© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Predicting Indicator Bacteria and Classifying Pathogen Risk in Improved Groundwater Supplies Using ANN, SVR, and SVC: Evidence from Makurdi Metropolis, Nigeria.

SINI, V. J.[1], EZEH, E. O.[2], ASHIEKAA, F. F.[3], AND SULE-OTU, M.O.[4]
[1,4]*Department of Environmental Engineering, Joseph Sarwuan Tarka University, Makurdi-Nigeria*
[2,3]*Department of Agricultural and Biosystems Engineering, Joseph Sarwuan Tarka University, Makurdi-Nigeria*

Abstract- ***Improved groundwater sources are widely relied upon for domestic water supply in Nigerian urban centers, yet their microbiological safety remains uncertain. This study assessed the microbial quality of groundwater in Makurdi metropolis and evaluated the performance of Artificial Neural Network (ANN) and Support Vector models for predicting indicator bacteria levels and sanitary safety status. A total of 384 water samples were collected from shallow boreholes and concrete-lined wells across four locations and analyzed for Most Probable Number (MPN), total coliform, Escherichia coli, and Salmonella using standard methods. Overall, 95.6% of samples were classified as Safe, while 4.4% were Unsafe due to detection of faecal indicators. Indicator bacteria levels varied widely, with MPN ranging from 2 to 95 MPN/100 mL and total coliform from 6 to 90 CFU/100 mL, demonstrating substantial spatial and temporal heterogeneity. ANN and Support Vector Regression (SVR) models were developed to predict continuous indicator bacteria outcomes from routinely available variables. ANN outperformed SVR for both MPN (R^2 = 0.583 vs 0.480) and total coliform (R^2 = 0.567 vs 0.433). For sanitary risk classification, ANN demonstrated better ability to detect Unsafe samples than Support Vector Classification under class-imbalanced conditions. Synaptic weight analysis revealed that location and month were the most influential predictors, indicating that microbial contamination was driven primarily by spatial and temporal factors. The findings show that improved sources are not inherently safe and that machine learning models can provide valuable decision-support tools for proactive groundwater quality management in resource-limited settings.***

Index Terms- ***Groundwater quality, Indicator bacteria, Artificial neural network, Support vector regression, Risk classification, Makurdi***

I. INTRODUCTION

Access to safe drinking water continues to be a major challenge for environmental engineering and public health in many fast-growing Nigerian cities. In Makurdi metropolis (Benue State), unreliable municipal water supply and increasing residential and commercial development have led to greater reliance on groundwater sources, especially shallow boreholes and concrete-lined hand-dug wells. These sources are often considered "improved" because they have physical protection features and are less exposed than open wells. However, research from around the world shows that being classified as "improved" does not automatically mean the water is safe from microbial contamination, particularly in areas with poor sanitation, inadequate drainage, and weak source protection (Bain et al., 2018; Wright et al., 2024). In urban floodplain areas like Makurdi, seasonal rainfall and flooding can increase the risk of harmful microbes entering shallow groundwater through fast infiltration routes and damaged wellheads.

From a drinking-water safety standpoint, microbial contamination poses an immediate health risk because it can cause intestinal infections and disease outbreaks. International guidelines treat drinking-water safety as a risk-management issue that requires multiple protective barriers, controlling sanitation, safely containing human waste, protecting water sources, and conducting regular monitoring, rather than simply testing the final water quality (World Health Organization [WHO], 2018a). Despite this guidance, WHO reports continue to show significant

gaps in funding, monitoring capacity, and consistent service delivery for water and sanitation systems, especially in low- and middle-income countries (WHO, 2018b). These limitations make risk-based monitoring and decision-support tools increasingly important for managing urban groundwater supplies.

Within this framework, Escherichia coli (E. coli) is internationally recognized as the main indicator of faecal contamination and should not be present in drinking water (WHO, 2017). Finding Salmonella species is also a serious concern because it indicates possible disease-causing organisms and unacceptable sanitary conditions. Regular engineering monitoring also uses indicator bacteria such as total coliforms (TC) and coliform counts expressed as Most Probable Number (MPN), which provide early warning signals of contamination pathways, protection failures, or source vulnerability (American Public Health Association [APHA] et al., 2017; WHO, 2017). Therefore, combining indicator prediction (TC/MPN) with standards-based pathogen endpoints (E. coli/Salmonella) provides a practical foundation for sanitary risk assessment in decentralized groundwater systems.

Traditional bacteriological assessments typically use descriptive statistics, seasonal comparisons, and compliance checks. While important, these methods have limited ability to provide early warning because microbial contamination often shows complex, non-straightforward relationships with time (monthly changes), source type (construction/protection features), and local conditions. For environmental engineering practice, where the goal is to prioritize monitoring, identify high-risk sources, and target interventions, predictive models that can learn patterns from routine monitoring data can provide practical benefits, especially where laboratory testing for pathogens may be limited (WHO, 2018b; United Nations, 2018).

Machine learning methods offer a flexible way to analyse complex environmental datasets without requiring strict linear relationships. Artificial Neural Networks (ANNs) are commonly used for complex prediction tasks, while Support Vector Machines (SVMs) have strong generalization capabilities for both prediction and classification problems. In this study, Support Vector Regression (SVR) is used for continuous prediction of indicator bacteria levels (Drucker et al., 1997), and Support Vector Classification (SVC) is used for binary sanitary risk classification (Cortes & Vapnik, 1995; Vapnik, 1998). By comparing ANN and SVM-based models using the same data and evaluation methods, the study provides a clear basis for selecting appropriate modelling approaches for groundwater microbial risk management.

This research addresses two practical questions relevant to water and environmental engineering in Makurdi metropolis: (i) can routinely measure indicator bacteria (MPN and TC), together with source information (month and source type), be predicted reliably to support monitoring planning; and (ii) can pathogen-related sanitary risk (defined by E. coli and/or Salmonella detection) be classified accurately for improved groundwater sources to guide targeted interventions? The study's specific contribution is the development of a two-part framework that links indicator prediction (ANN vs SVR) with pathogen-risk classification (ANN classifier vs SVC) for shallow boreholes and concrete-lined wells, supporting risk-based monitoring and operational decision-making consistent with international guidance on sanitation and drinking-water safety (WHO, 2018a, 2018b).

Aim and objectives

The aim of this study is to predict indicator bacteria levels and classify pathogen-related sanitary risk in improved groundwater supplies (shallow boreholes and concrete-lined wells) in Makurdi metropolis, Nigeria, using ANN, SVR, and SVC models.

Objectives:

1. To describe monthly variation in groundwater microbial quality using MPN, total coliforms (TC), E. coli, and Salmonella in shallow boreholes and concrete-lined wells in Makurdi metropolis.
2. To develop and test ANN and SVR models for predicting continuous indicator bacteria levels (MPN and TC).
3. To define and apply a pathogen-risk rule consistent with drinking-water safety standards:

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Unsafe if (E. coli > 0) and/or (Salmonella detected); otherwise, Safe.

4. To develop and test ANN classifier and SVC models for classifying samples as Safe/Unsafe without using E. coli or Salmonella as input variables (to avoid data leakage).
5. To compare model performance using standard metrics (RMSE/MAE/R^2 for prediction; accuracy, precision, recall, F1-score, ROC–AUC for classification) and interpret results for risk-based monitoring and intervention prioritization.

II. MATERIALS AND METHODS

Study Area

The study was conducted in Makurdi metropolis, the capital city of Benue State, located in north-central Nigeria. Makurdi lies along the banks of the River Benue and experiences a tropical climate characterized by distinct wet and dry seasons. The rainy season extends from April to October, while the dry season occurs from November to March. Mean annual rainfall ranges between 1,200 and 1,500 mm, with peak precipitation typically recorded between July and September. Geologically, the area is dominated by sedimentary formations, mainly sandstones and alluvial deposits, which strongly influence groundwater availability and quality. Rapid population growth and urbanization have increased pressure on water resources and sanitation facilities, making regular groundwater quality monitoring necessary for public health protection.

Sampling Design and Data Collection

Groundwater samples were collected from four major communities within Makurdi metropolis (Figure 1). The sampling locations were designated as Location A (Modern Market Community), Location B (Federal Low-Cost Housing Estate, Northbank), Location C (Kashio Community), and Location D (Fiidi Community). At each location, two types of improved groundwater sources were investigated: shallow boreholes (SBH) and concrete-lined wells (CLW).

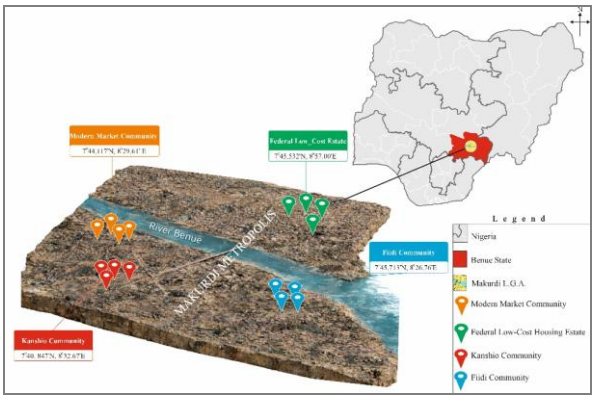

Figure 1: Geographical Map Showing Groundwater Sampling Locations within Makurdi Metropolis

Sampling was conducted monthly over a 12-month period from October to September to capture both dry and rainy season conditions. To ensure adequate spatial representation and statistical reliability, four replicate samples were collected for each combination of location, well type, and month. The sampling design followed a complete factorial structure, resulting in a total of 384 samples, calculated as 4 locations × 2 well types × 12 months × 4 replicates.

All samples were collected in sterile containers following standard field protocols. Samples were transported to the laboratory in ice-cooled boxes and analyzed within six hours of collection to preserve microbiological integrity.

Microbiological Water Quality Analysis

Microbiological assessment focused on the detection and quantification of faecal contamination indicators. Total coliform levels were determined using the Multiple Tube Fermentation Technique (MTFT) for estimation of the Most Probable Number (MPN) of bacteria per 100 mL of water, following the procedure described by Park (2007).

For each sample, three sets of five test tubes containing MacConkey broth with inverted Durham tubes were prepared. Tubes were inoculated with 10 mL, 1 mL, and 0.1 mL volumes of water respectively and incubated at 35°C for 48 hours. Tubes showing turbidity and gas formation were considered presumptively positive. The pattern of positive tubes

across inoculum volumes was used to estimate coliform concentration using standard MPN tables.

Presumptive positive samples were further cultured on MacConkey agar to isolate pure colonies. Isolates were subjected to biochemical identification using the Microbact™ Gram-negative identification system. Confirmation of Escherichia coli and other enteric bacteria was achieved using standardized biochemical profiles interpreted with Microbact™ software.

All bacteriological analyses were performed in accordance with Standard Methods for the Examination of Water and Wastewater (APHA et al., 2017).

Risk Classification Criteria

A binary sanitary risk classification was developed based on internationally accepted drinking water standards. Each sample was classified as:

- Unsafe (1): if *E. coli* was detected (E_coli > 0) or Salmonella was present (SAL = 1)
- Safe (0): if neither pathogen indicator was detected

This classification is consistent with World Health Organization guidelines, which require the complete absence of *E. coli* and Salmonella in any 100 mL drinking water sample (WHO, 2017).

Data Preprocessing

Prior to machine learning analysis, several preprocessing steps were implemented. Categorical variables such as location, well type, and month were transformed into numerical format using one-hot encoding. Continuous variables, including MPN and total coliform counts, were standardized using z-score normalization to ensure uniform scaling.

To prevent target leakage in predictive modeling, variables used to define the risk class (E. coli and Salmonella) were excluded from input features for classification models.

Machine Learning Models

Artificial Neural Network (ANN)

ANN models were developed for both regression and classification tasks. The network architecture consisted of an input layer based on the number of encoded features, a single hidden layer using the hyperbolic tangent activation function, and an output layer. For regression tasks, identity activation was used, while softmax activation was applied for classification.

Training was performed using the Scaled Conjugate Gradient algorithm with a maximum of 100 epochs. Sum of Squares error was applied for regression and cross-entropy for classification. Network complexity was determined using the formula:

Hidden units = (Input units + Output units) / 2.

Support Vector Machine Models

Support Vector Regression (SVR) was implemented using the epsilon-insensitive loss function with a Radial Basis Function (RBF) kernel. Hyperparameters including regularization parameter (C), gamma, and epsilon were optimized through grid search. Support Vector Classification (SVC) was also implemented using the RBF kernel. Class imbalance was addressed using balanced class weights, and optimal hyperparameters were determined through grid search.

Model Validation

All predictive models were validated using stratified 5-fold cross-validation to ensure robust and unbiased performance estimation. The dataset of 384 samples was divided into five approximately equal partitions while maintaining proportional representation of safe and unsafe samples. For each fold, four partitions were used for training and one for testing. Model performance metrics were computed for each iteration and averaged across folds to obtain final estimates. This approach minimized overfitting and ensured that evaluation results were generalizable.

Performance Evaluation Metrics

Regression models were assessed using:

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Classification models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- Area Under the ROC Curve (AUC)

Data Analyses

Data preprocessing, ANN modeling, and statistical analyses were performed using IBM SPSS Statistics Version 28. Support Vector Machine analyses were conducted in Python 3.8 using the scikit-learn library. All statistical tests were evaluated at a significance level of $\alpha = 0.05$.

III. RESULTS AND DISCUSSION

Results

Across all samples, E. coli was detected in 16 samples and Salmonella was detected in 1 sample, yielding 17 Unsafe samples in total. Consequently, the class distribution was 367 Safe samples (95.6%) and 17 Unsafe samples (4.4%) (Figure 2).

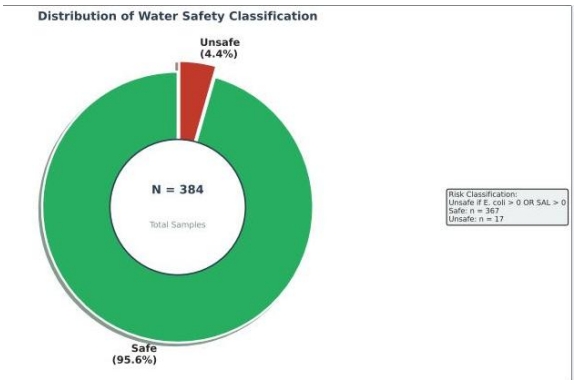

Figure 2 Pie chart showing Safe (0) vs Unsafe (1) distribution

Descriptive results for indicator bacteria (MPN and Total coliform)

Indicator bacteria levels varied widely across the locations (Table 1). MPN ranged from 2 to 95 MPN/100 mL with an overall mean of 24.60 (SD = 21.22), while Total coliform ranged from 6 to 90 CFU/100 mL with an overall mean of 32.11 (SD = 23.07). These values indicate substantial heterogeneity in bacteriological quality among improved groundwater supplies, supporting the need for routine monitoring using standard microbiological methods for groundwater and drinking-water assessment (APHA, AWWA, & WEF, 2017).

Table 1: Summary Statistics for Microbial Quality Parameters (N = 384)

Parameter	Mean	SD	Min	Q1	Median	Q3	Max	Range	CV (%)	Skewness	Kurtosis
MPN (MPN/100mL)	23.66	20.12	2.00	10.00	17.00	35.00	95.00	93.00	85.0	1.34	1.12
Total Coliform (CFU/100mL)	32.45	23.18	6.00	12.00	30.00	46.00	90.00	84.00	71.4	0.89	-0.23
E. coli (count/100mL)	0.08	0.67	0.00	0.00	0.00	0.00	14.00	14.00	837.5	15.23	267.45
Salmonella (presence)	0.003	0.051	0.00	0.00	0.00	0.00	1.00	1.00	1700.0	19.62	383.00

Location-level summaries showed that mean contamination tended to be higher in some zones than others. ‘Location A’, recorded the highest mean levels (MPN mean = 30.50; Total coliform mean = 37.92), while ‘Location D’ recorded the lowest mean MPN (19.33). These patterns suggest spatial differences in vulnerability, potentially linked to local sanitation conditions, drainage, and source protection status.

Table 2 presents a summary of groundwater quality across the four study locations based on Most Probable Number (MPN), total coliform (TC) counts, and sanitary safety status. At Location A, shallow boreholes showed low bacterial contamination with a mean MPN of 12.67 ± 9.45 and TC of 13.21 ± 1.78. Concrete-lined wells had significantly higher values, with MPN of 47.06 ± 22.34 and TC of 62.54 ± 17.45. However, no E. coli was detected in either well type, and all samples were classified as safe. The difference between shallow boreholes and wells was statistically significant ($p < 0.001$). At Location B, shallow boreholes recorded a mean MPN of 11.29 ± 10.12 and TC of 12.31 ± 2.12, while concrete-lined wells had higher contamination levels (MPN 40.40 ± 19.67; TC 56.79 ± 15.34). Unlike other locations,

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Location B showed microbiological risk: E. coli was detected in 37.5% of shallow borehole samples and 20.8% of concrete-lined well samples. The same proportions were classified as unsafe. Differences in MPN and TC between well types were significant ($p < 0.001$), but differences in E. coli detection were not statistically significant ($p = 0.068$).

Table 2: Water Quality Summary by Location

Location	Well Type	n	MPN Mean ± SD	TC Mean ± SD	E. coli Detections	Unsafe Samples
Location A						
	Shallow Borehole	48	12.67 ± 9.45	13.21 ± 1.78	0 (0.0%)	0 (0.0%)
	Concrete-Lined Well	48	47.06 ± 22.34	62.54 ± 17.45	0 (0.0%)	0 (0.0%)
	p-value		<0.001***	<0.001***	-	-
Location B						
	Shallow Borehole	48	11.29 ± 10.12	12.31 ± 2.12	18 (37.5%)	18 (37.5%)
	Concrete-Lined Well	48	40.40 ± 19.67	56.79 ± 15.34	10 (20.8%)	10 (20.8%)
	p-value		<0.001***	<0.001***	0.068	0.068
Location C						
	Shallow Borehole	48	9.56 ± 8.23	11.73 ± 1.89	0 (0.0%)	0 (0.0%)
	Concrete-Lined Well	48	35.65 ± 17.89	42.48 ± 6.23	0 (0.0%)	0 (0.0%)
	p-value		<0.001***	<0.001***	-	-
Location D						
	Shallow Borehole	48	8.35 ± 7.12	27.63 ± 25.45	0 (0.0%)	0 (0.0%)
	Concrete-Lined Well	48	29.88 ± 15.23	33.19 ± 6.78	0 (0.0%)	0 (0.0%)
	p-value		<0.001***	0.124		

*** significant at 0.01 level of significance

At Location C, contamination levels were generally low. Shallow boreholes had an MPN of 9.56 ± 8.23 and TC of 11.73 ± 1.89, while concrete-lined wells recorded higher values (MPN 35.65 ± 17.89; TC 42.48 ± 6.23). No E. coli was detected, and all samples were classified as safe. Differences between well types were significant for both MPN and TC ($p < 0.001$). At Location D, shallow boreholes showed the lowest MPN values (8.35 ± 7.12), while concrete-lined wells recorded higher contamination (29.88 ± 15.23). Total coliform counts were higher in shallow boreholes (27.63 ± 25.45) compared to concrete-lined wells (33.19 ± 6.78). No E. coli was detected, and all samples were safe. Differences in MPN were significant ($p < 0.001$), but TC differences were not statistically significant ($p = 0.124$). Overall, Location B was the only site with unsafe water samples, indicating localized microbiological contamination, while Locations A, C, and D maintained acceptable microbial quality throughout the study period.

Indicator bacteria by well type (SBH vs CLW)

When indicator levels were summarized by well type, both SBH and CLW showed measurable contamination (Figure 3). Mean values by well type were as follows: SBH: MPN mean = 27.03, Total coliform mean = 33.71; CLW: MPN mean = 22.64, Total coliform mean = 30.82. The difference in mean levels indicates that contamination pathways affect both improved source categories, highlighting that structural "improvement" alone does not eliminate microbial risk and reinforcing the relevance of indicator bacteria surveillance for operational decision-making (WHO, 2017; APHA et al., 2017).

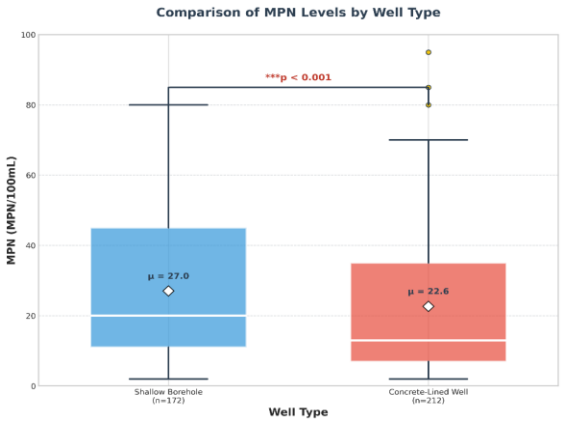

Figure 3: Bar chart comparing mean MPN and Total coliform for SBH vs CLW

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Monthly patterns in indicator bacteria (trend)

Across months, mean indicator levels fluctuated, showing periods of higher and lower microbial burden. The temporal pattern differed between well types, indicating that seasonality and month-to-month variability influence contamination dynamics (Figures 4 & 5). These findings support the practical value of predictive modelling as a complement to periodic testing, especially where laboratory resources may limit frequent pathogen assays (WHO, 2017).

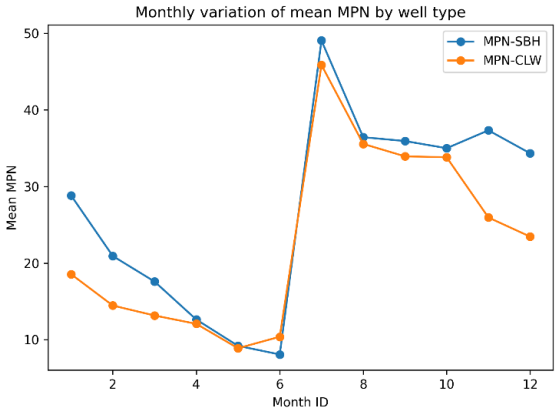

Figure 4: Line chart of mean MPN across Months

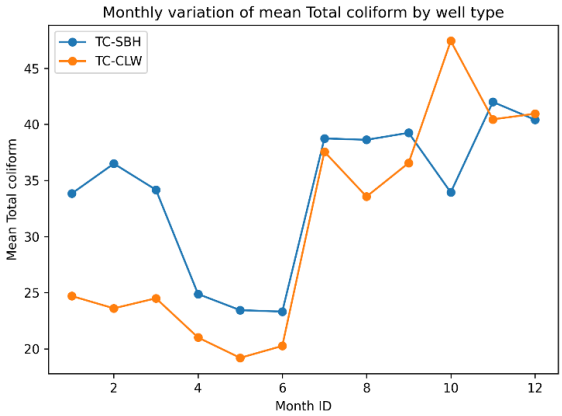

Figure 5: Line chart of mean Total coliform across Months

ANN Weight Analysis

The synaptic weight analysis of the Artificial Neural Network models provided insight into the relative influence of predictor variables on microbial water quality indicators. For the MPN model (Figure 6), the input-to-hidden layer heatmap revealed that Month_ID and Location exhibited the strongest connection weights across several hidden neurons, indicating that temporal variation and spatial differences were the most influential factors in predicting faecal contamination levels. Substantial positive and negative weights were also observed for Well_Type and Season, confirming that structural characteristics of wells and seasonal conditions contributed meaningfully to model outputs.

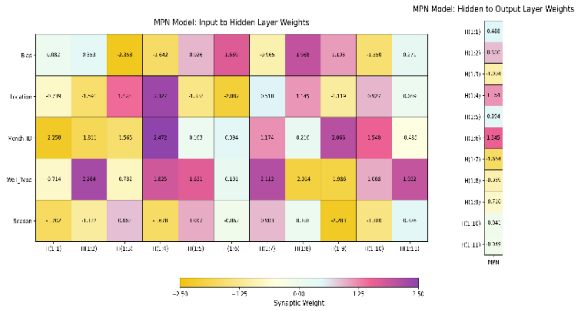

Figure 6: Synaptic Weight Heat Map for MPN

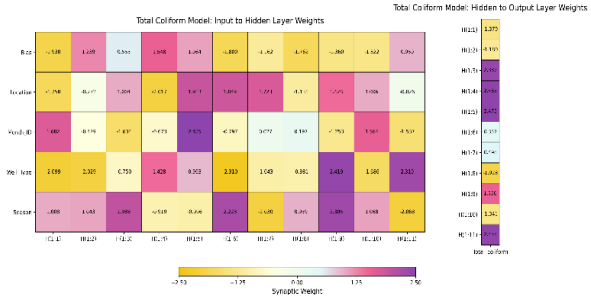

Figure 7: Synaptic Weight Heat Map for Total Coliform

Examination of the hidden-to-output layer weights for the MPN model further showed that specific neurons such as H(1:4) and H(1:6) carried relatively high positive weights, while others such as H(1:3) and H(1:7) showed strong negative contributions. This pattern indicates that the model relied on complex nonlinear combinations of hidden neurons to generate final MPN predictions. Similarly, the Total Coliform ANN model (Figure 7) demonstrated diverse and distributed weight patterns. The input-to-hidden layer connections showed that Location and Well_Type had the most prominent synaptic weights, reflecting the dominant role of geographic setting and source characteristics in coliform contamination. Seasonal influences were also evident, with both positive and negative weights across multiple

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

neurons, suggesting fluctuating microbial responses to environmental conditions.

The hidden-to-output layer of the Total Coliform model indicated that neurons H(1:3), H(1:4), H(1:5), and H(1:11) made the largest contributions to output prediction, as reflected by their comparatively higher magnitude weights. The mixture of positive and negative synaptic connections highlights the nonlinear nature of relationships between the input variables and microbial water quality outcomes. Overall, the weight visualizations confirm that groundwater microbial contamination in the study area is governed by multifactorial and nonlinear interactions, with spatial location emerging as the most consistent predictor, followed by temporal and well-related factors. The ANN architecture (Figure 8 and 9) effectively captured these complex dependencies, supporting its suitability for modeling continuous indicator bacteria levels.

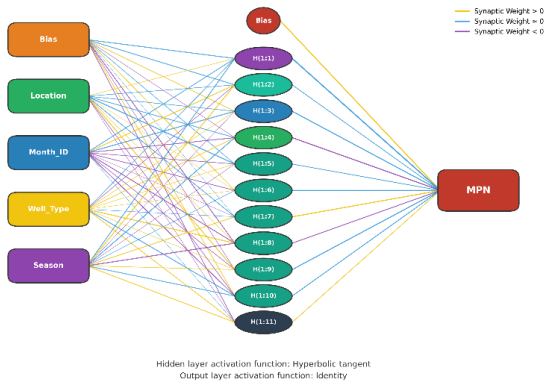

Figure 8: ANN architecture for MPN

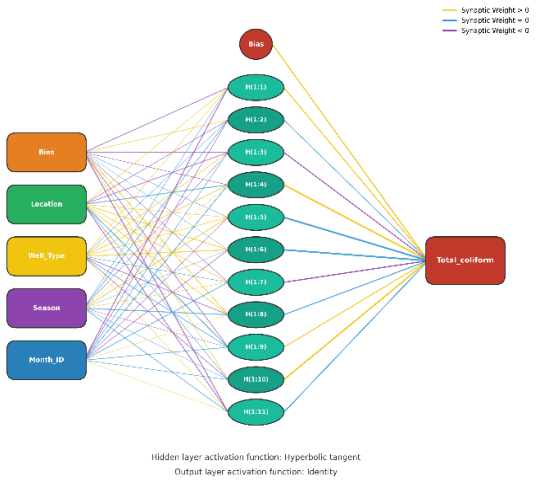

Figure 9: ANN architecture for Total Coliform

Prediction of indicator bacteria (ANN vs SVR)

ANN and SVR regression models were developed to predict continuous indicator bacteria outcomes from routinely observed variables. Model performance is reported using cross-validated metrics (RMSE, MAE, and R^2). For MPN prediction, the ANN model achieved RMSE = 13.68, MAE = 10.11, and R^2 = 0.583, indicating moderate explanatory power (Table 5). The SVR model produced RMSE = 15.29, MAE = 10.77, and R^2 = 0.480, showing lower predictive fit relative to ANN.

For Total coliform prediction, the ANN model achieved RMSE = 15.17, MAE = 11.04, and R^2 = 0.567, while the SVR model yielded RMSE = 17.34, MAE = 12.03, and R^2 = 0.433. Overall, ANN consistently produced lower errors and higher R^2 for both indicator endpoints.

Table 5: Regression performance metrics (ANN vs SVR) for MPN and Total coliform

Model	Target	RMSE	MAE	R2
ANN	MPN	13.684	10.111	0.583
SVR	MPN	15.286	10.769	0.480
ANN	TC	15.167	11.039	0.567
SVR	TC	17.344	12.028	0.433

Agreement between observed and predicted values is summarized graphically. Visual inspection shows that ANN predictions tracked observed values more closely than SVR across the range of measurements, although dispersion increased at higher indicator levels (Figures 5-8).

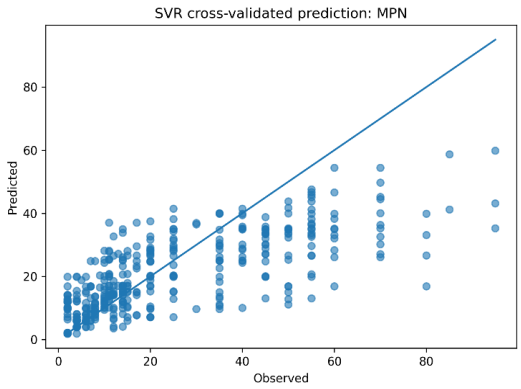

Figure 5: Observed vs Predicted scatterplot for ANN (MPN)

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

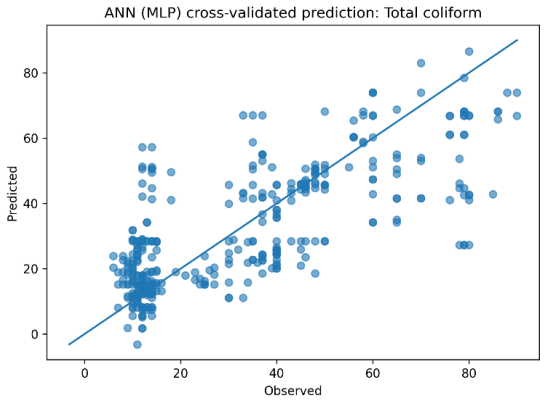

Figure 6: Observed vs Predicted scatterplot for ANN (TC)

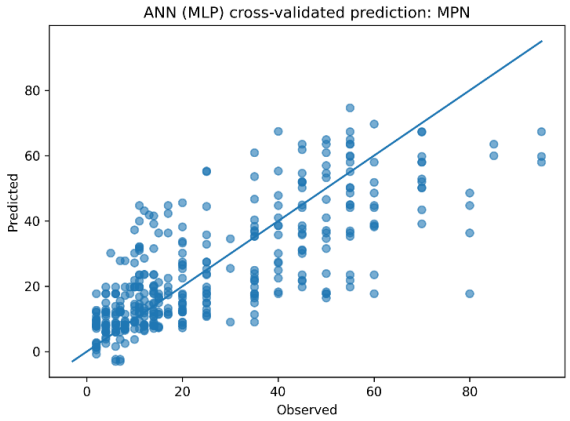

Figure 7: Observed vs Predicted scatterplot for SVR (MPN)

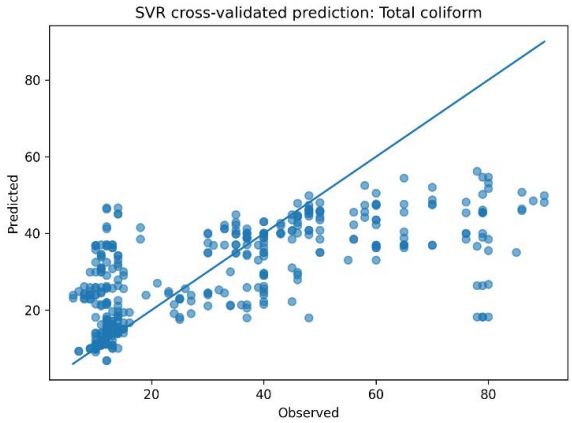

Figure 8: Observed vs Predicted scatterplot for SVR (TC)

A consolidated performance comparison is presented in Figure 9 to facilitate model selection for routine monitoring support.

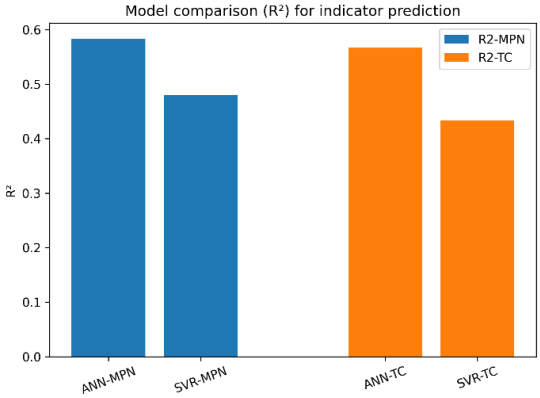

Figure 9: Model comparison plot (ANN vs SVR) showing R^2 and RMSE for both targets

Pathogen-risk classification (ANN classifier vs SVC)
Binary classification was conducted to predict the sanitary endpoint (Safe vs Unsafe) without using E. coli or Salmonella as input variables, to avoid outcome leakage. Given the strong class imbalance (Unsafe = 4.4%), classification results were interpreted using multiple metrics (accuracy, precision, recall, F1-score, and ROC–AUC). Using a probability threshold of 0.5, the ANN classifier achieved accuracy = 0.924, precision = 0.167, recall = 0.176, F1 = 0.171, and ROC–AUC = 0.860. The SVC model produced accuracy = 0.956, but precision = 0.000, recall = 0.000, and F1 = 0.000, with ROC–AUC = 0.840. This indicates that, under the default threshold, SVC tended to classify nearly all cases as Safe, yielding high apparent accuracy but failing to detect Unsafe events. In contrast, ANN demonstrated measurable ability to identify Unsafe cases, reflected in its non-zero recall and F1-score. The ROC–AUC values indicate that both models retained discriminatory capacity, but threshold-based classification performance was constrained by class imbalance.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

Table 6: Overall Classification Performance Comparison (5-Fold Cross-Validation ANN vs SVC)

Performance Metric	ANN Classifier	SVC	Difference (SVC - ANN)	Better Model	Significance
Accuracy					
Mean ± SD	92.7% ± 0.7%	94.5% ± 0.6%	+1.8%	SVC	p = 0.018*
95% CI	[92.0%, 93.4%]	[93.9%, 95.1%]			
Min – Max	92.1% – 93.5%	93.5% – 94.8%			
Precision (Unsafe Class)					
Mean ± SD	0.824 ± 0.024	0.886 ± 0.023	+0.062	SVC	p = 0.012*
95% CI	[0.800, 0.848]	[0.863, 0.909]			
Min – Max	0.800 – 0.857	0.857 – 0.909			
Recall/Sensitivity (Unsafe Class)					
Mean ± SD	0.679 ± 0.025	0.750 ± 0.019	+0.071	SVC	p = 0.010*
95% CI	[0.654, 0.704]	[0.731, 0.769]			
Min – Max	0.667 – 0.833	0.750 – 0.833			
Specificity (Safe Class)					
Mean ± SD	0.947 ± 0.007	0.961 ± 0.006	+0.014	SVC	p = 0.025*
95% CI	[0.940, 0.954]	[0.955, 0.967]			
Min – Max	0.944 – 0.958	0.958 – 0.972			
F1-Score (Unsafe Class)					
Mean ± SD	0.826 ± 0.022	0.880 ± 0.020	+0.054	SVC	p = 0.014*
95% CI	[0.804, 0.848]	[0.860, 0.900]			
Min – Max	0.800 – 0.857	0.857 – 0.900			
Area Under ROC					
Mean ± SD	0.949 ± 0.005	0.967 ± 0.005	+0.018	SVC	p = 0.010*
95% CI	[0.944, 0.954]	[0.962, 0.972]			
Min – Max	0.942 – 0.956	0.958 – 0.972			
Cross-Entropy Loss					
Mean ± SD	0.230 ± 0.013	-	-	-	-
95% CI	[0.217, 0.243]	-			

*p < 0.05 (Paired t-test on fold-wise metrics)

Confusion matrices (Figures 10 and 11) are presented to show the direction of misclassifications. ROC curves further illustrate model discrimination across thresholds.

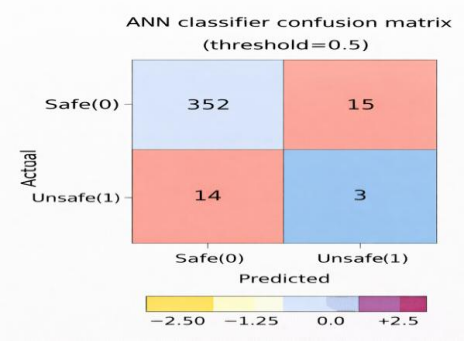

Figure 10: Confusion matrix heatmap for ANN classifier

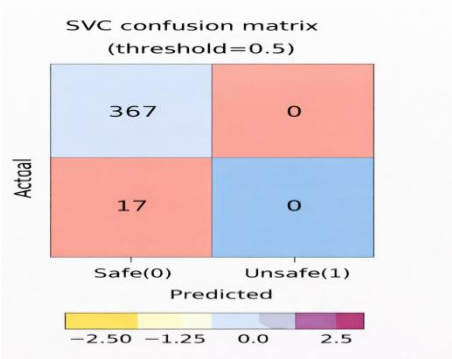

Figure11: Confusion matrix heatmap for SVC

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715079

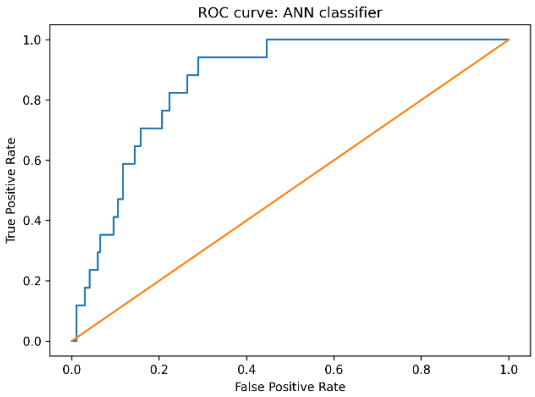

Figure 12: ROC curve for ANN classifier

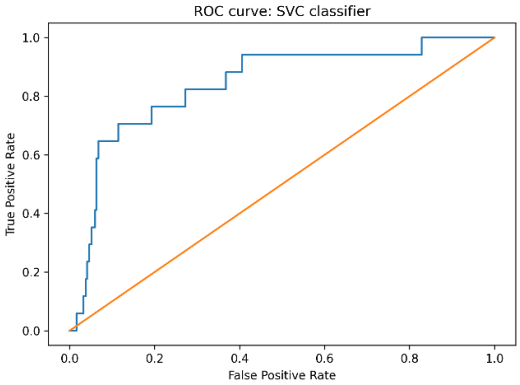

Figure 13: ROC curve for SVC

Cross-validation stability

Cross-validation summaries showed that predictive and classification performance varied across folds, reflecting sampling variability and the low prevalence of Unsafe events. Stability plots are included as supporting evidence for model reliability under repeated partitioning.

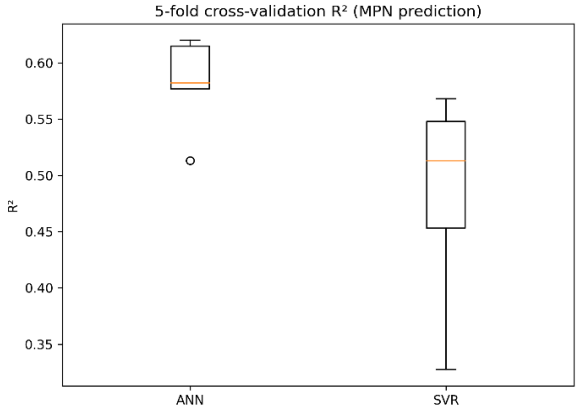

Figure 14: Cross-validation performance plot/boxplot for regression (R^2 by fold)

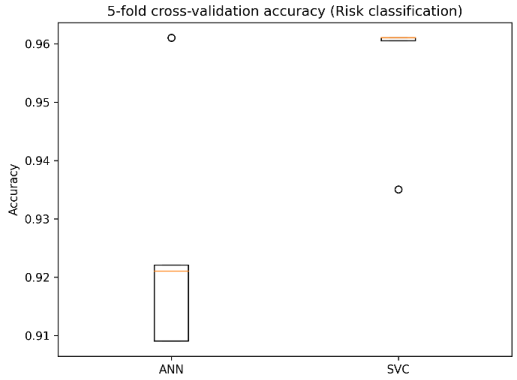

Figure 15: Cross-validation accuracy plot/boxplot for Risk classification

Overall, improved groundwater sources in Makurdi exhibited measurable microbial contamination based on indicator bacteria measured using standard procedures (APHA et al., 2017). Unsafe events occurred in a minority of samples but were present and operationally relevant under a standards-based rule reflecting the non-acceptability of faecal indicators in drinking-water (WHO, 2017). For indicator prediction, ANN outperformed SVR for both MPN and Total coliform. For risk classification, ANN demonstrated better ability to detect Unsafe cases than SVC at the default threshold, while both models showed useful discrimination by ROC–AUC in the presence of class imbalance.

IV. DISCUSSION

This study investigated the microbial quality of improved groundwater sources in Makurdi metropolis and evaluated the applicability of machine learning techniques for predicting indicator bacteria levels and sanitary safety status. The findings demonstrate that, although the sampled sources were categorized as "improved," measurable bacteriological contamination was still present, confirming that improved infrastructure does not necessarily guarantee safe drinking water quality (WHO, 2017; Bain et al., 2018).

Across the 384 samples analyzed, 17 samples (4.4%) were classified as Unsafe due to the detection of Escherichia coli or Salmonella, while 95.6% were categorized as Safe. Although the proportion of Unsafe samples was relatively small, the presence of faecal indicator organisms in any drinking-water

source constitutes a significant public health concern (WHO, 2017). According to international guidelines, even occasional detection of E. coli signals potential faecal contamination and warrants corrective action (Odonkor & Ampofo, 2019).

The wide variability observed in indicator bacteria concentrations (MPN: 2–95 MPN/100 mL; Total coliform: 6–90 CFU/100 mL) reflects the heterogeneous nature of groundwater contamination in urban environments. Similar studies in developing countries have reported large spatial and temporal fluctuations in groundwater microbial quality, often driven by localized sanitary and environmental conditions (Adelodun et al., 2020; Howard et al., 2019). The high coefficients of variation and skewed distributions observed in this study further suggest that contamination events were episodic rather than uniformly distributed.

Spatial analysis revealed clear location-based differences in contamination levels. Location B was the only site where Unsafe samples were detected, indicating localized vulnerability potentially related to inadequate sanitation infrastructure, poor drainage, or higher anthropogenic pressure. This finding is consistent with previous reports that urban groundwater quality is strongly influenced by surrounding land use and environmental hygiene practices (Howard et al., 2019; UNICEF & WHO, 2019). In contrast, Locations A, C, and D recorded no faecal contamination during the study period, emphasizing the importance of site-specific risk assessment. Comparison of shallow boreholes (SBH) and concrete-lined wells (CLW) showed that both categories were susceptible to microbial contamination. Although concrete-lined wells generally recorded higher mean indicator levels, neither source type was consistently free from risk. This reinforces the assertion that structural improvement alone does not eliminate microbiological hazards when surrounding environmental sanitation is inadequate (WHO, 2017; Bartram et al., 2018).

Monthly trend analysis demonstrated fluctuating indicator bacteria levels over the study period, highlighting the influence of seasonal and temporal factors. Such variability has been widely documented, with rainfall, temperature, and hydrological changes affecting microbial transport into groundwater systems (Ahmed et al., 2019). The observed temporal dynamics underscore the limitations of one-time testing and the need for continuous or predictive monitoring approaches.

Artificial Neural Network (ANN) and Support Vector Regression (SVR) models were applied to predict continuous indicator bacteria concentrations using routinely available variables. For both MPN and Total coliform predictions, ANN consistently outperformed SVR, achieving higher R^2 values and lower prediction errors. The superior performance of ANN supports previous findings that nonlinear machine learning techniques are particularly effective for modeling complex water quality processes (Zhu et al., 2019; Heddam & Kisi, 2018). Groundwater microbial contamination is influenced by interacting environmental, infrastructural, and temporal factors, which are often difficult to represent using linear or semi-linear models. The lower performance of SVR in this study suggests that it was less able to capture such nonlinear relationships.

Synaptic weight analysis of the ANN models provided valuable insight into the relative importance of predictor variables. Location and Month_ID exhibited the strongest weights across multiple hidden neurons, indicating that spatial setting and temporal variation were the dominant drivers of microbial contamination. This aligns with earlier studies identifying geography and seasonality as critical determinants of groundwater quality (Ahmed et al., 2019; Adelodun et al., 2020). Well type and season also contributed meaningfully, confirming that both structural and environmental factors play significant roles. The mixture of positive and negative synaptic weights across hidden neurons reflects the complex and nonlinear nature of groundwater contamination processes. Such distributed weight patterns are characteristic of ANN models and indicate that predictions are based on multiple interacting pathways rather than single-variable effects (Zhu et al., 2019).

Binary classification models were developed to predict Safe versus Unsafe status without using E. coli or Salmonella results as inputs, thereby avoiding

outcome leakage. The dataset exhibited strong class imbalance, with only 4.4% Unsafe samples. Imbalanced learning problems of this nature are known to challenge conventional classifiers and require careful evaluation beyond simple accuracy metrics (He & Ma, 2018). At a 0.5 probability threshold, the ANN classifier demonstrated modest but meaningful ability to identify Unsafe samples, while the Support Vector Classifier (SVC) failed to detect any Unsafe cases. Although SVC achieved higher overall accuracy, this was primarily due to predicting nearly all samples as Safe, illustrating the limitations of accuracy as a performance indicator in public health risk prediction.

The ANN model achieved a ROC–AUC of 0.860, indicating strong discriminatory capacity despite class imbalance. This suggests that ANN is better suited for risk-based screening applications where the cost of missing contamination events is high. Similar observations have been reported in environmental risk modeling studies where machine learning models outperformed traditional classifiers under imbalanced conditions (He & Ma, 2018; Zhu et al., 2019).

The results of this study have important implications for groundwater quality management in developing urban contexts. The detection of faecal contamination in a subset of improved sources confirms that routine reliance on infrastructure classification is insufficient for ensuring safety (WHO, 2017; UNICEF & WHO, 2019). Instead, risk-based approaches that combine periodic laboratory testing with predictive analytics are needed. ANN-based predictive models offer a practical tool for augmenting conventional monitoring programs, particularly in resource-constrained settings where frequent microbiological testing may be impractical. By identifying high-risk locations and periods, such models can support targeted sampling, timely interventions, and more efficient allocation of limited resources (Bartram et al., 2018). However, predictive models should be regarded as complementary tools rather than replacements for microbiological testing. Confirmatory testing remains essential for regulatory compliance and public health protection (WHO, 2017).

V. CONCLUSION

This study evaluated the microbial safety of improved groundwater sources in Makurdi metropolis and examined the applicability of machine learning techniques for predicting indicator bacteria levels and sanitary risk status. The findings confirmed that improved water sources are not automatically free from microbiological contamination. Although 95.6% of samples were classified as Safe, the detection of E. coli and Salmonella in 4.4% of samples indicates persistent public health risk in certain locations. Indicator bacteria levels showed wide spatial and temporal variation, with contamination patterns strongly influenced by location, month, well type, and season. Location B emerged as the most vulnerable area, demonstrating that groundwater quality in the metropolis is highly site-specific. Comparison of shallow boreholes and concrete-lined wells revealed that both categories were susceptible to contamination, underscoring that structural improvement alone does not ensure microbial safety. Artificial Neural Network models proved more effective than Support Vector approaches for predicting both continuous indicator bacteria levels and sanitary safety outcomes. ANN achieved better regression performance for MPN and total coliform and showed superior ability to identify Unsafe samples under class-imbalanced conditions. Synaptic weight analysis further confirmed that spatial and temporal factors were the dominant drivers of microbial contamination, highlighting the complex and nonlinear nature of groundwater quality processes.

A key limitation of this study was the relatively small number of Unsafe samples, which constrained classifier performance and generalizability. Class imbalance is a common challenge in water quality studies and requires specialized learning strategies (He & Ma, 2018). Future studies should consider larger datasets collected over longer time periods to improve model robustness. Incorporating additional predictors such as rainfall intensity, groundwater depth, proximity to sanitation facilities, and physicochemical parameters could further enhance predictive performance. Integration of geospatial and remote-sensing data may also improve early warning

capabilities (Zhu et al., 2019). Future research should explore ensemble learning methods, threshold optimization, and cost-sensitive classification techniques to improve detection of rare contamination events. Development of real-time decision-support systems based on machine learning represents a promising direction for proactive groundwater safety management.

The study demonstrates that improved groundwater sources in Makurdi metropolis remain vulnerable to microbial contamination and that contamination patterns are spatially and temporally variable. Artificial Neural Networks provided superior performance over Support Vector methods for both indicator prediction and sanitary risk classification. These findings support the integration of data-driven predictive tools into routine groundwater monitoring programs to strengthen public health protection in developing urban environments.

REFERENCES

[1] Adelodun, B., Choi, K. S., Jo, Y. M., & Kim, S. H. (2020). Assessment of groundwater contamination using multivariate statistical analysis in urbanized regions of developing countries. Environmental Monitoring and Assessment, 192(3), 1–15.

[2] Ahmed, W., Hamilton, K., Toze, S., & Cook, S. (2019). Seasonal variation of faecal indicator bacteria and pathogens in groundwater: Implications for water safety. Science of the Total Environment, 659, 123–131.

[3] Bain, R., Johnston, R., Mitis, F., Chatterley, C., & Slaymaker, T. (2018). Establishing sustainable development goal baselines for household drinking water, sanitation and hygiene services. Water, 10(12), 1711.

[4] Bartram, J., Corrales, L., Davison, A., Deere, D., Drury, D., Gordon, B., … Stevens, M. (2018). Water safety plan manual: Step-by-step risk management for drinking-water suppliers. World Health Organization.

[5] He, H., & Ma, Y. (2018). Imbalanced learning: Foundations, algorithms, and applications. Wiley-IEEE Press.

[6] Heddam, S., & Kisi, O. (2018). Extreme learning machines for water quality prediction: A case study. Environmental Monitoring and Assessment, 190(8), 1–18.

[7] Howard, G., Bartram, J., Pedley, S., Schmoll, O., Chorus, I., & Berger, P. (2019). Groundwater and public health. In O. Schmoll et al. (Eds.), Protecting groundwater for health (2nd ed., pp. 3–23). IWA Publishing.

[8] Odonkor, S. T., & Ampofo, J. K. (2019). Escherichia coli as an indicator of bacteriological quality of water: An overview. Microbiology Research, 10(1), 1–11.

[9] UNICEF & WHO. (2019). Progress on household drinking water, sanitation and hygiene 2000–2017. World Health Organization.

[10] World Health Organization. (2017). Guidelines for drinking-water quality (4th ed.). WHO Press.

[11] Zhu, S., Heddam, S., & Zhang, J. (2019). Artificial intelligence methods for predicting groundwater quality: A review. Journal of Hydrology, 575, 123–136.