and test sets. While implementing the machine learning algorithms on the training set, the random state was set to 42 in all cases. In Logistic Regression, K-fold with 3 splits was also applied. Grid Search was also applied for hyperparameter tuning in all the models. After successfully training the models, it is crucial to find which model is more suitable. This was done by comparing the parameters calculated from the confusion matrix of all the models. The parameters were "Accuracy", "AUC", "Recall", "Precision", "Specificity", "F1-score", "MCC", "TN", "FP", "FN" and "TP". The abbreviations used in Table IV have been defined in Table III.

TABLE III. Abbreviations used in Metrics

Abbreviation	Meaning
MCC	Mathews correlation coefficient
AUC	Area Under Curve
CM	Confusion Matrix

Table IV shows that the Decision Tree model performed exceedingly well and had a solid and steady relation with the dataset that was used. It had an accuracy of 0.9358, followed by RF with 0.8246, and RF being followed by LightGBM with 0.8033. Decision Tree performed well in every field when compared to other models. From the observation of Table-IV, we can see that XGBoost almost took the lead in specificity and LightGBM in Recall than Decision Tree. SVM and Logistic Regression are the least scorer and have the same scores in every parameter. This happened because the data got overfitted in the models. They both lacked in building a good relationship with the dataset, which could be because of the enormous dataset having a dimension of 3276x10Units. True Negative (TN) are values of true negative facts through a negative description. False-negative (FN) are values of negative data which is characterized as positive. A false positive (FP) is the value of positive data that has been described as negative. The equations (1), (2), (3), and (4) give us the formularies for the above-used metrics in assessing the results of different models as follows.

$$Specificity = \frac{TN}{FP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$F1\,Score = \frac{2TP}{2TP + FP + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]} \quad (4)$$

TABLE IV. Performance Metrics of All Machine Learning Algorithms

Model	Accuracy	Precision	AUC	Recall	Specificity	F1 Score	MCC	CM	
XGBoost	0.8018	0.8475	0.7544	0.5696	0.9393	0.6813	0.5681	387	25
								105	139
LightGBM	0.8033	0.7751	0.7749	0.6639	0.8859	0.715232	0.5703	365	47
								82	162
Random Forest	0.8246	0.8568	0.7860	0.6352	0.9368	0.7294	0.6186	386	26
								89	155
Gradient Boost	0.8003	0.8304	0.7557	0.5819	0.9296	0.6843	0.5632	383	29
								102	142
Decision Tree	**0.9358**	**0.9493**	**0.9220**	**0.9258**	**0.9466**	**0.9374**	**0.8719**	**487**	**26**
								39	**461**
Bernoulli NB	0.6478	0.5343	0.6001	0.4139	0.7864	0.4665	0.2138	324	88
								143	101
Gaussian NB	0.6310	0.5096	0.5467	0.2172	0.8762	0.3045	0.1236	361	51
								191	53
KNN	0.6295	0.5024	0.5872	0.4221	0.7524	0.4587	0.1820	310	102
								141	103
Logistic Regression	0.6280	0.0000	0.5000	0.0000	1.0000	0.0000	0.0000	412	0
								244	0
SVM	0.6280	0.0000	0.5000	0.0000	1.0000	0.0000	0.0000	412	0
								244	0

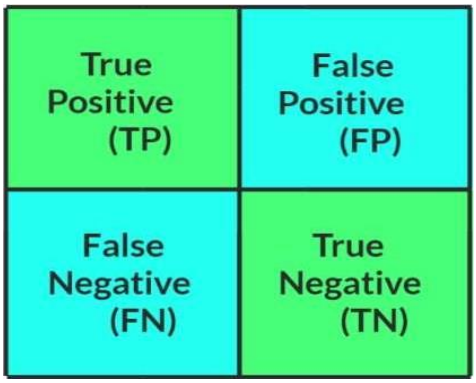

Fig. 3. Confusion Matrix

From the experimentation, we also noticed that Linear regression and Support Vector Machine overfits the dataset. They did not have good relationships with the dataset and they achieved the minimum accuracy of 0.6280. Bernoulli Naïve Bayes, KNN, and Gaussian Naïve Bayes also not performed well and had accuracy below 70% which is not good at all. The decision tree achieved the highest accuracy of 0.9358 and performed outstandingly in all the performance metrics and was the best classifier for water potability. It also had an F1 score of 0.9374 which is very high and the decision tree was the best and most stable algorithm for the classification of water potability. In terms of results, all the metrics are included in Table IV. Fig. 3 is a visualization of the confusion matrix which plays an important role in performance evaluation. Random forest was also performing well with the dataset but due to its complex nature, it had an accuracy of 0.8246 which was approximately 0.11 less than the decision tree.

VI. Discussion

From the Implementation and Results section, we can conclude that the Decision Tree classifier performed well in terms of all metrics analysis. Every algorithm was implemented by setting the random state to 42. This ensures that the results are reproducible and do not change every time we retrain the model. So, the maximum predictions which were made correctly implying no contaminated water were predicted as potable, will not change on recompile. Fig. 4 shows a graphical analysis of the outcomes of the model.

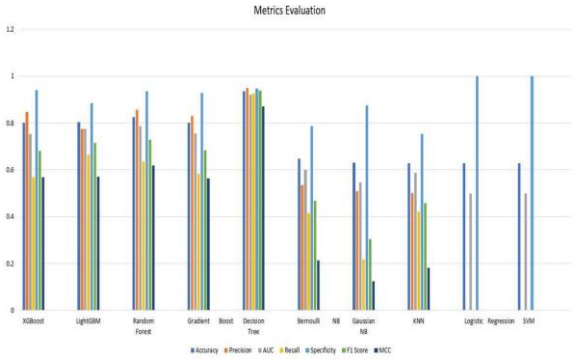

Fig. 4. Graphical representation of models

From the above representation, we can see that the Decision Tree model was able to process the input data and predict the values better than other models. It was selected as it performed well in all the metrics as compared to other models. All the models were able to learn the features except for the SVM and Logistic regression models. These models could not perform well as they were able to grasp the details and noise in the training dataset due to the complexity of the large dataset and were overfitted. Random Forest, LightGBM and XGBoost were also able to form a steady relationship with the input features. However, the RF model got too complex while making trees in the large dataset. In the case of K-NN, even after scaling the features, it remains a little sensitive toward noise in the dataset and could not perform well.

So, as compared to other models, Decision Tree neither overfits nor underfit the dataset. It did not get affected by the large dataset and the noise present in it. The classifier was able to form a more stable relationship between the dataset due to its simplicity and less effort for data preparation during pre-processing.

VII. Conclusion and Future Work

Contamination of water is a very serious issue as it not only affects our health but also puts other living organisms at risk. In this paper, we have used the power of machine learning to classify whether the given sample of water is potable or not. Numerous machine learning techniques like Decision Tree, Support Vector Machine, LightGBM, KNN, Random Forest, Bernoulli NB, XGBoost, Gradient Boost, and Logistic Regression were applied and on comparing, Random Forest performed superbly giving high accuracy and precision than any other models.

The decision tree gained the highest accuracy of 0.9358 according to our study and gained better results in almost all the metrics evaluation. From the Table-IV, we can see that Random Forest, LightGBM, and XGBoost were the close competitors of the Decision tree, followed by Gradient Boost.

The main purpose of this paper was to implement the ability of machine learning to classify the water sample based on the components present and chemical factors like “ph” and “turbidity” as potable or not and we compared the algorithms used and determined that Random Forest gave the best accuracy out of all other models.

For future work, we can implement other algorithms like AdaBoost or CatBoost, or ANN. Data-preprocessing and dimension reduction using techniques like PCA and LDA can also help in giving good results.

Acknowledgment

The author would like to express sincere gratitude for the support given to him by his co-authors for the completion of this research project especially Sir M. K. Gourisaria for their hard work, constant support, and guidance.

References

[1] C. Xu, D. Xing, J. Wang, and G. Xiao, "The lag effect of water pollution on the mortality rate for esophageal cancer in a rapidly industrialized region in China.," Environmental Science and Pollution Research, vol. 26(32), pp. 32852-32858, 2019.

[2] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, and D.M. Parkin, "Estimates of worldwide burden of cancer in 2008," GLOBOCAN 2008. Int J Cancer, vol. 127(12), pp. 2893–2917, 2010.

[3] T.A Arnon, S. Ezra, and B. Fishbain, "Water characterization and early contamination detection in highly varying stochastic background water, based on Machine Learning methodology for processing real-time UV-Spectrophotometry.," Water Research, vol. 155, pp. 333-342, 2019.

[4] M. Hameed, S.S Sharqi, Z.M. Yaseen, H.A. Afan, A. Hussain, and A. Elshafie, "Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia.," Neural Computing and Applications, vol. 28(1), pp. 893-905, 2017.

[5] M.K. Gourisaria, S. Das, R. Sharma, S.S. Rautaray, M. Pandey, "A Deep Learning Model for Malaria Disease Detection and Analysis using Deep Convolutional Neural Networks.," International Journal of Emerging Technologies, vol. 11(2), pp. 699-704, 2020.

[6] J. H. AlKhateeb, J. Ren, J. Jiang, and H. Al-Muhtaseb, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking," Pattern Recognition Letters, vol. 32(8), pp. 1081-1088, 2011

[7] G. Jee, H. GM, and M. K. Gourisaria, "Juxtaposing inference capabilities of deep neural models over posteroanterior chest radiographs facilitating COVID-19 detection," Journal of Interdisciplinary Mathematics, pp.1-27, 2021.

[8] H. Das, B. Naik, H. S. Naik, S. Jaiswal, P. Mahato, and M. Rout, "Biomedical data analysis using a neuro-fuzzy model with post-feature reduction.," Journal of King Saud University-Computer and Information Sciences. 2020.

[9] S. Sarah, V. Singh, M. K. Gourisaria, and P. K. Singh, "Retinal Disease Detection using CNN through Optical Coherence Tomography Images," In 2021 5th International Conference on Information Systems and Computer Networks (ISCON), pp. 1-7, Oct 2021.

[10] V. Singh, M. K. Gourisaria, and H. Das, "Performance Analysis of Machine Learning Algorithms for Prediction of Liver Disease," In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), pp. 1-7, Sep 2021, IEEE.

[11] K. P. Panigrahi, H. Das, A. K. Sahoo, and S. C. Moharana, "Maize leaf disease detection and classification using machine learning algorithms," In Progress in Computing, Analytics, and Networking, pp. 659-669, Springer, Singapore, 2021.

[12] S. Chandra, M. K. Gourisaria, G. M. Harshvardhan, D. Konar, X. Gao, T. Wang, and M. Xu, "Prolificacy Assessment of Spermatozoan via state-of-the-art Deep Learning Frameworks," 2020.

[13] R. Pramanik, S. Khare, M.K. Gourisaria, "Inferring the Occurrence of Chronic Kidney Failure: A Data Mining Solution." In: Gupta D., Khanna A., Kansal V., Fortino G., Hassanien A.E. (eds) Proceedings of Second Doctoral Symposium on Computational Intelligence. Advances in Intelligent Systems and Computing, vol. 1374, Springer, Singapore,2022.DOI https://doi.org/10.1007/978-981-16-3346-1_5.

[14] F.J. Chang, Y.C. Chen, 2001. "A counterpropagation fuzzy-neural network modeling approach to real-time streamflow prediction.," J. Hydrol. vol. 245, pp. 153–164, 2001. DOI https://doi.org/10.1016/S0022-1694(01)00350-X.

[15] Y.T Chang, L.C. Chang, and F.J. Chang, " Intelligent control for modeling of real – time reservoir operation, part II: artificial neural network with operating rule curves., " Hydrological Processes: An International Journal, vol. 19(7), pp. 1431-1444, 2005.

[16] F.J. Chang, Y.T. Chang, "Adaptive neuro-fuzzy inference system for prediction of water level in reservoir.," Adv. Water Resource, vol. 29, pp. 1–10, 2006. DOI https://doi.org/10.1016/j. advwatres.2005.04.015

[17] Z. Yang, Y. Liu, D. Hou, T. Feng, Y. Wei, J. Zhang, and G. Zhang, "Water quality event detection based on Multivariate empirical mode decomposition. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2663-2668, Oct 2014, IEEE.

[18] E. M. Dogo, N. I. Nwulu, B. Twala, and C. Aigbavboa, "A survey of machine learning methods applied to anomaly detection on drinking-water quality data," Urban Water Journal, 2019. DOI: 10.1080/1573062X.2019.1637002

[19] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set.," Journal of Information and Telecommunication, vol. 3(3), pp. 294-307, 2019.

[20] F. Muharemi, D. Logofătu, C. Andersson, and F. Leon, "Approaches to Building A Detection Model for Water Quality: A Case Study.", In Modern Approaches for Intelligent Information and Database Systems, edited by A. Sieminski, A. Kozierkiewicz, M. Nunez, and Q. Ha, 173–183, Vol. 769. Springer, Cham, Switzerland, 2018.

[21] N. Oliker, & A. Ostfeld, "Minimum volume ellipsoid classification model for contamination event detection in water distribution systems.," Environmental modelling & software, vol. 57, pp. 1-12, 2014.

[22] D. Hou, H. He, P. Huang, G. Zhang & H. Loaiciga, "Detection of water-quality contamination events based on multi-sensor fusion using an extended Dempster–Shafer method.," Measurement Science and Technology, vol. 24, pp. 1-18, 2013.

[23] P.A. Riyantoko, and I.G.S.M Diyasa, ""FQAM" Feyn-QLattice Automation Modelling: Python Module of Machine Learning for Data Classification in Water Potability.," In 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), pp. 135-141, Oct 2021, IEEE

[24] H. Lu, and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction.," Chemosphere, vol. 249, pp. 126169, 2020.

[25] V. Singh, M. K. Gourisaria, G. M. Harshvardhan, and V. Singh, "Mycobacterium Tuberculosis Detection Using CNN Ranking Approach," In Gandhi T.K., Konar D., Sen B., Sharma K. (eds) Advanced Computational Paradigms and Hybrid Intelligent Computing. Advances in Intelligent Systems and Computing, vol. 1373,Springer, Singapore, 2020. DOI: https://doi.org/10.1007/978-981-16-4369-9_56

[26] K.P. Panigrahi, A.K. Sahoo, and H. Das, "A cnn approach for corn leaves disease detection to support digital agricultural system.," In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI), vol. 48184, pp. 678-683, June 2020, IEEE.

[27] M.K Gourisaria, G.M. Harshvardhan, R. Agrawal, S.S. Patra, S.S. Rautaray, and M. Pandey, "Arrhythmia Detection Using Deep Belief Network Extracted Features From ECG Signals.," International Journal of E-Health and Medical Communications (IJEHMC), vol. 12(6), pp. 1-24, 2021, DOI http://doi.org/10.4018/IJEHMC.20211101.oa9.

[28] Yue, W., Wang, Z., Chen, H., Payne, A., Liu, X., "Machine learning with applications in breast cancer diagnosis and prognosis.," Designs, vol. 2(2), pp. 13, 2018.

[29] A. Sannigrahi, V. Singh, M. K. Gourisaria and R. Srivastava, "Diagnosis of Skin Cancer Using Feature Engineering Techniques," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 405-411, DOI: 10.1109/ICAC3N53548.2021.9725420.

[30] Kadiwal. A,. (2021).Water Quality, Version 3.Retrieved from https://www.kaggle.com/datasets/adityakadiwal/water-potability.

[31] M. K. Gourisaria, G. Jee, G. M. Harshvardhan, V. Singh, P. K. Singh, and T. C. Workneh, "Data science appositeness in diabetes mellitus diagnosis for healthcare systems of developing nations," IET Communications, 2022.

[32] W. D. Yu, S. Nargundkar, and N. Tiruthani, "PhishCatch—a phishing detection tool," in Proceedings of the 33rd Annual IEEE International Computer Software and Applications Conference (COMPSAC '09), vol. 2, pp. 451–456, Seattle, Wash, USA, July 2009.

[33] A. Chordia, S. Sarah, M. K. Gourisaria, R. Agrawal, and P. Adhikary (2021, September). "Surface Crack Detection Using Data Mining and Feature Engineering Techniques." In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 1-7). IEEE.