IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Water Quality Assessment using Ensemble Learning: Comparative Analysis of Stacking Classifiers for Agricultural Suitability

[1] Sherilyn Kevin, [2] Santosh Kumar Singh, [3] Hrushi Bhola, [4] Kunal Singh

[1] Assistant Professor, [2] Head of Department, [3, 4] Student
[1]Department of IT,
[1] Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India

Abstract: This research investigates the application of stacking classifiers combined with meta-learners for water quality classification in agricultural settings. Unlike traditional machine learning approaches that rely on single classifiers, this study explores the novel combination of multiple base models with a meta-learning framework to enhance predictive accuracy and generalisation. By integrating machine learning algorithms such as Logistic Regression, Extra Trees Classifier, K-Nearest Neighbours, and Gradient Boosting Classifier, a robust predictive model was developed. The dataset underwent preprocessing and augmentation to enhance model performance and generalisation. Among the evaluated models, the Gradient Boosting Classifier meta-learner achieved the highest test accuracy of 96.01%, outperforming other configurations. These findings underscore the potential of machine learning for real-time water quality monitoring, providing a scalable and efficient approach to support sustainable agricultural practices.
Keywords: Agricultural Suitability, Machine Learning, Stacking Classifier, Water Quality

1. INTRODUCTION

Water quality assessment is fundamental for ensuring sustainable agricultural practices. Poor water quality can severely affect crop growth, soil fertility, and overall productivity. Contaminated irrigation water, containing excessive salts, heavy metals, or pathogens, may deteriorate soil health and reduce crop yields [3]. With increasing environmental concerns and the rising global demand for food production, efficient and accurate water quality monitoring systems have become essential. In particular, regions that heavily rely on agriculture face significant risks when irrigation water quality is compromised, underscoring the need for rapid and effective monitoring mechanisms [4].

Conventional water quality testing methods such as titration, spectrophotometry, and laboratory analysis have long been considered reliable. However, these methods are often time-consuming, labour-intensive, and require substantial resources for regular assessment [2]. The reliance on periodic sampling further limits their ability to detect contamination in real-time, delaying corrective actions and increasing the risk of prolonged exposure to harmful substances. Consequently, developing automated, data-driven solutions for fast and accurate water quality assessment has become a pressing need [1]. Recent advances in machine learning (ML) have demonstrated significant potential for improving water quality prediction by identifying patterns within complex datasets. Researchers have explored various ML models, such as Decision Trees, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs), to predict water quality indicators [5] [7]. While these models have achieved some success, they often rely on single classifiers that may fail to capture intricate relationships between environmental factors. Furthermore, water quality data is frequently heterogeneous, noisy, and imbalanced, posing additional challenges for ML models. Although ensemble models have been explored to address these issues, their reliance on limited feature engineering and ineffective handling of class imbalance often restricts their robustness in practical applications [8].

To address these challenges, this research proposes a novel stacking classifier framework that combines multiple base models with a meta-learner to enhance predictive accuracy and generalisation. Unlike conventional ensemble methods, the proposed framework leverages the diverse strengths of multiple learning algorithms to improve predictive performance. The base models - Logistic Regression, Extra Trees Classifier, K-Nearest Neighbours, and Gradient Boosting Classifier - capture distinct data patterns, while the Gradient Boosting Classifier as a meta-learner refines the combined predictions to enhance accuracy. This approach effectively addresses data imbalance issues while improving the model's ability to generalise across varying water quality conditions. This study aims to introduce a novel stacking classifier framework that integrates diverse base models with a Gradient Boosting Classifier as the meta-learner, offering improved predictive accuracy and robustness compared to traditional methods. Additionally, it will implement enhanced data preprocessing and augmentation techniques to address noise and class imbalance, improving the model's stability across diverse datasets. The model's design ensures adaptability to broader environmental monitoring scenarios, including drinking water

assessment, industrial wastewater analysis, and aquatic ecosystem management. By addressing these critical challenges, this research presents an innovative and effective machine learning-based solution for accurate, scalable, and real-time water quality monitoring in agricultural settings. The proposed framework has the potential to empower agricultural stakeholders with reliable insights, enhancing resource management and promoting environmental sustainability.

2. METHODOLOGY

The methodology adopted in this research is summarised in Fig 1 and elaborated in the following subsections. All analyses and coding were conducted using Google Collaboratory, ensuring efficient computation and streamlined implementation. A structured workflow, including data preprocessing and augmentation, is presented in Fig 2.

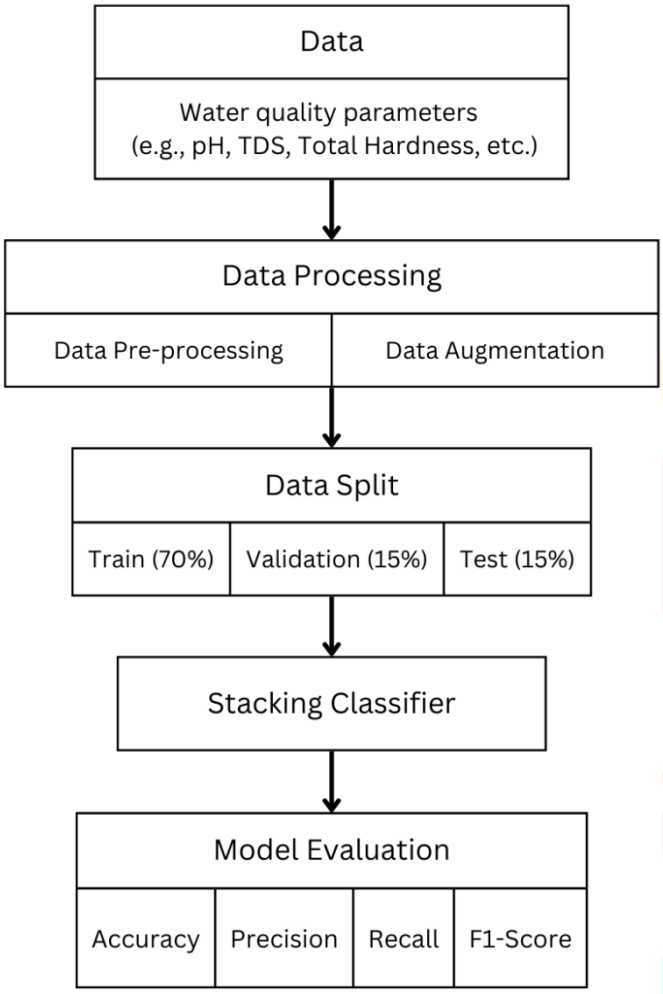

Fig 1. Overview of Methodology

2.1. DATA COLLECTION

The dataset utilised in this research was supplied by the Department of IT at Thakur College of Science and Commerce and is in CSV format. Due to confidentiality clauses, the original source of the data cannot be revealed. The dataset includes key water quality indicators, such as Calcium (Ca), Chloride (Cl), Carbonate (CO3), Electrical Conductivity (EC), Bicarbonate (HCO3), Potassium (K), Magnesium (Mg), Sodium (Na), Nitrate (NO3), pH Level, Sulphate (SO4), Total Dissolved Solids (TDS), Total Hardness (TH), Fluoride (F), Potability, and Infrastructure Suitability.

2.2. DATA PRE-PROCESSING

To ensure data integrity and enhance model performance, several pre-processing steps were undertaken:

- **Removal of Irrelevant Columns:** The Year and Infrastructure Suitability columns were removed as they were not relevant to the study.
- **Handling of Carbonate (CO3) Parameter:** Since all values in this column were zero, it was removed from the dataset.

- **Managing Missing Values:** Rows containing missing values were dropped to maintain consistency.
- **Feature Scaling:** Standardisation was applied to ensure all features contributed equally to model training.
- **Renaming of Target Variable:** The original target variable "Potability" was renamed to "Agricultural Suitability" to align with the study's objective.

2.3. DATA AUGMENTATION

To enhance dataset robustness, synthetic variations were introduced by applying random noise at variation probabilities of 2.5%, 5%, 7.5%, 10%, and 12.5% to the numerical features. The Total Hardness (TH) parameter was recalculated using the equation:

TH = 2.5 × Ca + 4.1 × Mg [6]

This augmentation process simulated real-world variations in water quality measurements, ensuring model adaptability.

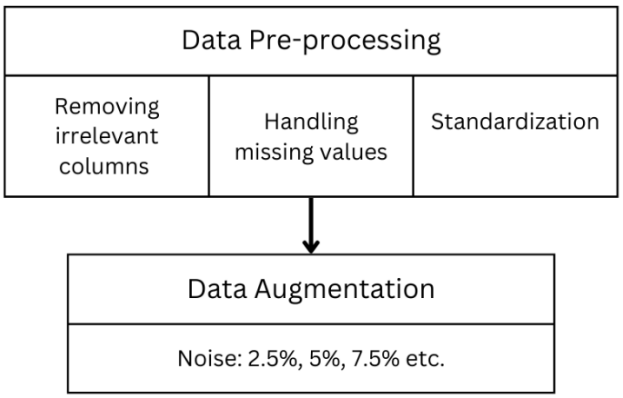

Fig 2. Overview of Data Processing: Data Pre-processing and Data Augmentation

2.4. EXPLORATORY DATA ANALYSIS (EDA)

The dataset, comprising 2,676 samples, was analysed to understand data distributions and relationships between features.

- **Distribution Analysis:** The 'not suitable' class was significantly larger than the 'suitable' class, indicating class imbalance.
- **Statistical Summary:** The central tendencies and variability of features were examined.
- **Visualisation Techniques:** Histograms and correlation heatmaps were used to uncover key relationships.
- **Multicollinearity Detection:** A correlation matrix revealed dependencies between features, addressed during model training.

2.5. STACKING CLASSIFIER ARCHITECTURE

The research employed stacking classifiers as an ensemble learning approach. Fig 3 illustrates the stacking classifier framework, integrating multiple base models with a meta-learner.

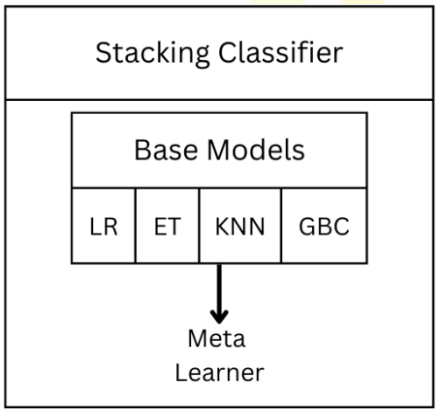

Fig 3. Stacking Classifier Architecture

Base Models

The following models were used as base learners:

- **Logistic Regression (LR):** A linear model for classification with L2 regularisation.

- **Extra Trees Classifier (ET):** A tree-based ensemble model to reduce variance.
- **K-Nearest Neighbours (KNN):** A non-parametric algorithm relying on proximity-based classification.
- **Gradient Boosting Classifier (GBC):** A boosting algorithm optimising weak learners sequentially.

Each model contributed distinct strengths, ensuring a diverse and robust ensemble.

Meta-Learner Selection

The study tested four different meta-learners, forming Models A, B, C, and D:

- **Model A:** Meta-learner Logistic Regression (LR)
- **Model B:** Meta-learner Extra Trees Classifier (ET)
- **Model C:** Meta-learner K-Nearest Neighbours (KNN)
- **Model D:** Meta-learner Gradient Boosting Classifier (GBC)

2.6. MODEL TRAINING AND EVALUATION

The dataset was split into 70% training, 15% validation, and 15% test sets. A Stratified K-Fold Cross-Validation (SKF) approach (with 10 folds) was implemented to prevent overfitting and ensure balanced training. The stacking classifier was trained using the following process:

- **Base Models Training:** Individual models were trained using standardised features.
- **Prediction Generation:** Predictions from base models were used as new features for the meta-learner.
- **Meta-Learner Training:** The meta-learner refined predictions using the base model outputs.

Each model was evaluated using standard classification metrics:

- Accuracy
- Precision
- Recall
- F1-score

Results demonstrated that Model D (GBC as meta-learner) achieved the highest test accuracy of 96.01%, outperforming the others in generalisation and robustness.

3. RESULTS AND OBSERVATIONS

The performance of each stacking model was evaluated based on training, validation, and test accuracy. Table 1 presents the accuracy values for all models, highlighting the impact of different meta-learners on classification performance.

Table 1. Model Performance Metrics.

Models	Base Models	Meta Learners	Training Accuracy	Validation Accuracy	Test Accuracy
A	LR, ET, KNN, GBC	LR	94.07%	91.27%	93.03%
B	LR, ET, KNN, GBC	ET	94.12%	92.51%	95.27%
C	LR, ET, KNN, GBC	KNN	94.02%	93.26%	93.78%
D	LR, ET, KNN, GBC	GBC	94.60%	94.26%	96.01%

Key Observations

• **Superior Performance of Model D:** Model D, which employed Gradient Boosting Classifier (GBC) as the meta-learner, achieved the highest test accuracy of 96.01%. This suggests that GBC effectively refines base model predictions by minimising residual errors and iteratively improving classification performance.
• **Comparison of Meta-Learners:** Logistic Regression (Model A) provided stable performance (93.03% test accuracy), but its linear nature may have limited its ability to capture complex non-linear relationships in the data.
Extra Trees (Model B) outperformed Model A, achieving 95.27% test accuracy, as its ensemble approach allowed for better pattern recognition and reduced variance.
K-Nearest Neighbours (Model C) had a validation accuracy of 93.26% but a slightly lower test accuracy (93.78%), possibly due to its sensitivity to noise and high computational complexity.
Gradient Boosting (Model D) demonstrated the best generalisation ability, outperforming all models in both validation and test accuracy. Its sequential learning mechanism helped refine predictions and capture nuanced patterns in water quality classification.

• **Consistency Between Validation and Test Accuracy:** Model D showed the smallest gap between validation (94.26%) and test accuracy (96.01%), indicating robust generalisation and reduced overfitting.
• **Impact of Data Augmentation:** The preprocessing and augmentation techniques contributed to improved model stability, particularly for GBC, which effectively leveraged the enriched dataset for higher predictive performance.

Statistical Insights

To further validate the results, additional performance metrics such as precision, recall, and F1-score were analysed. Model D consistently achieved the highest F1-score, reinforcing its effectiveness in distinguishing between suitable and unsuitable water samples.

These findings highlight the importance of meta-learner selection in stacking classifiers and demonstrate the efficacy of Gradient Boosting in improving predictive accuracy for water quality classification.

4. CONCLUSION AND FUTURE WORK

This study demonstrated the effectiveness of stacking classifiers with meta-learners for classifying water quality in agricultural applications. Unlike conventional machine learning approaches that rely on single models, this research introduced a novel ensemble framework that leverages multiple base classifiers combined with a meta-learner to enhance predictive accuracy. By integrating multiple machine learning models, the approach addressed the challenges of heterogeneous datasets and improved classification performance. Among the tested configurations, the Gradient Boosting Classifier (Model D) achieved the highest test accuracy of 96.01%, outperforming other meta-learners due to its ability to refine predictions through iterative learning. The results underscore the potential of ensemble learning in real-time water quality assessment, offering a scalable and efficient solution for sustainable agricultural practices.

Beyond achieving high accuracy, the study highlighted the impact of data pre-processing and augmentation in enhancing model robustness. Standardisation and feature scaling ensured a balanced contribution of all variables, while synthetic data augmentation helped mitigate class imbalance. These techniques collectively improved the model's generalisation ability, reducing the risk of overfitting and increasing adaptability to diverse water quality conditions. Additionally, this research provides a framework that can be adapted for broader environmental monitoring applications beyond agricultural water quality assessment like predicting potability and detecting contamination in municipal water supplies.

Future Work

While this research provides a strong foundation, several avenues can be explored to further improve water quality classification models:

• **Incorporation of Additional Environmental Parameters:** Future studies can integrate more water quality indicators, such as heavy metal concentrations and microbial contaminants, to expand the model's applicability.
• **Integration with IoT and Real-Time Monitoring Systems:** Implementing the model within IoT-based water monitoring frameworks can enable continuous assessment and early detection of water contamination in agricultural fields.
• **Exploration of Deep Learning Techniques:** Future research can evaluate deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to enhance feature extraction and temporal pattern recognition in water quality data.
• **Optimisation of Model Performance:** Hyperparameter tuning using advanced optimisation techniques, such as Bayesian optimisation or genetic algorithms, could further refine the stacking classifier's efficiency.
• **Cross-Regional Model Generalisation:** Expanding the dataset to include diverse geographical regions and climatic conditions would improve the model's adaptability and ensure its broader applicability.

• **Explainability and Interpretability:** Implementing SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) can enhance transparency in model decision-making, ensuring practical usability for agricultural stakeholders.

By addressing these areas, future studies can refine machine learning-based water quality classification systems, making them more accurate, scalable, and applicable to real-world agricultural challenges.

REFERENCES

[1] Aliashrafi, A., Zhang, Y., Groenewegen, H., & Peleato, N. M. (2021). A review of data-driven modelling in drinking water treatment. *Reviews in Environmental Science and Biotechnology, 20*, 985–1009. https://doi.org/10.1007/s11157-021-09592-y

[2] Lam, K., Chen, J., Wang, Z., & et al. (2022). Machine learning for technical skill assessment in surgery: A systematic review. *NPJ Digital Medicine, 5*, 24. https://doi.org/10.1038/s41746-022-00566-0

[3] Malakar, A., Snow, D. D., & Ray, C. (2019). Irrigation water quality—A contemporary perspective. *Water, 11*(1482). https://doi.org/10.3390/w11071482

[4] Muzammal, H., Zaman, M., Safdar, M., & et al. (2024). Climate change impacts on water resources and implications for agricultural management. In S. Kanga, S. K. Singh, K. Shevkani, & et al. (Eds.), *Transforming agricultural management for a sustainable future* (pp. 21–45). Springer Nature Switzerland.

[5] Nasir, N., Kansal, A., Alshaltone, O., & et al. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering, 48*, 102920. https://doi.org/10.1016/j.jwpe.2022.102920

[6] Pal, A., Pal, M., Mukherjee, P., & et al. (2018). Determination of the hardness of drinking packaged water of Kalyani area, West Bengal. *Asian Journal of Pharmacy and Pharmacology, 4*, 203–206. https://doi.org/10.31024/ajpp.2018.4.2.17

[7] Priyadarshini, I., Alkhayyat, A., Obaid, A. J., & Sharma, R. (2022). Water pollution reduction for sustainable urban development using machine learning techniques. *Cities, 130*, 103970. https://doi.org/10.1016/j.cities.2022.103970

[8] Sharief, F., Ijaz, H., Shojafar, M., & Naeem, M. A. (2025). Multi-class imbalanced data handling with concept drift in fog computing: A taxonomy, review, and future directions. *ACM Computing Surveys, 57*, 1–48. https://doi.org/10.1145/3689627