International Journal of Enhanced Research in Management & Computer Applications,
ISSN: 2319-7471, Volume 14 Issue 7, July 2025

To Study Water Quality Prediction for Different Applications

Sonali Abhijeet Padalkar[1], Dr Nilesh Vashant Ingale[2]

[1]Research Scholar, Dept of Computer Science & Engineering, Vikrant University
[2]Supervisor, Dept of Computer Science & Engineering, Vikrant University

ABSTRACT

Artificial Intelligence (AI) is increasingly being used to address the challenges faced in sustainable agriculture, creating a transformative impact. This revolutionary technology is helping farmers and agricultural businesses make smarter decisions, increase productivity, and reduce environmental impact, thus playing a pivotal role in sustainable farming practices. One of the most significant ways AI is contributing to sustainable agriculture is through precision farming. It involves using AI technologies like machine learning, robotics, and drones to gather detailed information about the condition of crops, soil, weather, and other essential factors.

This data is then analyzed to provide farmers with valuable insights, enabling them to make informed decisions about planting, fertilization, pest control, and harvesting. As a result, precision farming significantly reduces the use of resources such as water, fertilizer, and pesticides, enhancing the sustainability of agriculture. AI is also being used to develop predictive models for crop diseases and pests. These models use AI algorithms to analyze various data, including weather conditions, historical disease patterns, and crop health data, to predict the likelihood of crop diseases or pest infestations. Farmers can use these predictions to take preventive measures, reducing crop loss and the need for chemical treatments.

Keyword: Crops, Soil, Weather, Weather Conditions, Planting, Fertilization

INTRODUCTION

A well-balanced aquaponic system must be established by carefully taking into account a number of essential elements that cooperate to maximise the system's overall production and functionality. These crucial standards, which are the result of a thorough synthesis of study and professional knowledge, guarantee that the aquaponic system operates well while optimising the growth of both fish and plants. The following are the essential components that support an aquaponic system's effective design and operation:

- **Fish Type and Feed Rate:** Growth traits, temperature tolerance, and aquaponic environment adaptability all play a role in the selection of fish. In order to satisfy the fish's nutritional requirements without sacrificing the quality of the water, it is also essential to create a balanced feed rate.

- **Food Composition:** The fish's excretion is influenced by the feed's nutritional value, which in turn impacts the system's overall nutrient balance.

- **Feeding Frequency:** To satisfy the fish population's nutritional needs while avoiding overfeeding, which can result in excessive nutrient accumulation and possible disturbances to the system's equilibrium, feeding frequency must be balanced.

- **Hydroponic system design:** elements like planting bed placement, nutrient delivery systems, and water circulation patterns have a direct effect on how well plants absorb nutrients and, as a result, the system's overall balance.

- **Plant Types:** The symbiotic interaction between hydroponics and aquaculture is strengthened by choosing plants that complement the nutrient dynamics produced by fish faeces.

- **Plant Sowing Density:** In order to minimise overcrowding and encourage ideal growth, each plant must have enough space to absorb nutrients.

- **Chemical Composition of the Recirculated Water:** To establish an environment that is favourable for the health of fish, plants, and vital microorganisms, it is crucial to regularly check and change water parameters, such as pH, ammonia levels, and nutrient concentrations.

- **Environmental Parameters:** The aquaponic system's overall balance is significantly influenced by environmental parameters, which include air temperature, relative humidity, CO2 levels, and light intensity. Stable circumstances for the growth and development of both aquatic and plant components are ensured by monitoring and controlling these parameters.

A productive, sustainable aquaponic environment that supports both aquatic and plant life may be created by addressing each of these factors, which will eventually result in increased efficiency and better food production outcomes.

Innovative technologies that can further improve the sustainability and efficiency of food production have been developed as a result of recent developments in aquaponics. New methods for integrating and enhancing conventional aquaponic systems include algaeponics, aeroponics, aeroaquaponics, biofloc technology, dige-ponics, haloponics, maraponics, and vermiponics. These new techniques offer promising ways to boost output and maximise resource utilisation in aquaponic systems.

REVIEW OF LITERATURE

Goddek et al. (2015) offer a thorough examination of the difficulties related to sustainable aquaponics, including a variety of topics that go beyond environmental concerns to include commercial ones as well. Their research sheds insight on the challenges aquaponic systems confront, including scalability, market acceptability, and economic viability—all of which are essential to the long-term sustainability of these integrated farming methods. The authors provide a thorough review of the obstacles that must be removed in order to successfully promote sustainable aquaponics by looking at both the ecological and commercial dimensions.

A comprehensive analysis of aquaponic systems is provided by Palm et al. (2018), covering a number of important topics such design considerations, operational sizes, and industry nomenclature. Their research explores the many aquaponic system configurations, looking at how various designs might affect production and efficiency. They also investigate the range of scales at which aquaponics can be used, from modest home-based systems to expansive commercial enterprises. In order to give readers a thorough grasp of the terminology and ideas used in this cutting-edge farming method, the authors also explain the nomenclature related to aquaponic activities. All things considered, academics and practitioners interested in the diverse aspects of aquaponics will find this review to be a useful resource.

OBJECTIVES OF THE STUDY

1. To Study the significance of various aquaponic parameters and their optimal ranges, presenting them in a detailed tabular format. This serves as a useful reference tool for aquaponic farmers, helping them manage and optimize their systems effectively.

2. To Study the strategies and methodologies for effectively monitoring and controlling the key parameters of aquaponic systems. This section explores different approaches to ensure that all conditions remain optimal for plant and fish growth, thereby enhancing overall system efficiency.

RESULT AND DATA INTERPRETATION

The suggested approach uses data cleansing to deal with outliers and missing numbers. Because the percentage of missing data is slightly higher, the suggested method retains the samples within the dataset by using the substitution approach. The missing value is replaced using the average of the three closest values that were present before and adjacent to the missing data. Data analysis determines how many outliers there are in a dataset. The suggested strategy removes outliers from the dataset since the percentage of outliers in the dataset is approximately 0.186 percent.

Finding a subset of the top-performing features for the greatest model fit is done using the Forward Feature Selection (FFS) technique. The attributes of a dataset with n attributes are chosen based on earlier research. The suggested approach uses the FFS method to assess the model's performance after training each of the N features. The feature (or group of characteristics) that performs the best is then chosen.

Until the suggested system has the required number of characteristics, this procedure is repeated. The FFS methodology is therefore a wrapper method for choosing the optimal subset of features. To obtain the optimal subset of features, the suggested method makes use of the FFS technique with numerous trials. Features including pH, TDS, DO, BOD, ammonia-N, conductivity, sodium, boron, nitrate-N, and chloride are provided by this procedure.

Improved G-SMOTE for handling the Imbalanced Dataset

Ten features (dimensions) having numerical values are chosen by the suggested system using a water quality dataset of 12456 samples, as shown in Table 3.4. The suggested approach separates 12456 samples into six classes using the data cleaning procedure. Tables 3.3 and 3.4 indicate that the first five of these six classes are from Class A to Class E, while the sixth class is determined to be a not appropriate class (NS).

The dataset is then split into a 70:20:10 ratio, meaning that 70% of the samples are used for model training, 20% are used for model validation, and 10% are used for model testing. Analysis reveals that the distribution of classes in the training dataset is not uniform. The proposed system offers multiple methods for handling unbalanced dataset issues by employing a range of performance metrics, including precision, recall, F-score, and Cohen's kappa, as accuracy is not a desirable metric for imbalanced datasets. The issue of an unbalanced dataset is successfully resolved by sampling the dataset using oversampling and undersampling techniques.

Samples (rows) from classes with more copies of samples (rows) are eliminated through under-sampling. Oversampling randomly adds samples (rows) with substitutions from the minority class to the training dataset. Previous SMOTE techniques employ samples from various classes instead of examining neighbouring samples to create new samples for minority classes, which increases class overlap and outliers (noise). Additionally, SMOTE creates multiple samples for minority classes and is unable to handle high-dimensional data. As a result, after applying SMOTE, the dataset may contain a large number of duplicate samples, which could impact the accuracy of the model.

These are a major drawback of the SMOTE approach. It is a challenging optimisation task to choose samples, sampling features, sampling type, sampling rate, and sampling distribution technique from a minority class. The training dataset for the suggested system is unbalanced due to the non-uniform distribution of the classes for multi-class classification. The Artificial Neural Network model can be strengthened with further data. The model's capacity to generalise on the new input data points is enhanced by the G-SMOTE data augmentation method, which creates variances in the data points. Thus, the suggested enhanced G-SMOTE approach is only used to the training examples (70 percent samples from the original dataset) in order to produce a balanced dataset.

Mathematical modelling is necessary to determine the optimal solution for these problems. A Genetic Algorithm (GA) is a stochastic optimisation technique that uses randomised search techniques to solve a challenging multidimensional nonlinear problem. A Genetic Algorithm (GA) is frequently used in operations and engineering optimisation challenges. In order to address the imbalance issue, the suggested approach introduces the enhanced G-SMOTE technique, which hybridises the smote method with a GA. The enhanced G-SMOTE method's algorithm is explained below.

- **G-SMOTE**

Step 1: Begin
Read the dataset in step two.

Step 3: Examine the dataset's distribution

Step 4: Initialisation: Use samples that include both majority and minority populations to start the population. Additionally, set the sample range by initialising the decision variable.

Step 5: SMOTE: This step uses the support vector machine to add artificially created samples to the minority class.

Step 6: Get the updated dataset.

Step 7: Selection: The fitness function is created to establish the cutoff point for minority class instances' fitness scores. For every minority class, this function identifies the unique and duplicate samples.

Step 8: Check for Converge: The dataset is considered balanced and the algorithm ends if the likelihood of duplicate samples in each minority class is less than the fitness score threshold; if not, eliminate the duplicate samples.

Step 9: Crossover: The suggested method selects k-samples from each minority class instance using the Euclidean distance, then calculates the average of the k-samples to provide new samples to be added.

Step 10: Mutation: To modify the sample produced by the crossover phase and update the samples of minority classes, it randomly chooses a mutation probability between 0.98 and 0.99.

Step 11: To verify for converge conditions, proceed to step 8.

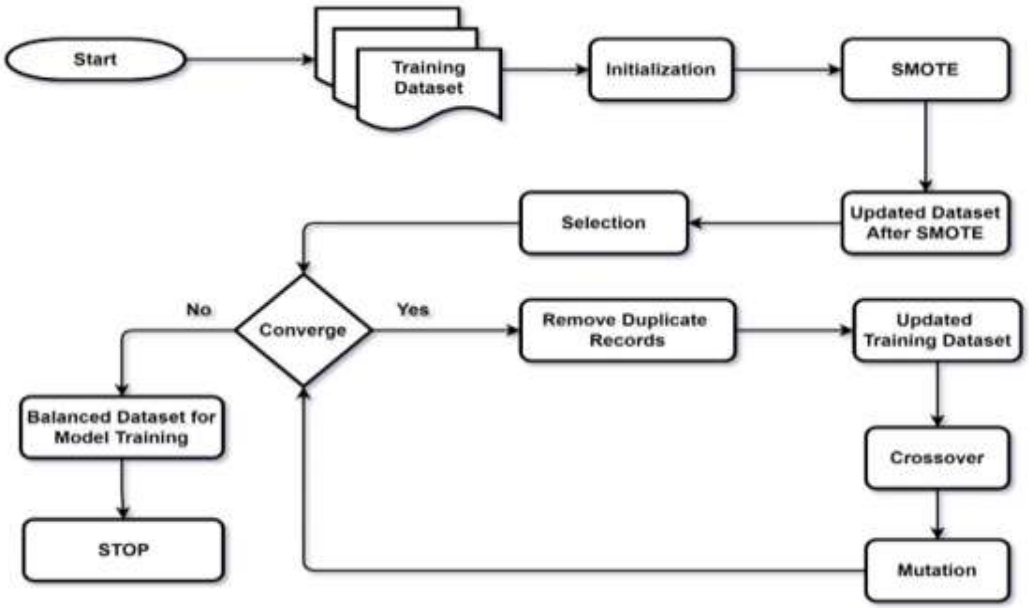


Figure 4.1: Working of G-SMOTE Algorithm

To get the best accuracy rate for the classification, the suggested method links many instances (rows) of the original training dataset with the sampling rate. The mathematical modelling of the suggested G-SMOTE approach is displayed below.

The suggested method uses the genetic algorithm (GA) to obtain an ideal sampling rate for several occurrences, as indicated in equation 3.18.

Maximise: Z = G(H); minX << Xi << maxX **(3.18)**
H = (X1, X2, X3,..., XP); i = 1, 2, 3,..., P.
The accuracy rate for classifying minority classes and the total dataset is represented by the objective function G(H). The sampling rate is denoted by H. P stands for the total number of minority class instances. Xi represents the minority class instance Hi's sampling rate. MinX is the lower bound of sampling rate Xi, and maxX is the upper bound. To improve sample rates for oversampling, the G-SMOTE technique is employed. For multiclass classification, it creates a fresh (balanced) dataset with a uniform distribution of samples in each minority class. The enhanced G-SMOTE's mathematical modelling is displayed below.

- **Mathematical Modelling of G-SMOTE**

I. Initialization

An N-sized population for the GA is produced using this procedure. Xj represents the sampling rate for minority class samples Hj. From the standpoint of a genetic algorithm, the technique uses a member of a population to specify combinations of the sampling rates for each sample, as shown in equation (3.19).

$$Hi = (X'_1, X'_2, X'_3, \ldots\ldots, X'_p) \quad \text{where } i=1,2,3\ldots\ldots N. \qquad (3.19)$$

Equation 3.19 illustrates that P is the length of a chromosome, which is equivalent to the number of instances of a minority class. To initialise an individual, each chromosomal node is assigned to a random number (whole number) within the top and lower boundaries of the sampling rate, as indicated by equation (3.20).

$$Xi = round(\, minX + (maxX\,, minX) \times random(0,1)) \qquad (3.20)$$

with j = 1,2,3,4, N and i = 1,2,3,4.... P.
The rounding (round-off) is done using the round() method. A distinct person is represented by each Hi column.

II. Selection and fitness function

The population is sorted in descending order based on the fitness function value, which is calculated for each member of the population. The letter Ps stands for the selected probability. After then, the process creates two copies. The first is N X Ps, a sorted population of people. In contrast, the second creates a new population by removing the final N X Ps people from the sorted populations while retaining the individuals in the center.

III. Crossover Operation

As shown in equations (3.21 and 3.22), respectively, two individuals are chosen at random from the population each time and are denoted by Hi and Hj.

$$Hi = (X^i_1, X^i_2, X^i_k \ldots X^i_{k+1}, X^i_{k+2} \ldots . X^i_p) \text{ where } i=1.2.3 \ldots N \quad (3.21)$$

$$Hj = (X^j_1, X^j_2, X^j_k \ldots X^j_{k+1}, X^j_{k+2} \ldots . X^j_p) \text{ where } i=1.2 \ldots N. \quad (3.22)$$

Equations (3.23) and (3.24) illustrate how the nodes for the early crossing are chosen at random and represented by the letter K. Hi and Hj crossings take place immediately after K nodes, and new individuals are derived as soon as these crossings take place.

$$Hi = (X^i_1, X^i_2, X^i_k \ldots \ldots X^j_{k+1}, X^j_{k+2} \ldots \ldots X^j_p) \quad (3.23)$$

$$Hj = (X^j_1, X^j_2, X^j_k \ldots \ldots X^i_{k+1}, X^i_{k+2} \ldots \ldots X^i_p) \quad (3.24)$$

IV. Mutation Operation

For every member of the population, a random number between 0 and 1 is created. A random number between 0 and 1 is used to calculate the fitness probability; non-uniform mutation happens when each individual's fitness probability is less than the mutation probability. According to equation (3.25), no mutation happens whenever the fitness probability of all individuals is greater than the mutation probability.

$$X^i_k = \begin{cases} X^i_k + (MaxX - X^i_k \times (1 - rand(0,1)^{(1-\frac{ge}{Ge})^3}), rad(2) = 0 \\ X^i_k - (X^i_k - minX) \times (1 - rand(0,1)^{(1-\frac{ge}{Ge})^3}), rad(2) = 1 \end{cases} \quad (3.25)$$

Ge stands for the current generation and Ge for the maximum generation in equation (3.25).

V. Termination Check

The algorithm satisfies the terminating condition and produces the best individual if the Current Generation (ge) exceeds the maximum generation; if not, continue step 2. SMOTE is used to produce the new dataset if the optimal sampling rate is reached. Larger and less precise decision boundaries are thus produced by G-SMOTE, which enhances classifier performance and generalisation capacities. As a result, it addresses the problem of overgeneralisation and the potential to enhance noisy areas.

CONCLUSION

The large dimensionality and imbalance of the dataset, as well as the issues of unseen data to the machine learning model, are the primary concerns of the suggested solution. To solve the problems with large dimensional and unbalanced datasets, innovative data augmentation techniques like M-SMOTE, G-SMOTE, and JBO-SMOTE are developed and put into practice. These techniques manage high-dimensional datasets, produce a uniform distribution of the dataset's samples, prevent class overlapping, eliminate duplicate samples, and produce a dataset free of outliers and noise. As a result, with a high dimensional dataset, these data augmentation techniques produce consistent samples with distinct values for every minority class. The balanced dataset is created after using the innovative data augmentation techniques, and it is then fed into machine learning or deep learning techniques to enhance performance. Then, by drastically cutting the training time, the suggested solution has optimised the deep learning algorithm, enhancing accuracy, improving generalisation, and lowering the likelihood of overfitting the data. Lastly, to solve the problem of unseen data in the machine learning models, a novel adaptive incremental learning framework is created and put into practice.

Six cutting-edge classifiers are used in the initial implementation of the voting-based water quality prediction system for aquaponic farming. The voting principle is used to create an ideal water quality prediction model. Based on the voting concept, the suggested system finds that XGBoost and Random Forest are the best classifiers using the suggested M-SMOTE. With an accuracy of 97.35%, the Random Forest classifier produced an optimised water quality prediction model for aquaponic farming that is deployed. The Random Forest classifier is used to test the suggested data augmentation methods, such as M-SMOTE, G-SMOTE, and JBO-SMOTE, independently for aquaponic farming water quality prediction. When it comes to Random Forest classifier accuracy, M-SMOTE and JBO-SMOTE have outperformed G-SMOTE.

REFERENCES

1. M. K. Srivastava, S. Gaur, and A. Ohri (2024), "Analysing the effectiveness of mcdm and integrated weighting approaches in groundwater quality index develop-ment," *Water Conservation Science and Engineering*, vol. 9, no. 2, p. 35,.
2. H. Azarnivand (2024), "Assessment of groundwater resources potential using improved water quality index (impwqi) and entropy-weighted topsis model," *Sustainable Water Resources Management*, vol. 10, no. 1, p. 7,.
3. M. Saqlain, H. Garg, P. Kumam, and W. Kumam (2023), "Uncertainty and decision-making with multi-polar interval-valued neutrosophic hypersoft set: A distance, similarity measure and machine learning approach," *Alexandria Engineering Journal*, vol. 84, pp. 323–332,.
4. V. Acharya, S. K. Sharma, and S. K. Gupta (2018), "Analyzing the factors in in-dustrial automation using analytic hierarchy process," *Computers & Electrical Engineering*, vol. 71, pp. 877–886,.
5. M. R. Z. Banadkouki (2023), "Selection of strategies to improve energy efficiency in industry: A hybrid approach using entropy weight method and fuzzy topsis," *Energy*, vol. 279, p. 128 070,.
6. Y. Lu, Y. Ge, G. Zhang, *et al.*(2023), "Evaluation of waste management and energy saving for sustainable green building through analytic hierarchy process and artificial neural network model," *Chemosphere*, vol. 318, p. 137 708.
7. M. Aghaie and S. Fazli (2012), "An improved mcdm method for maintenance approach selection: A case study of auto industry," *Management Science Letters*, vol. 2, no. 1, pp. 137–146,.
8. F. Vergara-Solana, M. E. Araneda, and G. Ponce-D´ıaz (2019), "Opportunities for strengthening aquaculture industry through multicriteria decision-making," *Re-views in aquaculture*, vol. 11, no. 1, pp. 105–118,.
9. H. Liao, Y. He, X. Wu, Z. Wu, and R. Bausys (2023), "Reimagining multi-criterion decision making by data-driven methods based on machine learning: A literature review," *Information Fusion*, p. 101 970.
10. W. Alqaisi, C. Z. El-Bayeh, K. Alzaareer, M. Zellagui, A. Flah, and B. Brahmi,(2024) "Multi-criteria decision-making (mcdm) methods for micro-grids: An overview," *Innovation and Technological Advances for Sustainability*, pp. 341–351.
11. O. A. Sianaki and M. A. Masoum (2013), "A multi-agent intelligent decision making support system for home energy management in smart grid: A fuzzy topsis approach," *Multiagent and Grid systems*, vol. 9, no. 3, pp. 181–195.
12. H. Zhao and N. Li (2016), "Performance evaluation for sustainability of strong smart grid by using stochastic ahp and fuzzy topsis methods," *Sustainability*, vol. 8, no. 2, p. 129.
13. B. Yilmaz and N. Harmancioglu (2010), "Multi-criteria decision making for water re-source management: A case study of the gediz river basin, turkey," *Water SA*, vol. 36, no. 5.
14. S. G. Meshram, E. Alvandi, C. Meshram, E. Kahya, and A. M. Fadhil Al-Quraishi (2020), "Application of saw and topsis in prioritizing watersheds," *Water Resources Management*, vol. 34, pp. 715–732.
15. B. Khodabakhshi and H. R. Jafari (2010), "Environmental impact assessment of wa-ter resources development projects using the electre tri model (a case study of ardebil reservoir, drainage, and irrigation network)," *Journal of Water and Wastewater; Ab va Fazilab (in persian)*, vol. 21, no. 3, pp. 64–74,.
16. W. Vernandhes, N. S. Salahuddin, A. Kowanda, and S. P. Sari (2017), "Smart aquaponic with monitoring and control system based on iot," in *2017 second international conference on informatics and computing (ICIC)*, IEEE, 2017, pp. 1–6.
17. M. A. Zamora-Izquierdo, J. Santa, J. A. Mart´ınez, V. Mart´ınez, and A. F. Skarmeta (2019), "Smart farming iot platform based on edge and cloud computing," *Biosystems engineering*, vol. 177, pp. 4–17,.