IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 10, October, 2025, pp. 3519-3530
Journal homepage: http://iieta.org/journals/mmep

The Modified Gamma Distribution and Machine Learning for Modeling and Classification of Groundwater Potability

Ahmad Abubakar Suleiman[1*], Mohamed A.F. Elbarkawy[2], Hanita Daud[3], Aliyu Ismail Ishaq[4], Charuai Suwanbamrung[5], Ehab M. Almetwally[6], Mohammed Elgarhy[7,8]

[1] Department of Statistics, Aliko Dangote University of Science and Technology, Wudil 713281, Nigeria
[2] Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[3] Fundamental and Applied Sciences Department, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia
[4] Department of Statistics, Ahmadu Bello University, Zaria 810107, Nigeria
[5] School of Public Health, Walailak University, Nakhon Si Thammarat 80160, Thailand
[6] Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[7] Department of Basic Sciences, Higher Institute of Administrative Sciences, Nasr City 11765, Egypt
[8] Department of Computer Engineering, Biruni University, Istanbul 34010, Turkey

Corresponding Author Email: ahmadabubakar31@gmail.com

https://doi.org/10.18280/mmep.121018

Received: 27 July 2025
Revised: 3 October 2025
Accepted: 11 October 2025
Available online: 31 October 2025

Keywords:
odd beta prime family, Gamma distribution, artificial intelligence, machine learning, groundwater, environmental sustainability, public health

ABSTRACT

Groundwater is vital for public health, industry, and agriculture, and it is found under the surface in soil pores and rock fissures. Accurate modeling and prediction of groundwater parameters are required to ensure effective resource management and environmental sustainability. While the Gamma distribution is commonly used for forecasting groundwater features, it is limited to describing data with a right-skewed shape (where most values are low, but a few are very large). In this paper, we introduce the Odd Beta Prime Gamma (OBP-Gamma) distribution, a flexible statistical model that can describe both left- and right-skewed patterns as well as hazard rates (the probability that failure or contamination occurs at a given time). The OBP-Gamma distribution is applied to two groundwater parameters, pH and conductivity, and compared with classical Gamma and Weibull-Gamma models. Results show that OBP-Gamma provides a better fit for the observed data. In addition, we evaluated the use of machine learning models to classify groundwater potability using a small dataset of 30 water samples collected in Jaen, Kano State, Nigeria. Fourteen models were tested, and Gaussian Naive Bayes achieved the highest classification accuracy (90%), followed by Gradient Boosting (83.3%). Other models, such as Passive Aggressive and AdaBoost, performed poorly, with accuracy below 50%. These results highlight that the OBP-Gamma model offers improved flexibility for groundwater data analysis and that machine learning methods, particularly Gaussian Naive Bayes, show potential for assessing groundwater potability. However, due to the small sample size, the findings should be viewed as a proof-of-concept, with future research needed on larger datasets to confirm generalizability.

1. INTRODUCTION

As a vital resource for domestic, industrial, and agricultural uses, groundwater is essential to maintaining human activity. Because this resource directly affects the health and well-being of dependent populations, its quality must be guaranteed. Considerable studies have been done over the years to monitor and analyze the water quality parameters in various regions. These attempts have produced sufficient information related to water quality characteristics, offering a starting point for additional study and well-informed decision-making [1].

To ensure community safety and well-being, groundwater quality needs to be meticulously assessed and monitored continuously. The health and well-being of the population are directly related to the quality and safety of groundwater supplies. Contaminants that surpass set quality requirements offer considerable dangers, potentially leading to waterborne diseases and other adverse health consequences [2]. Anthropogenic activities such as urbanization and industrialization pose a rising threat to groundwater quality. These variables present considerable obstacles to maintaining groundwater purity, particularly for potable usage. As a result, several studies have been conducted to assess the acceptability of groundwater for human consumption, using a wide range of approaches and analytical techniques. For instance,

multivariate statistics by Thomas [3], the automatic exponential smoothing model by Nsabimana et al. [4], and so on.

Statistical analysis in environmental research has advanced significantly in recent years, especially with the increasing application of probability distribution models Ishaq et al. [5]. When analyzing and interpreting environmental data, these statistical frameworks are essential. Even though using probability models to measure extreme weather and hydrological events is a relatively new idea in environmental studies, many researchers have carefully evaluated several probability distributions to find the one that best fits actual data. For example, an Australian study by Haddad [6] found that the normal and generalized extreme value models performed better than other distributions assessed in the study in terms of representing annual maximum temperatures. Furthermore, probability distribution models have found application in modeling monthly maximum temperatures in Bangladesh by Hossian et al. [7] and Hossain [8], as well as the average daily maximum temperature in regions such as South Africa by Nemukula and Sigauke [9] and Thailand by Busababodhin et al. [10]. Another study by Shakil et al. [11] assessed the suitability of five probability models to find the one that best fit temperature data: Weibull, Gumbel, Cauchy, Logistic, and normal distribution. Three cities' daily extreme temperatures were successfully modeled using a mixed Gaussian model by Al-Hemyari and Abbasi [12]. Five probability distributions were used to examine the temperature by Chen et al. [13]. Furthermore, Poonia and Azad [14] employed the Gamma, Gumbel, log-normal, normal, and Weibull distributions to predict the yearly maximum temperature in India's Northwest Himalayas.

The Gamma distribution is useful in a variety of fields, including finance, environmental research, and engineering, especially when modeling continuous variables with positive skewness. It has strong similarities to the beta distribution and is naturally relevant in situations where the waiting time between Poisson-distributed events is connected. Furthermore, the Gamma distribution has similarities with other well-known distributions such as the normal, exponential, chi-squared, and Erlang distributions. This study aims to enhance the Gamma distribution by creating a more adaptable variation that can accommodate both positive and negative skewness. Such an extension would make a significant contribution to the current research on the Gamma distribution. The cumulative distribution function (CDF) and probability density function (PDF) of the Gamma distribution are described in Eqs. (1) and (2), respectively.

$$G(x) = \frac{\gamma\left(p, \frac{x}{\lambda}\right)}{\Gamma(p)}; x > 0 \tag{1}$$

and

$$g(x) = \frac{x^{p-1}\, exp\left(-\frac{x}{\lambda}\right)}{\lambda^p \Gamma(p)}; x > 0 \tag{2}$$

where, λ, $p > 0$, λ is the scale parameter, p is the shape parameter, and $\Gamma(p)$ is the Gamma function, which has the following formula by Wadi et al. [15]:

$$\Gamma(p) = \int_0^\infty t^{p-1}\, exp(-t)\, dt \tag{3}$$

Classical Gamma distributions have limitations in their capacity for effectively analyzing datasets with left-skewness. This limitation necessitates the creation of an improved Gamma distribution that is more flexible and adaptable in capturing the distinct characteristics of left-skewed data, which are common in a variety of domains, including environmental science, reliability engineering, medicine, finance, insurance, and engineering [16, 17].

In recent years, there has been an increased interest in creating flexible probability distributions by expanding current models with extra shape factors [18]. This method has proven extremely useful in capturing the many different aspects of real-world data, such as skewness and changing tail behavior. A wide range of generalized families of univariate probability distributions have emerged in the literature, including noteworthy examples such as the Beta-G family by Eugene et al. [19], the new odd reparameterized exponential transformed-X family by Orji et al. [20], the Transmuted-G family by Shaw and Buckley [21], and various extensions of the Kumaraswamy distribution, such as the Kumaraswamy-Pareto by Bourguignon et al. [22], the exponentiated Kumaraswamy distribution by Lemonte et al. [23], the Kumaraswamy Marshal-Olkin family by Alizadeh et al. [24], the Kumaraswamy Marshall-Olkin Fréchet by Afify et al. [25], the Kumaraswamy power function by Abdul-Moniem [26], the exponentiated Kumaraswamy-power function by Bursa and Ozel [27], the updated Lindley by Onyekwere et al. [28], the exponentiated Kumaraswamy-G class by Gomes-Silva et al. [29], the Kumaraswamy inverted Topp–Leone by Hassan et al. [30], the modified exponentiated Kumaraswamy by Arshad et al. [31], entropy analysis of the Kumaraswamy distribution by Al-Babtain et al. [32], the extended generalized inverted Kumaraswamy-G by Ramzan et al. [33], and many others. These generalized families provide a rich framework for modeling complex data in diverse fields.

Suleiman et al. [34] presented the Odd Beta Prime Generalized (OBP-G) class of distributions, which is an extension of the widely used beta prime distribution. The beta prime distribution has proven useful for modeling lifetime data in a variety of fields, including biomedical science and engineering. Despite its importance, the beta prime distribution has gotten little attention in the literature. The OBP-G class provides a fresh framework for analyzing data with properties that may not be fully reflected by the usual beta prime distribution. Eq. (4) defines the CDF for the OBP-G class.

$$F(x) = \frac{B_{\frac{G(x,\varepsilon)}{1-G(x,\varepsilon)}}(c,d)}{B(c,d)}; x > 0, c, d > 0 \tag{4}$$

where, c, d are shape parameters, ε is a vector parameter. The corresponding PDF is expressed in Eq. (5) as follows:

$$f(x) = \frac{g(x,\varepsilon)}{B(c,d)\{1-G(x,\varepsilon)\}^2} \frac{\left\{\frac{G(x,\varepsilon)}{1-G(x,\varepsilon)}\right\}^{c-1}}{\left\{1+\left(\frac{G(x,\varepsilon)}{1-G(x,\varepsilon)}\right)\right\}^{c+d}}$$
$$x > 0, c, d > 0 \tag{5}$$
$$f(x) = \frac{g(x,\varepsilon)G(x,\varepsilon)^{c-1}}{B(c,d)\{1-G(x,\varepsilon)\}^{1-d}}$$

where, $g(x,\varepsilon)$ is the baseline PDF.

The OBP-G class has proved its adaptability by generalizing

different baseline models, resulting in the creation of innovative and versatile compound distributions with broad application across multiple domains. This expansion has resulted in a diverse set of distributions, including the OBP-logistic model developed by Suleiman et al. [34], which effectively simulates environmental and engineering data. Building on this achievement, Suleiman et al. [35] proposed the OBP-inverted Kumaraswamy distribution and demonstrated its usefulness in medical data analysis. Continuing this line of research, Suleiman et al. [36] defined the OBP-Burr X distribution, which has applications in geological and COVID-19 data. Following these advances, this study presents a four-parameter Gamma distribution derived from the OBP-G class for analyzing groundwater datasets. However, previous extensions of the Gamma distribution have improved flexibility but still face limitations when applied to datasets with diverse skewness and hazard rate behaviors. A more adaptable distribution is therefore needed.

In recent years, machine learning models have emerged as useful tools for classification problems in a variety of fields. These models have the potential to improve and supplement traditional methods, resulting in more precise and effective assessments. Several studies have investigated the application of various machine learning models in the context of water quality classification. Readers may refer to the references [37-40] for more details. While some studies have applied single classifiers such as Support Vector Machines or Decision Trees, few have conducted systematic comparisons across multiple machine learning approaches. This limits our understanding of which algorithms perform best under small-sample groundwater datasets, where generalization is difficult.

This study addresses these two gaps. First, we introduce the OBP-Gamma distribution, a new extension that provides greater flexibility in modeling both left- and right-skewed data and diverse hazard rate structures. Second, we compare the performance of multiple machine learning classifiers for groundwater potability prediction using a small dataset of 30 samples. Unlike earlier studies, our approach combines probability distribution with comparative machine learning analysis. This approach offers a more comprehensive framework for groundwater modeling.

The primary motivations for this investigation are outlined below:

(a) Development of the OBP-Gamma distribution as a flexible alternative to classical Gamma and related models.
(b) Application of the OBP-Gamma distribution to real groundwater data (pH and conductivity).
(c) Comparative evaluation of 14 machine learning models for groundwater potability classification.

2. METHODOLOGY

The OBP-Gamma distribution is described in this section, along with its basic ideas, significant applications, and classification methods for groundwater potability using machine learning models.

2.1 The OBP-Gamma distribution

An expansion of the Gamma distribution, the OBP-Gamma distribution introduces two additional shape parameters, *c* and d, from the OBP-G family. Eq. (1) is substituted into Eq. (4) to obtain the CDF of the OBP-Gamma distribution in the manner described below:

$$F(x) = \frac{B\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}(c,d)}{B(c,d)\left(1-\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right)} \qquad x>0, c>0, d>0, p>0, \lambda>0 \tag{6}$$

where, λ is scale parameter and *c*, *d*, *p* are shape parameters. Inserting Eq. (2) into Eq. (5) yields the PDF as follows:

$$f(x) = \frac{x^{p-1}\exp\left(-\frac{x}{\lambda}\right)\left(\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right)^{c-1}}{B(c,d)\lambda^{p}\Gamma(p)\left\{1-\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right\}^{1-d}}; x>0 \tag{7}$$

Some of the potential forms of the OBP-Gamma distribution PDF for specific values of the parameters *c*, *d*, *p* and λ, respectively, are shown in Figure 1.

For reproducibility, the general definition of the OBP–G family is provided in Eqs. (4) and (5), which specify the generator function and general probability density form. These expressions serve as the foundation for the OBP–Gamma model developed in this study. This ensures that the method can be replicated and extended by other researchers.

The survival function for the OBP-Gamma distribution is as follows:

$$S(x) = 1 - F(x) \tag{8}$$

Substituting Eq. (6) into Eq. (8) yields the following survival function for the OBP-Gamma model:

$$S(x) = 1 - \frac{B\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}(c,d)}{B(c,d)\left(1-\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right)}; x>0 \tag{9}$$

The OBP-Gamma distribution's hazard function is as follows:

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1-F(x)} \tag{10}$$

The hazard function of the suggested distribution can be obtained by replacing Eqs. (6) and (7) in Eq. (10):

$$h(x) = \frac{x^{p-1}\exp\left(-\frac{x}{\lambda}\right)\left(\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right)^{c-1}}{\left(1-B\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}(c,d)\right)\left(B(c,d)\lambda^{p}\Gamma(p)\left\{1-\dfrac{\gamma\left(p,\frac{x}{\lambda}\right)}{\Gamma(p)}\right\}^{1-d}\right)}; x>0. \tag{11}$$

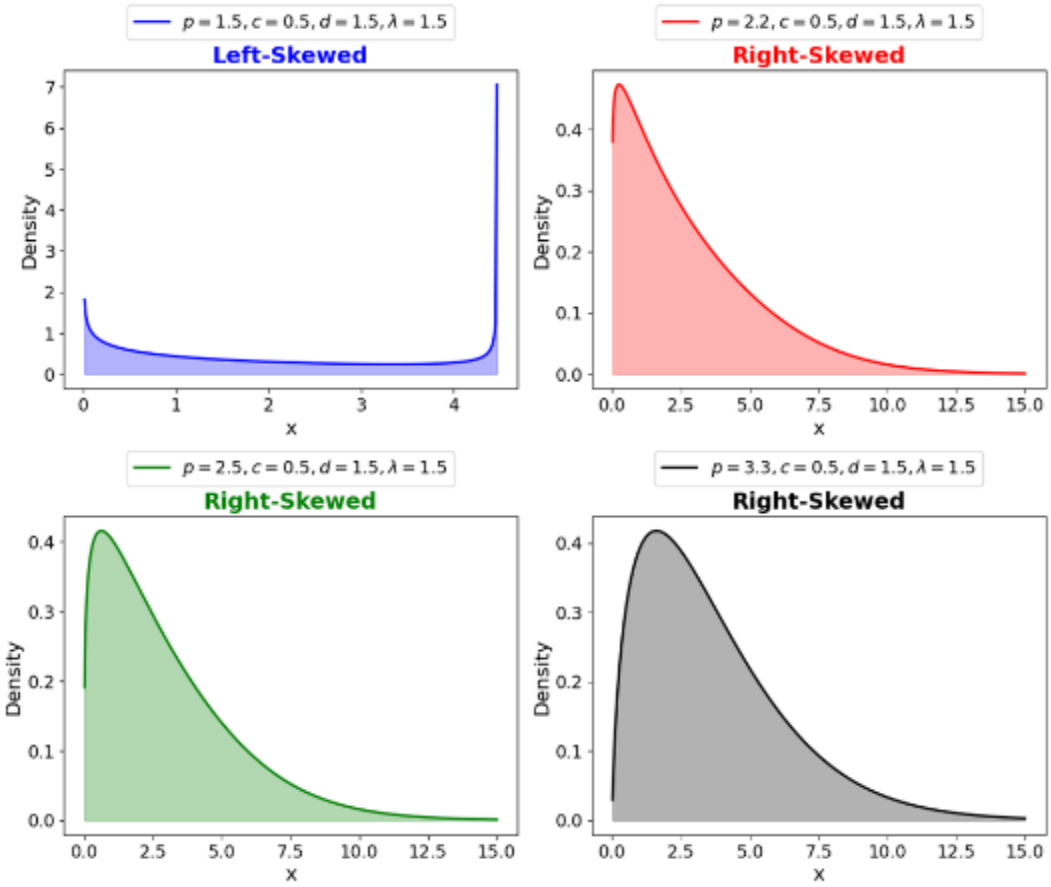

Figure 1. PDF of the OBP-Gamma model

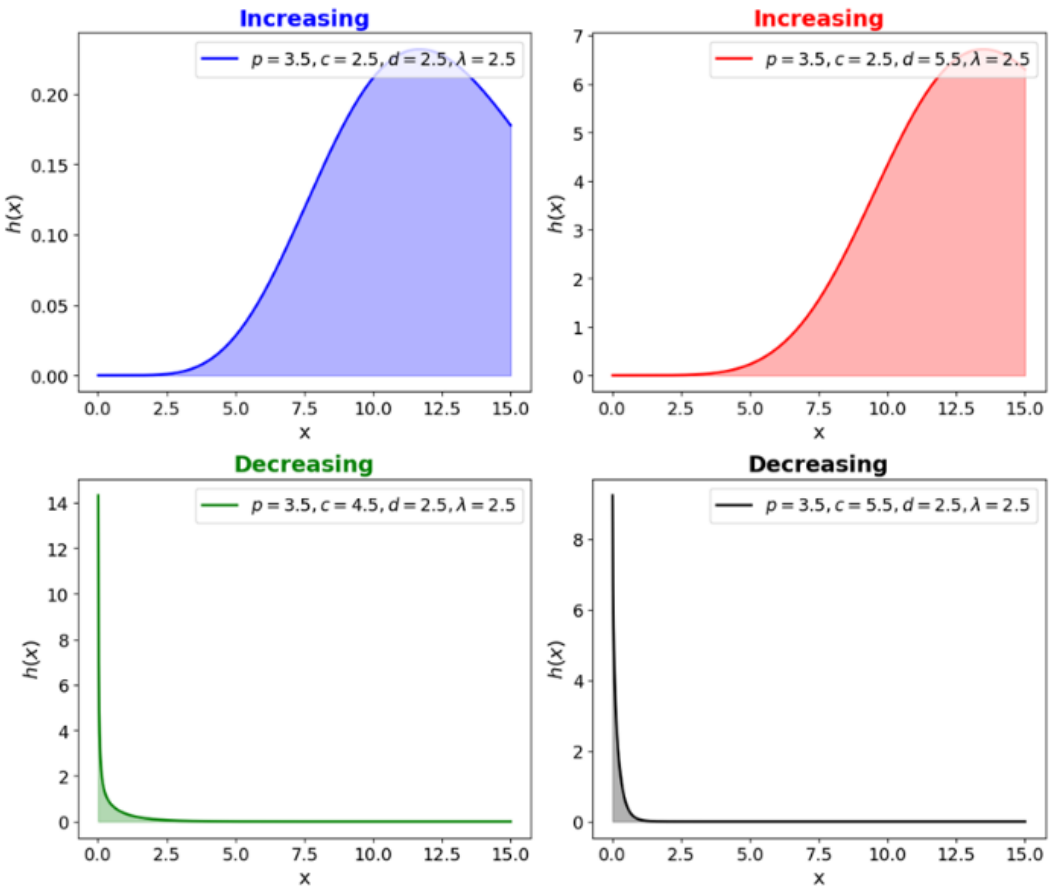

Figure 2. Hazard function of the OBP-Gamma distribution

For certain values of the parameters c, d, p and λ, respectively, Figure 2 shows several potential forms of the hazard functions of the OBP-Gamma distribution.

The proposed OBP-Gamma distribution is a fundamental novelty in this research. As shown in Figure 1, this distribution is extremely flexible, capturing both positive and negative skewness, which the typical Gamma distribution lacks. Figure 2 further demonstrates the dynamic nature of the OBP-Gamma distribution's hazard function, which can show either increasing or decreasing patterns. These curvature qualities improve the model's applicability in a variety of statistical circumstances. The OBP-Gamma distribution's practical significance is demonstrated by its successful application to real-world datasets from a variety of fields, as shown in the application section. This distribution has a potential application in various domains, including biomedical sciences, engineering, physics, reliability, economics, environmental sciences, biology, and survival analysis.

2.1.1 Application of the OBP-Gamma distribution to groundwater datasets

This section describes the practical implementation of the newly suggested OBP-Gamma distribution. We analyze its performance using two real-world groundwater datasets on pH and conductivity. To determine its acceptability, we make a comparison with different distributions. This comparison includes the traditional Gamma distribution and the Weibull-Gamma distribution by Klakattawi [41].

To find the most suitable model, we used R statistical software to evaluate some important metrics. These metrics included the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the negative log-likelihood $(-\ell)$. These metrics provide insights into the model's goodness of fit. These measures provide a rigorous statistical basis for model comparison and were consistently applied across the Gamma, Weibull–Gamma, and OBP–Gamma models. The model with the smallest values for these metrics is selected as the best model.

Dataset I, sourced from Suleiman et al. [42], provides information on the pH concentration in groundwater. The dataset is given below:

6.27	6.14	5.59	6.07	5.76	6.97
6.65	6.50	7.17	6.97	6.99	6.90
6.92	6.82	6.35	7.12	6.85	6.71
6.85	7.22	6.66	7.07	6.14	6.41
6.47	6.32	6.62	7.20	6.76	7.04

Dataset II, obtained from Suleiman et al. [42], includes information on the conductivity concentration in groundwater. This dataset is given below:

654	514	626	796	902	209
796	780	234	812	998	422
394	542	594	422	326	446
452	562	854	306	1003	356
922	1490	1210	1030	790	740

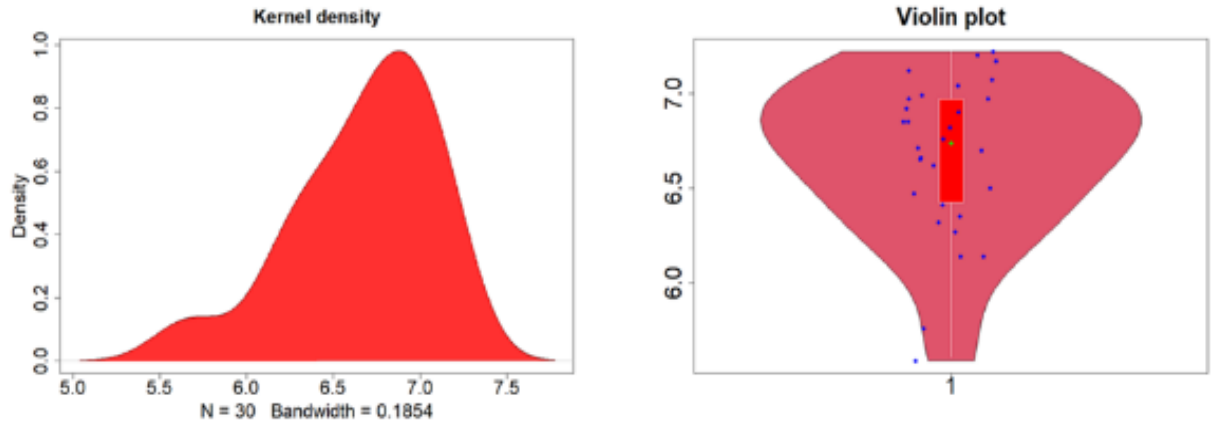

Figure 3. Descriptive plots of dataset I

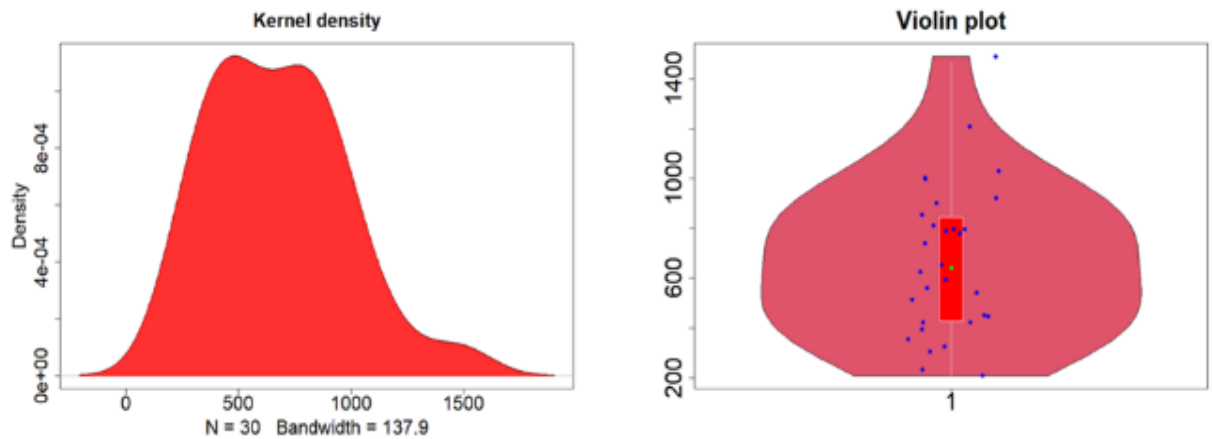

Figure 4. Descriptive plots of dataset II

Table 1. Maximum likelihood estimates (MLEs) and goodness-of-fit measures for the first dataset

Models	Estimate	$-\ell$	AIC	BIC
OBP-Gamma	$\hat{\lambda} = 0.1054$ $\hat{p} = 83.6986$ $\hat{c} = 0.8394$ $\hat{d} = 59.3355$	-13.5052	35.0104	40.6152
Weibull-Gamma	$\hat{\lambda} = 0.5569$ $\hat{p} = 8.3549$ $\hat{b} = 22.8730$ $\hat{a} = 3.3212$	-224.2258	456.4516	462.0564
Gamma	$\hat{\lambda} = 1.6678$ $\hat{p} = 7.1523$	-37.4451	82.8903	88.4951

Table 2. MLEs and goodness-of-fit measures for the second dataset

Models	Estimate	$-\ell$	AIC	BIC
OBP-Gamma	$\hat{\lambda} = 0.1214$ $\hat{p} = 72.0785$ $\hat{c} = 0.6521$ $\hat{d} = 46.4180$	-19.6125	47.2251	52.8299
Weibull- Gamma	$\hat{\lambda} = 0.82726$ $\hat{p} = 10.6411$ $\hat{b} = 3.4379$ $\hat{a} = 0.2426$	-19.6352	47.2705	52.8753
Gamma	$\hat{\lambda} = 1.6015$ $\hat{p} = 6.8813$	-40.0649	84.1299	86.9323

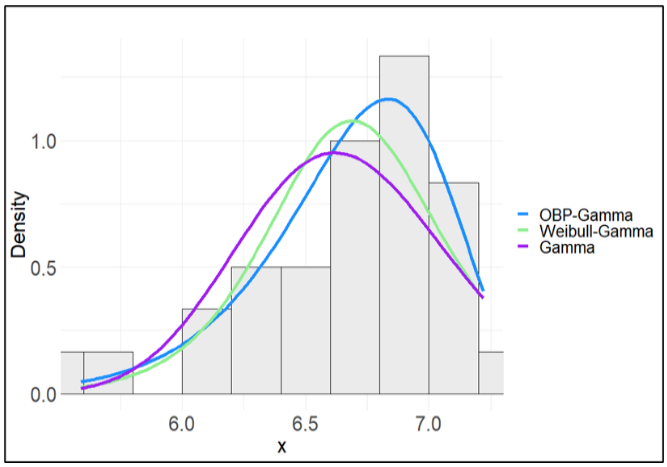

Figure 5. Empirical PDFs of dataset I

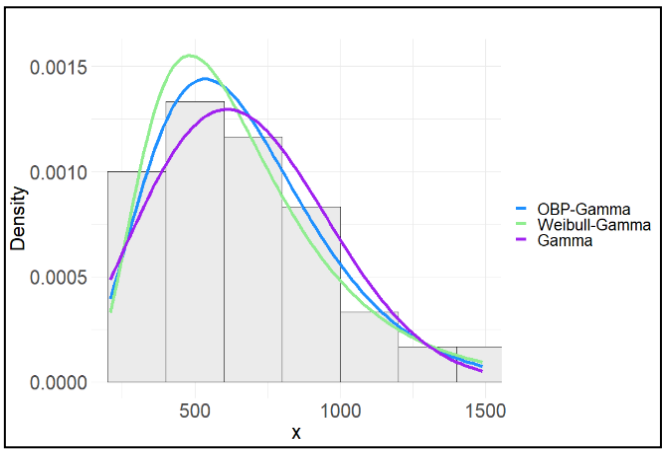

Figure 6. Empirical PDFs of dataset II

Figures 3 and 4 show visual representations of datasets I and II using kernel density and violin plots. The kernel density corresponds to the dataset I shows a left-skewed distribution, whereas the violin plot indicates the existence of outliers. Similarly, dataset II has a right-skewed distribution based on the kernel density plot, with the violin plot confirming the presence of extreme values. Figure 1 shows how the suggested OBP-Gamma distribution may adequately capture both right and left-skewed features in the datasets.

Tables 1 and 2 show the comparison criteria for datasets I and II, respectively. The results show that the OBP-Gamma model outperforms the Gamma and Weibull-Gamma models. This superiority is clear across both datasets, as the OBP-Gamma model consistently has lower information criterion values.

Figures 5 and 6 demonstrate how well each model fits the data. The findings show that the OBP-Gamma model outperforms the Weibull-Gamma and Gamma models in capturing the patterns in both datasets compared to the Weibull-Gamma and Gamma models.

2.2 Machine learning models

The previous subsection evaluated the appropriateness of the OBP-Gamma model for groundwater parameter datasets; this subsection focuses on classification algorithms. A wide range of machine learning models, including Random Forest Classifier (RF), AdaBoost Classifier (ADA), Gradient Boosting Classifier (XGB), K-Nearest Neighbours (KNN), Decision Tree Classifier (DTC), Gaussian Naive Bayes (GNB), Stochastic Gradient Descent Classifier (SGDC), Perceptron (Perc), Nearest Centroid (NC), Ridge Classifier (Ridge), Bernoulli Naive Bayes (BNB), Gradient Boosting Classifier (XGB), and Passive Aggressive Classifier (PAC), are tested for their ability to classify groundwater potability.

All nine groundwater features available in the dataset were included in the analysis. Before modeling, the data were cleaned and standardized; however, no advanced feature selection or outlier removal procedure was applied. While conductivity displayed some extreme values, these were retained to reflect the real distribution of the data and to avoid introducing bias in such a small dataset. All computational operations, including training and testing, are carried out using the Python programming environment. Figure 7 illustrates the methodology used in this subsection.

2.2.1 Water sampling and preprocessing of data

This study examined 30 groundwater samples obtained from open wells and boreholes in Jaen, Kano State, Nigeria, in August 2020. Each sample had fifteen physicochemical parameters tested, with a focus on nine essential properties such as conductivity, pH, and sulphate concentration. Each sample was labeled as "potable" (1) or "non-potable" (0) based on Nigerian drinking water quality standards outlined in NSDWQ [43].

The preprocessing procedures listed below were carried out in order to get the data ready for analysis:

i. Data cleaning: Removed inaccurate and missing data points from the dataset.
ii. Feature scaling: To ensure equal contribution to model training, continuous variables were scaled to a mean of 0 and a standard deviation of 1.
iii. Data split: An 80-20 split between training and testing sets ensures consistent results across models.

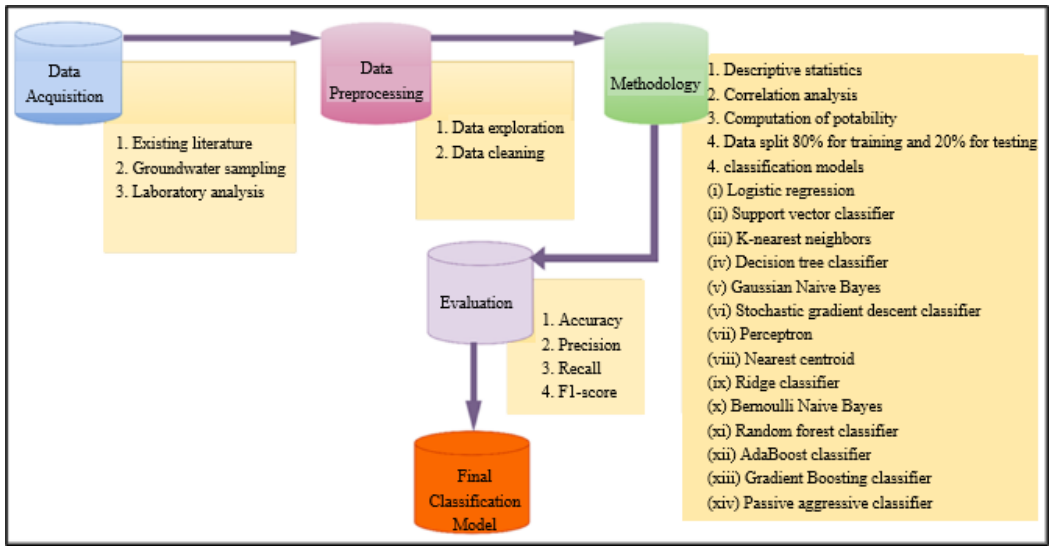

Figure 7. Methodology of the machine learning classification model

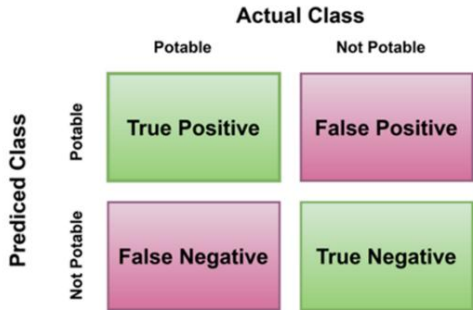

Figure 8. Confusion matrix for binary classification for groundwater potability

2.2.2 Model selection and training

The comparison analysis included fourteen distinct machine learning models: LR, SVC, KNN, DTC, GNB, SGDC, Perc, NC, Ridge, BNB, RF, ADA, XGB, and PAC algorithms. The following studies [44-47] provide a comprehensive review of these algorithms.

The following steps were taken for each model:

1. Model training: The model was trained with the training dataset and default hyperparameters.

2. Model evaluation: Metrics such as accuracy, precision, recall, and F-1 score were used to assess the model's performance on the test dataset.

2.2.3 Evaluation metrics

This study employs a binary confusion matrix to evaluate the efficacy of water quality classification systems. The groundwater is classified as "potable" or "not potable." A confusion matrix is a useful tool for evaluating model predictions by comparing them to actual class labels, which provides information about the model's accuracy in identifying various water quality classes.

Figure 8 shows the concept of a confusion matrix. True Positive (TP) refers to the correct identification of samples classified as "potable". True Negative (TN) indicates that samples are correctly classified as "not potable". False Positives (FP) occur when a sample from the "not potable" class is incorrectly identified as "potable." Similarly, False Negative (FN) refers to the incorrect classification of samples from the "potable" class as "not potable." These metrics are essential in assessing a classification model's ability to distinguish between potable and non-potable water sources.

To evaluate model performance, we use a set of common metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score, which are calculated using the values for TP, TN, FP, and FN as outlined in Eqs. (12)-(15), respectively.

$$Accuracy\ Score = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

3. RESULTS

3.1 Descriptive statistics

Table 3 presents the statistical features of the groundwater dataset. The results show significant variability in important parameters, as shown by large standard deviations. Groundwater conductivity varied greatly, with values ranging from 209 to 1490 units. The slightly acidic nature of the water is reflected in the pH range of 5.59 to 7.22. The skewness and kurtosis values indicate that the distribution of these properties is not perfectly symmetrical and may have longer tails than a normal distribution. These measures of variation play a crucial role in evaluating groundwater quality assessment and monitoring strategies.

Figure 9 depicts the spatial distribution and trends in groundwater parameter concentrations at the sampling stations. The visualization effectively displays how these concentrations fluctuate spatially, including gradients that may indicate directional trends. This geographical representation improves our understanding of the groundwater system by assisting in the identification of regions with potentially higher parameter concentrations and guiding future research as needed.

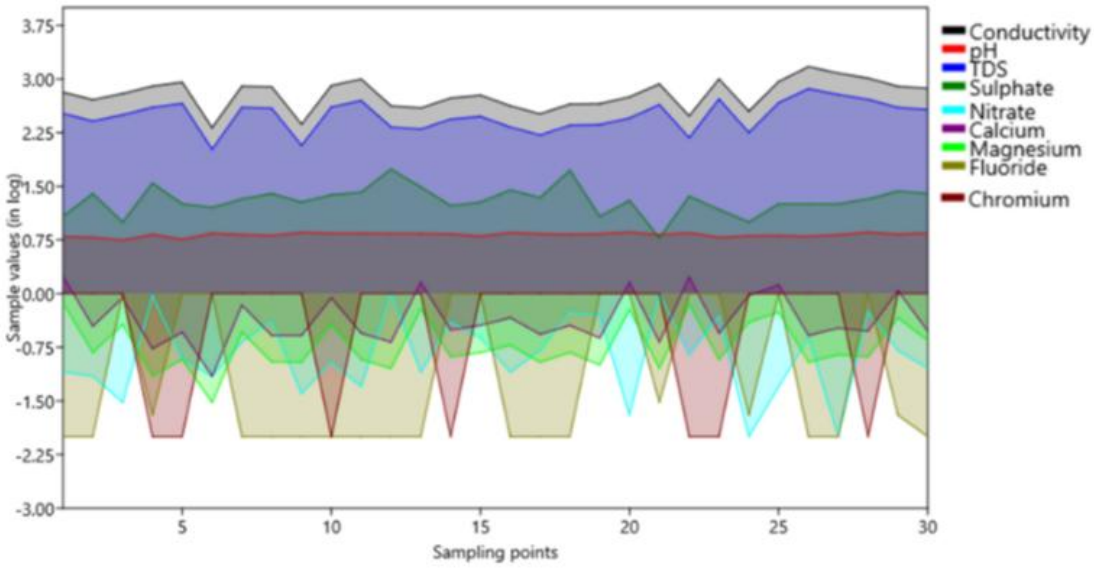

Figure 9. Graph of sampling points showing trend values

Table 3. Descriptive statistics for the groundwater dataset

Parameters	Min.	Max.	Std. Dev.	Skewness	Kurtosis
Conductivity	209	1490	302.61	0.63	0.32
Ph	5.59	7.22	0.41	-0.92	0.55
TDS	104	731	151.26	0.52	0.04
Sulphate	6.00	55.00	10.75	1.66	3.70
Nitrate	0.01	0.93	0.22	1.68	2.84
Calcium	0.07	1.75	0.50	1.21	0.14
Magnesium	0.03	0.74	0.21	1.20	0.14
Fluoride	0.00	0.03	0.01	0.87	0.98
Chromium	0.00	0.01	0.01	1.33	-0.26

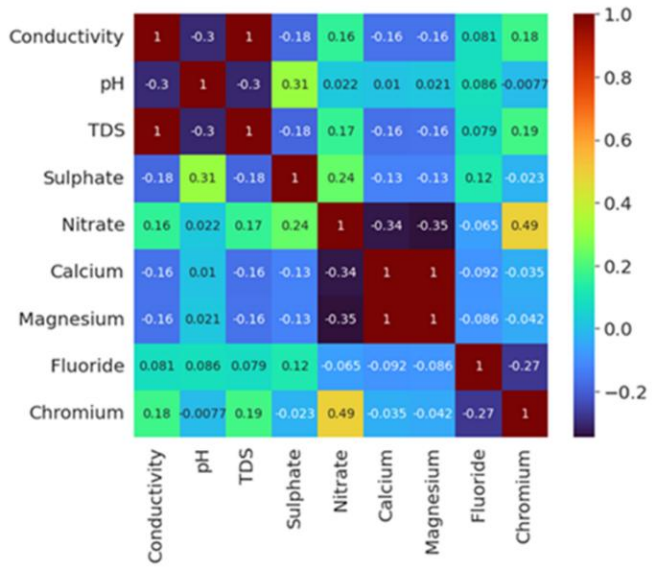

Figure 10. Groundwater parameter correlation matrix

3.2 Measure of correlation

Correlation analysis is a useful approach for determining the correlations between different water quality measures. By investigating these relationships, we can obtain insight into the complex mechanisms that influence groundwater quality [45]. Figure 10 depicts the correlation analysis results, which demonstrate strong positive and negative connections between the analysed parameters. In particular, a strong correlation of 0.99 was found between Calcium and Magnesium, demonstrating a near-linear relationship and that variations in their amounts are closely related in groundwater samples. Similarly, a considerable positive correlation (0.99) was reported between conductivity and TDS, indicating a close relationship. These findings help us understand how groundwater quality metrics are interrelated.

3.3 Evaluation of model performance

To measure the performance of each machine learning algorithm, we used a set of key evaluation metrics, such as accuracy, precision, recall, and F1-score, all reported as percentages. This comprehensive method gave a holistic picture of each model's performance in classifying the groundwater dataset. Table 4 compares these models, demonstrating that GNB achieved the highest accuracy of 90%, while ADA had a lower accuracy of 33.33%. This investigation demonstrates GNB's higher effectiveness in classifying groundwater potability when compared to other assessed approaches.

Table 4. Comparative analysis of various model-algorithm accuracy

S/No.	Algorithm	Accuracy	Precision	Recall	F1-score
1	GNB	0.9000	0.9000	0.9000	0.8889
2	XGB	0.8333	0.6750	0.6750	0.6667
3	NC	0.7750	0.7750	0.7750	0.7750
4	BNB	0.7750	0.7750	0.7750	0.7750
5	SVC	0.7500	0.7500	0.6000	0.5000
6	KNN	0.7500	0.7500	0.6000	0.5000
7	DTC	0.7500	0.7500	0.6000	0.5000
8	RF	0.7500	0.5833	0.5750	0.5555
9	SGDC	0.6750	0.5833	0.5750	0.5500
10	Perc	0.6750	0.6750	0.6750	0.6667
11	LR	0.5833	0.5833	0.5750	0.5500
12	Ridge	0.4643	0.4642	0.4750	0.4156
13	PAC	0.4643	0.4642	0.4750	0.4756
14	ADA	0.3333	0.3333	0.3500	0.3250

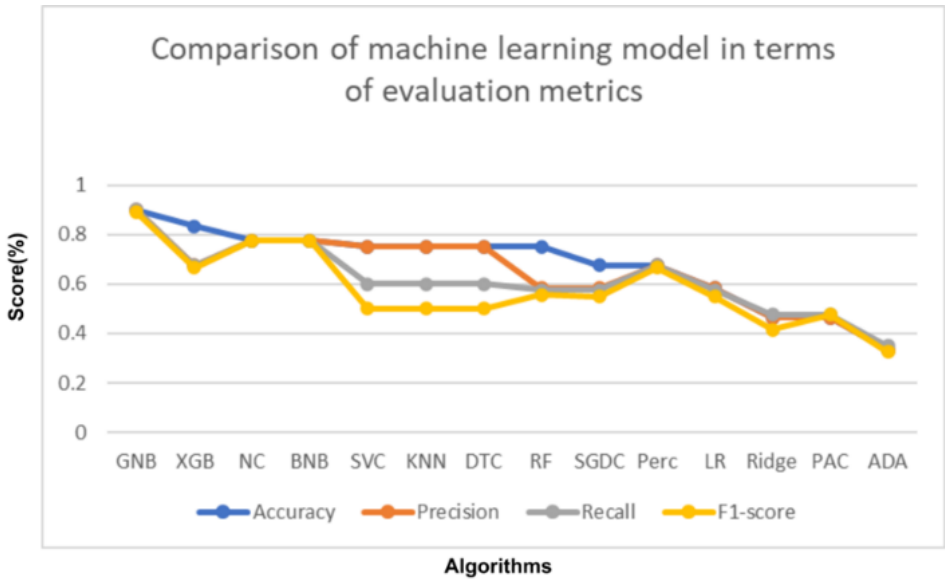

Figure 11. Comparison of machine learning models in terms of evaluation metrics

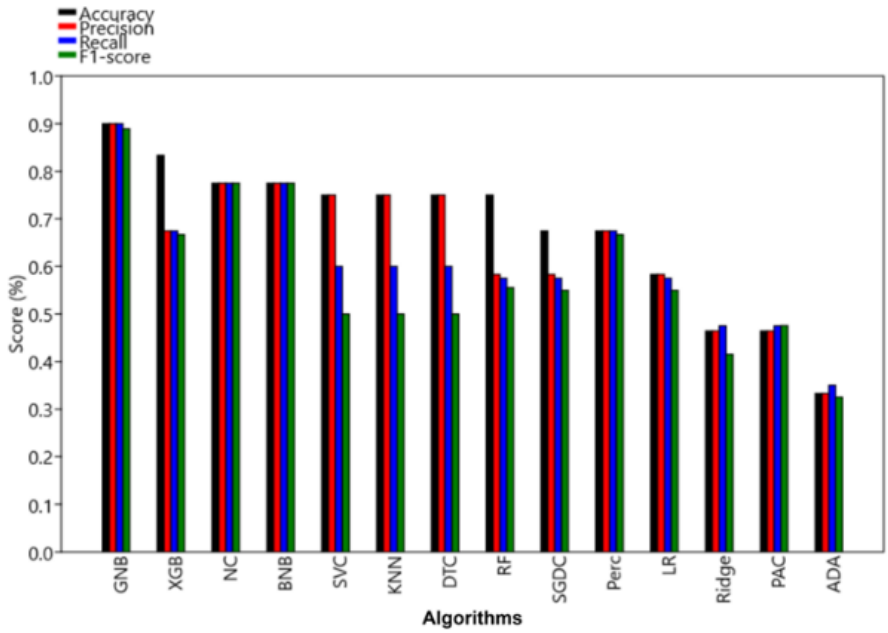

Figure 12. Comparison of machine learning models in terms of evaluation metrics

Figures 11 and 12 present a visual comparison of the accuracy of several model algorithms. These numbers enable a rapid and easy evaluation of model performance, showing the algorithms with the best accuracy scores.

The results reported above indicate the effectiveness of the GNB classifier in predicting groundwater potability. The GNB performed well, especially when correlations across groundwater quality measurements were modest. These findings indicate that GNB is a realistic and competitive approach for real-world groundwater potability assessments.

It is important to note that the groundwater dataset analyzed here is relatively small (n = 30), which limits the robustness and generalizability of the statistical and machine learning results. Moreover, A key limitation of the machine learning component of this study is the very small dataset size. With an 80–20 train-test split, only six samples were used for testing, which limits the statistical reliability of performance estimates and may lead to overfitting. A more robust approach would have been to use k-fold cross-validation or leave-one-out cross-validation; however, this was not implemented here due to dataset size constraints.

Another limitation of the present study is that confidence intervals, hypothesis testing, or resampling approaches such as bootstrapping were not applied to formally quantify the uncertainty of the model comparisons. Future work should incorporate these statistical tools to more rigorously validate differences between models.

4. CONCLUSIONS

This study introduced the OBP–Gamma distribution as a flexible extension of the classical Gamma model for groundwater data analysis. The OBP–Gamma distribution was shown to capture a wider range of shapes and hazard behaviors compared with the Gamma and Weibull–Gamma

distributions, providing improved model fit to the observed groundwater parameters. In parallel, we conducted a comparative evaluation of machine learning models for groundwater potability classification.

The findings suggest that the OBP–Gamma distribution offers enhanced flexibility and that Gaussian Naive Bayes performed best among the machine learning classifiers tested. However, given the small dataset size (n = 30) and the reliance on simple train-test splits.

Future studies will focus on refining the OBP-Gamma distribution's mathematical features and examining its relevance in numerous medical fields, including COVID-19 and cancer data analysis. In addition, further research should focus on improving classification accuracy through model refining and ensemble approaches in this vital subject. Furthermore, the OBP–Gamma distribution was compared with the classical Gamma and Weibull–Gamma models, as these are among the most widely used baselines in groundwater analysis and provide a direct benchmark for evaluating improvements. We acknowledge, however, that this represents a limited set of competing models. Future studies will extend the comparison to more recent and flexible distributions such as the generalized Gamma and Log-logistic, as well as apply the OBP–Gamma model to larger datasets to further validate its performance.

REFERENCES

[1] Yusri, H.I.H., Ab Rahim, A.Z.A., Hassan, S.L.M., Halim, I.S.A., Abdullah, N.E. (2022). Water quality classification using SVM and XGBoost method. In 2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, pp. 231-236. https://doi.org/10.1109/ICSGRC55096.2022.9845143

[2] Suleiman, A.A., Abdullahi, U.A., Suleiman, A., Suleiman, S.A., Abubakar, H.U. (2022). Correlation and regression model for physicochemical quality of groundwater in the Jaen district of Kano State, Nigeria. Journal of Statistical Modeling and Analytics, 4(1). https://doi.org/10.22452/josma.vol4no1.2

[3] Thomas, E.O. (2023). Evaluation of groundwater quality using multivariate, parametric and non-parametric statistics, and GWQI in Ibadan, Nigeria. Water Science, 37(1): 117-130. https://doi.org/10.1080/23570008.2023.2221493

[4] Nsabimana, A., Wu, J., Wu, J., Xu, F. (2023). Forecasting groundwater quality using automatic exponential smoothing model (AESM) in Xianyang City, China. Human and Ecological Risk Assessment: An International Journal, 29(2): 347-368. https://doi.org/10.1080/10807039.2022.2087176

[5] Ishaq, A.I., Usman, A., Tasi'u, M., Suleiman, A.A., Ahmad, A.G. (2022). A new odd F-Weibull distribution: Properties and application of the monthly Nigerian naira to British pound exchange rate data. In 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, pp. 326-332. https://doi.org/10.1109/ICDABI56818.2022.10041527

[6] Haddad, K. (2021). Selection of the best fit probability distributions for temperature data and the use of L-moment ratio diagram method: A case study for NSW in Australia. Theoretical & Applied Climatology, 143(3-4): 1261-1284. https://doi.org/10.1007/s00704-020-03455-2

[7] Hossian, M.M., Abdulla, F., Rahman, M.H. (2016). Selecting the probability distribution of monthly maximum temperature of Dhaka (capital city) in Bangladesh. Jahangirnagar University Journal of Statistical Studies, 33: 33-45.

[8] Hossain, M. (2018). Fitting the probability distribution of monthly maximum temperature of some selected stations from the northern part of Bangladesh. International Journal of Ecological Economics & Statistics, 39(4): 80-91.

[9] Nemukula, M.M., Sigauke, C. (2018). Modelling average maximum daily temperature using r largest order statistics: An application to South African data. Jàmbá: Journal of Disaster Risk Studies, 10(1): 1-11.

[10] Busababodhin, P., Chiangpradit, M., Papukdee, N., Ruechairam, J., Ruanthaisong, K., Guayjarempanishk, P. (2021). Extreme value modeling of daily maximum temperature with the r-largest order statistics. Journal of Applied Science and Emerging Technology, 20(1): 28-38.

[11] Shakil, M., Sirajo, M., Khadim, A., Aliyu, M.A., Singh, J.N., Kibria, B.M.G. (2022). Probability modeling of lifetime and temperature data of the black holes existing in X-ray binaries. World Scientific News, 173: 78-93.

[12] Al-Hemyari, Z.A., Abbasi, J.N A. (2023). Modelling and analysing the daily temperature of several cities using mixture Gaussian distributions. International Journal of Computing Science and Mathematics, 17(4): 320-341. https://doi.org/10.1504/IJCSM.2023.131630

[13] Chen, D., Chen, X., Wang, J., Zhang, Z., Wang, Y., Jia, C., Hu, X. (2021). Estimation of thermal time model parameters for seed germination in 15 species: The importance of distribution function. Seed Science Research, 31(2): 83-90. https://doi.org/10.1017/S0960258521000040

[14] Poonia, N., Azad, S. (2022). Projection of annual maximum temperature over Northwest Himalayas using probability distribution models. Theoretical and Applied Climatology, 149(3): 1599-1627. https://doi.org/10.1007/s00704-022-04121-5

[15] Wadi, M., Elmasry, W., Tamyigit, F.A. (2023). Important considerations while evaluating wind energy potential. Journal of the Faculty of Engineering and Architecture of Gazi University, 38(2): 947-962. https://doi.org/10.17341/gazimmfd.1066351

[16] Panitanarak, U., Ishaq, A.I., Usman, A., Sadiq, I.A., Mohammed, A.S. (2025). The modified sine distribution and machine learning models for enhancing crude oil production prediction. Journal of Statistical Sciences and Computational Intelligence, 1(1): 29-45. https://doi.org/10.64497/jssci.3

[17] Ode, O., Musa, T., Usman, A., Sadiq, I.A. (2025). A novel odd Rayleigh-exponential distribution (OR-ED) and its application to lifetime datasets. Journal of Statistical Sciences and Computational Intelligence, 1(3): 262-282. https://doi.org/10.64497/jssci.91

[18] Suleiman, A.A., Daud, H., Usman, A.G., Abba, S.I., Othman, M., Elgarhy, M. (2025). A new two-parameter half-logistic distribution with numerical analysis and applications. Journal of Statistical Sciences and Computational Intelligence, 1(1): 1-28. https://doi.org/10.64497/jssci.2

[19] Eugene, N., Lee, C., Famoye, F. (2002). Beta-normal

distribution and its applications. Communications in Statistics-Theory and Methods, 31(4): 497-512. https://doi.org/10.1081/STA-120003130

[20] Orji, G.O., Etaga, H.O., Almetwally, E.M., Igbokwe, C.P., Aguwa, O.C., Obulezi, O.J. (2025). A new odd reparameterized exponential transformed-x family of distributions with applications to public health data. Innovation in Statistics and Probability, 1(1): 88-118. https://doi.org/10.64389/isp.2025.01107

[21] Shaw, W.T., Buckley, I.R. (2009). The alchemy of probability distributions: Beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map. arXiv preprint arXiv:0901.0434. https://doi.org/10.48550/arXiv.0901.0434

[22] Bourguignon, M., Silva, R. B., Zea, L.M., Cordeiro, G.M. (2013). The kumaraswamy Pareto distribution. Journal of statistical theory and applications, 12(2): 129-144. https://doi.org/10.2991/jsta.2013.12.2.1

[23] Lemonte, A.J., Barreto-Souza, W., Cordeiro, G.M. (2013). The exponentiated Kumaraswamy distribution and its log-transform. Brazilian Journal of Probability and Statistics, 27(1): 31-53. https://doi.org/10.1214/11-BJPS149

[24] Alizadeh, M., Tahir, M.H., Cordeiro, G.M., Mansoor, M., Zubair, M., Hamedani, G. (2015). The Kumaraswamy marshal-Olkin family of distributions. Journal of the Egyptian Mathematical Society, 23(3): 546-557. https://doi.org/10.1016/j.joems.2014.12.002

[25] Afify, A.Z., Yousof, H. M., Cordeiro, G.M., Ahmad, M. (2016). The Kumaraswamy Marshall-Olkin Fréchet distribution with applications. Journal of ISOSS, 2(1): 41-58.

[26] Abdul-Moniem, I.B. (2017). The Kumaraswamy power function distribution. Journal of Statistics Applications & Probability, 6(1): 81-90. http://doi.org/10.18576/jsap/060107

[27] Bursa, N., Ozel, G. (2017). The exponentiated Kumaraswamy-power function distribution. Hacettepe Journal of Mathematics and Statistics, 46(2): 277-292.

[28] Onyekwere, C.K., Aguwa, O.C., Obulezi, O.J. (2025). An updated Lindley distribution: Properties, estimation, acceptance sampling, actuarial risk assessment and applications. Innovation in Statistics and Probability, 1(1): 1-27. https://doi.org/10.64389/isp.2025.01103

[29] Gomes-Silva, F., Ramos, M., Cordeiro, G.M., Marinho, P., Andrade, T. (2019). The exponentiated Kumaraswamy-G class: General properties and application. Revista Colombiana de Estadística, 42(1): 1-33.

[30] Hassan, A.S., Almetwally, E.M., Ibrahim, G.M. (2021). Kumaraswamy inverted Topp-Leone distribution with applications to COVID-19 Data. Computers, Materials & Continua, 68(1): 337-358. http://doi.org/10.32604/cmc.2021.013971

[31] Arshad, M.Z., Iqbal, M.Z., Al Mutairi, A. (2021). Statistical properties of a new bathtub shaped failure rate model with applications in survival and failure rate data. International Journal of Statistics and Probability, 10(3): 49-68. https://doi.org/10.5539/ijsp.v10n3p49

[32] Al-Babtain, A.A., Elbatal, I., Chesneau, C., Elgarhy, M. (2021). Estimation of different types of entropies for the Kumaraswamy distribution. PLoS One, 16(3): e0249027. https://doi.org/10.1371/journal.pone.0249027

[33] Ramzan, Q., Qamar, S., Amin, M., Alshanbari, H.M., Nazeer, A., Elhassanein, A. (2022). On the extended generalized inverted Kumaraswamy distribution. Computational Intelligence and Neuroscience, 2022(1): 1612959. https://doi.org/10.1155/2022/1612959

[34] Suleiman, A.A., Daud, H., Singh, N.S.S., Othman, M., Ishaq, A.I., Sokkalingam, R. (2023). A novel odd beta prime-logistic distribution: Desirable mathematical properties and applications to engineering and environmental data. Sustainability, 15(13): 10239. https://doi.org/10.3390/su151310239

[35] Suleiman, A.A., Daud, H., Ishaq, A.I., Othman, M., Sokkalingam, R., Usman, A., Osi, A.A. (2023). The odd beta prime inverted Kumaraswamy distribution with application to COVID-19 mortality rate in Italy. Engineering Proceedings, 56(1): 218. https://doi.org/10.3390/ASEC2023-16310

[36] Suleiman, A.A., Daud, H., Singh, N.S.S., Ishaq, A.I., Othman, M. (2023). A new odd beta prime-burr X distribution with applications to petroleum rock sample data and COVID-19 mortality rate. Data, 8(9): 143. https://doi.org/10.3390/data8090143

[37] Bouamar, M., Ladjal, M. (2007). Evaluation of the performances of ANN and SVM techniques used in water quality classification. In 2007 14th IEEE International Conference on Electronics, Circuits and Systems, Marrakech, Morocco, pp. 1047-1050. https://doi.org/10.1109/ICECS.2007.4511173

[38] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. (2016). Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In 2016 6th International Conference on System Engineering and Technology (ICSET), Bandung, Indonesia, pp. 137-141. https://doi.org/10.1109/ICSEngT.2016.7849638

[39] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. Journal of Water Process Engineering, 48: 102920. https://doi.org/10.1016/j.jwpe.2022.102920

[40] Li, L., Qiao, J., Yu, G., Wang, L., Li, H.Y., Liao, C., Zhu, Z. (2022). Interpretable tree-based ensemble model for predicting beach water quality. Water Research, 211: 118078. https://doi.org/10.1016/j.watres.2022.118078

[41] Klakattawi, H.S. (2019). The Weibull-gamma distribution: Properties and applications. Entropy, 21(5): 438. https://doi.org/10.3390/e21050438

[42] Suleiman, A.A., Yousafzai, A.K., Zubair, M. (2023). Comparative analysis of machine learning and deep learning models for groundwater potability classification. Engineering Proceedings, 56(1): 249. https://doi.org/10.3390/ASEC2023-15506

[43] Suleiman, A.A., Ibrahim, A., Abdullahi, U.A. (2020). Statistical explanatory assessment of groundwater quality in Gwale LGA, Kano State, Northwest Nigeria. Hydrospatial Analysis, 4(1): 1-13. https://doi.org/10.21523/gcj3.2020040101

[44] Anantrasirichai, N., Bull, D. (2022). Artificial intelligence in the creative industries: A review. Artificial Intelligence Review, 55: 589-656. https://doi.org/10.1007/s10462-021-10039-7

[45] Salleh, S.F., Suleiman, A.A., Daud, H., Othman, M., Sokkalingam, R., Wagner, K. (2023). Tropically adapted passive building: A descriptive-analytical approach

using multiple linear regression and probability models to predict indoor temperature. Sustainability, 15(18): 13647. https://doi.org/10.3390/su151813647
[46] Salau, A.O., Markus, E.D., Assegie, T.A., Omeje, C.O., Eneh, J.N. (2023). Influence of class imbalance and resampling on classification accuracy of chronic kidney disease detection. Mathematical Modelling of Engineering Problems, 10(1): 48-54. https://doi.org/10.18280/mmep.100106
[47] Macherla, H., Muvva, V.R., Lella, K.K., Palisetti, J.R., Pulugu, D., Vatambeti, R. (2024). A deep reinforcement learning-based RNN model in a traffic control system for 5G-envisioned internet of vehicles. Mathematical Modelling of Engineering Problems, 11(1): 75-83. https://doi.org/10.18280/mmep.110107