
RESEARCH ARTICLE

Performance Evaluation of Nature-Inspired Optimization Algorithms vs. the Tweedy Model in Rainfall Data Forecasting: A Case Study in Kirkuk, Iraq

Taha Darwassh Hanawy Hussein

Clinical Laboratory Sciences, College of Pharmacy, University of Kirkuk, Kirkuk, Iraq

Corresponding Author: Taha Darwassh Hanawy Hussein **E-mail:** taha_hussein@uokirkuk.edu.iq

ABSTRACT

Accurate rainfall prediction is crucial for agricultural water management. However, complex data pose a challenge to traditional statistical models. Therefore, this study assesses the effectiveness of a series of predictive algorithms in analysing continuous data. We looked at traditional and contemporary algorithms and compared four major algorithms: Tweedie, WPO, XGBoost and HBOA. These models were evaluated with the help of statistical parameters such as coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE). According to the results, the WPO algorithm performed better in predicting the results. It showed the peak value for the coefficient of determination ($R^2 = 0.684$) and the minimal value for the root mean square error (RMSE = 10.58). Conversely, the Tweedie model demonstrated consistent and stable performance ($R^2 = 0.596$ and RMSE = 11.97). It constitutes an appropriate option for predictions that require a balance between accuracy and reliability. The HBOA algorithm attained ordinary performance metrics, characterised by a comparatively balanced statistical profile ($R^2 = 0.621$, MAE = 7.71, and RMSE = 11.59). However, the XGBoost algorithm performed worst, as evidenced by the lowest coefficients of determination and the highest error metrics ($R^2 = 0.35$, RMSE = 15.18, and MAE = 10.85). These results indicate a restricted efficacy when applied to this specific dataset. Accordingly, the WPO algorithm is recommended for predicting continuous data due to its highest performance. Nevertheless, the Tweedie model is a viable option when a balance between performance and accuracy is required.

KEYWORDS

Regression Algorithms, Tweedie Model, Whale Optimization Algorithm, XGBoost, Hybrid Bio-inspired Optimization Algorithms.

ARTICLE INFORMATION

ACCEPTED: 19 September 2025

PUBLISHED: 27 October 2025

DOI: 10.32996/jcsts.2026.8.12.2

1.Introduction:

Predicting future weather patterns is of paramount importance due to the impact of climatic conditions on vital sectors such as agriculture, air and sea transport, defense, and disaster management. Historically, weather forecasting has been based on numerical models, which are developed by solving complex mathematical equations describing atmospheric dynamics, based on remote sensing from satellites, such as temperature, humidity, precipitation and wind. These models are efficient but very demanding of computer resources, and are subject to inaccuracies due to initial conditions. In this context, AI and machine learning are becoming vital complementary components. They enable the analysis of massive amounts of data and the discovery of hidden patterns thanks to their ability to model complex and nonlinear relationships within the data. This contributes to improving the accuracy, speed, and efficiency of forecasts in modern systems. Today, machine learning is the foundation of contemporary weather prediction [1].

In recent years, the need for advanced analytical techniques has become essential due to the increasing volume and complexity of data, particularly in fields such as data science and artificial intelligence[2]. Agriculture under rain-fed conditions is very climate

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (<https://creativecommons.org/licenses/by/4.0/>). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

sensitive and the production of crops is directly dependent on the quantity and timing of rainfall. Rainfall modeling plays a pivotal role in influencing crop growth, as well as in managing droughts and floods. These models contribute to providing accurate data that supports advanced studies on the impacts of climate change. However, rainfall modeling faces several challenges, most notably the difficulty of accurately measuring precipitation. Rainfall data are characterized by their discontinuity, high dispersion, and high proportion of zero values resulting from low rainfall, especially during the summer. This complexity poses significant challenges for traditional statistical models, which often do not have the capacity to deal with it. [3].

Regression models are a crucial tool for predicting numerical values, ranging from simple linear models to more sophisticated and flexible models like the Tweedie model. The Tweedie model is distinguished by its ability to handle complex, highly dispersed distributions, such as those associated with insurance, loss, and rainfall data. As a result of these challenges, considerable efforts have been dedicated to developing advanced analytical techniques capable of processing and predicting data with greater accuracy. The Tweedie family of distributions is an effective choice for modeling datasets that include both discrete (zero) and continuous (non-zero) values. This family represents a special case of the exponential dispersion model, which is commonly used as a distribution within generalized linear models. The Tweedie family provides a framework that allows for the easy application of the generalized linear model, as auxiliary variables can be incorporated to influence the response variables. The Poisson-Gamma distribution within this family is suitable for representing rainfall components in a single, unified model, a point discussed in several studies such as [4,5,6].

[7]. This research explores modern algorithms for analyzing highly complex data and addresses modeling challenges. Mathematical tools and precise algorithms are employed to achieve optimal results.

The article is structured as follows. Section 2 provides a comprehensive review of the existing literature, addressing the theoretical foundations of the Tweedie model and the XGBoost algorithm, as well as nature-inspired optimization algorithms, including the whale optimization algorithm and hybrid models. The Tweedie model is unique in its capability in managing complex highly dispersed distributions such as those that occur in insurance, loss and rainfall data. The work concludes with Section 4, which presents experimental results accompanied by a detailed analysis and discussion. Finally, Section 5 concludes this research and outlines future directions.

2. Literature Review

Despite extensive research on the XGBoost and Tweedie models individually, few studies have systematically explored their integration to develop a novel approach [8]. Current research rarely focuses on using biologically inspired optimizers, such as WPO, to fine-tune Tweedie model parameters, particularly in hydrological contexts [9]. Comparative research is also limited in its ability to thoroughly investigate robustness for data with zero-inflation and for data with extreme events.

Tweedie models, part of the exponential dispersion family, are commonly used when the target variables are non-negative and have zero mass, such as in precipitation data that often includes rainless days. Furthermore, these models demonstrate efficiency in handling excessive dispersion or heavy-tailed data. These characteristics make Tweedie models an ideal choice for modeling rainfall data. Studies have proven their effectiveness in handling this type of data and achieving accurate results. Current studies rarely address biologically inspired enhancers, such as WPO, for adjusting Tweedie model parameters within hydrological applications. Rainfall models are of great importance in forecasting and simulation, with applications ranging from surface runoff modeling to soil water content analysis, and drought and flood forecasting [4,5]. Studies indicate that using Tweedie distributions within generalized estimation equations provides a robust framework for modeling rainfall across multiple locations. This approach allows for addressing high variability, precise zero repetitions, and the non-independent structure present in rainfall data.[10]. The authors in [11] found that Tweedie distributions, especially gamma distributions, are among the most accurate models for representing monthly rainfall and effectively handling zeros.

Studies have indicated that Tweedie family distributions, such as the gamma and Poisson-gamma distributions, are effective models for representing seasonal rainfall at Australian stations. These distributions are particularly well-suited for accurately simulating zeros and heavy rainfall events [12]. Models based on the Tweedie model in Australia demonstrated effective in explaining the effects of climate on monthly rainfall rates, by integrating cyclical and climatic variables within the GLM modeling framework, with the aim of enhancing understanding and improving prediction [13]. Research in Malaysia has shown that the Tweedie models (along with seasonal and climatic factors) accurately describe monthly rainfall. They further allow for more realistic statistical simulations of the effects of different geographical and climatic conditions [4]. The results of the studies showed that the Tweedie composite model using Lasso reduction was a good statistical model for predicting rainfall intensity. Additionally, it is useful to solve the problems of multicollinearity in large-scale climate data [14]. Research in Sweden has verified that the Tweedie

distribution is an appropriate model for the analysis of monthly precipitation, and that the parameters of the distribution are similar in Swedish environments to those found in Australian environments, especially the influence of coastal location [15].

Whale Optimization (WOA) is a bioengineering algorithm that implements whale hunting strategy. Known for its capacity to find a fine balance between exploration and exploitation in a search space. It has been proven to be effective and efficient in tuning parameters of statistical models and machine learning algorithms, especially in complex applications like climate models or rainfall forecasting. A variant of WOA is called Whale Particle Optimization (WPO), which combines the best of both worlds – WOA and PSO – to increase the diversity of particles and allow them to circumvent local solutions to obtain better solution. This algorithm is a novel and enhanced progress in systems optimization and engineering applications worldwide. This paper presents a novel, nature-inspired optimization algorithm, the WOA, which is based on simulating the social behavior of humpback whales. The main idea of this algorithm is based on the strategy of the bubble-hunting used by these creatures. To test the performance of the WOA algorithm, 29 mathematical optimization problems and 6 structural design problems were used. The results showed that the WOA algorithm is very efficient and competitive with the latest optimization algorithms and traditional algorithms [13,14]. Hybrid Bio-inspired Optimization Algorithms (HBOA)

HBOA is a family of computational techniques inspired by natural and biological phenomena to solve problems that are complex. The features from more than one algorithm are combined together to form the HBOA algorithm, typically evolutionary or swarm algorithms like PSO, GA, WOA and GWO. In the last two decades, there have been amazing developments of metaheuristic algorithms inspired by biological systems that excel at solving complex and nonlinear optimization problems, especially in dynamic and flexible environments[7][18][19]. These algorithms are based on principles of collective behavior of living organisms, natural evolution and physiological activities. Some of the remarked algorithms are the Harris hawk algorithm, the Walrus-like algorithm and the snarling fish algorithm[15,16,17]. One of the most popular machine learning algorithms is extreme gradient boosting (XGBoost), which is very efficient, scalable, and can be used for complex and nonlinear data. This algorithm is based on incremental optimization of tree models through a differentiable loss-based boosting technique, which can significantly decrease errors and increase overall model stability in the environmental field. Research has shown it to be an effective method for estimating temperature patterns in forests from open-field meteorological data [22][23][24].

Moreover, the prediction of solar radiation and groundwater level in the Malaysian state of Selangor has been done successfully using XGBoost model which is well capable of handling high natural noise time-series and climate data [19,20,21]. XGBoost is a popular method for prediction of climate and environmental sensitive variables [25][26]. In addition to XGBoost, other tree-boosting methods have been developed, For example, CatBoost has shown to be effective for analyzing insurance data, which is often unbalanced or zero-inflation, with a Tweedie Boosting model tailored for such situations. These models are suitable for complementing XGBoost and yield more precise solutions for data that has an abnormal distribution and high diversity [27].

This article is structured as follows. Section 2 provides a brief overview of the methods utilized: Tweedie regression model, WPO, HBOA, and XGBoost. The methodology and evaluation metrics are described in Section 3. The collecting and preparation of the required data are explained in Section 4. Sections 5 and 6 present the results and analyze them, respectively. Finally, Section 7 concludes the research and paves the way for future work

2.1 Tweedie Regression

The Tweedie regression model is a type of Generalized Linear Model (GLM) where the response variable follows a Tweedie distribution. It is useful when modeling non-negative data with many zeros and continuous positive values, such as rainfall, insurance claims, or trading volumes [28], [29], and [30]. It is used for modeling continuous rainfall data that may contain zeros or exhibit overdispersion. This class of models is a broad family of distributions defined by their form [31][32].

$$f(y|\theta, \varphi) = a(y, \varphi) \exp \left[\frac{y\theta - \kappa(\theta)}{\varphi} \right] \quad (1)$$

where the natural parameter is θ and the parameter φ is positive, the parameter φ is termed the dispersion and μ means the variation of random distributions is a special case of EDMS in which $\text{Var}(y) = \kappa(\theta) \varphi$, respectively. The Tweedie family is defined by the EDM variable $E(Y) = \mu = \kappa(\theta)$. There are multiple well-known distributions that describe the ventilation medium relationship. For example, natural ($P = 0$), $\text{Var}(\mu) = \varphi \mu^P$ for $P \in (0, 1)$. The Tweedie family ($p = 3$) includes. Poisson ($P = 1$), Gamma ($P = 2$) and the reverse Gaussian. The upper choice (1, 2) is most conducive to the semi-continuous results. Tweedie EDMs are summarised in

Table 1, and fall into the class of so-called Poisson-Gamma distributions [20, 21, 22]. are characterized by the parameter P. Table 1 summaries Tweedie EDMs [33].

$$E(X) = \lambda K'(\theta) \quad (2)$$

$$Var(X) = \lambda K''(\theta) \quad (3)$$

$$\text{Conditional mean:} \quad E(Y_i | X_i) = \mu_i \quad (4)$$

$$\text{Variance function:} \quad Var(Y_i) = \varphi \mu_i^p \quad (5)$$

$$\text{Link function (log link):} \quad g(\mu_i) = \log(\mu_i) = X_i^T \beta \quad (6)$$

$$\mu_i = \exp(X_i^T \beta) \quad (7)$$

$$\text{Log-likelihood function:} \quad \ell(\beta, \varphi, p) = \sum \log f(y_i | \mu_i, \varphi, p) \quad (8)$$

Table 1: A summary of Tweedie EDMs [33].

Distribution	p	α	M	
Extreme stable	$p < 0$	$1 < \alpha < 2$	$(0, \infty)$	
Normal	$P=0$	$\alpha = 2$	R	R
[Do not exist]	$0 < p < 1$	$2 < \alpha < \infty$		
Poisson	$P=1$	$\alpha \rightarrow -\infty$	$(0, \infty)$	N
Compound Poisson	$1 < p < 2$	$\alpha < 0$	$(0, \infty)$	$(0, \infty)$
Gamma	$P=2$	$\alpha = 0$	$(0, \infty)$	$(0, \infty)$
Positive stable	$p > 2$	$0 < \alpha < 1$	$(0, \infty)$	$(0, \infty)$
Extreme stable	$p \rightarrow \infty$	$\alpha = 1$	R	R

2.2 Whale Optimization Algorithm (WPO)

The Whale Optimization Algorithm (WOA) mimics the social behavior of humpback whales. The algorithm is inspired by the bubble net fishing strategy[34]. It belongs to the baleen whale family, which comprises seven species, and is known for its unique sleep pattern, in which one hemisphere of its brain remains active while the other rests. Humpback whales typically live in groups and primarily feed on krill and small fish. The humpback whale's distinctive hunting behavior involves a unique technique of forming circular patterns of bubbles around schools of fish to corral them, facilitating easier capture. The WPO draws on this natural behavior to provide effective strategies that combine exploration and exploitation to optimize the performance of various models. This algorithm is one of the most important nature-inspired optimization techniques, proven effective across a wide range of applications. The Key concepts of the WPO algorithm are as follows [35]:

First: Prey Encircling

Whales are known as the largest mammals in the world. An adult whale can reach 30 meters in length and weigh up to 180 tons. They hunt small schools of fish.[1] Merghelli and Lewis [34].[36] developed the Whale Optimization Algorithm (WOA) as a

recent addition to the family of nature inspired optimization algorithms. The WOA algorithm is based on the bubble-hunting behavior that humpback whales employ when targeting schools of krill or small fish near the water's surface. To execute this method, humpback whales swim around their prey in a gradually shrinking circle, while simultaneously spiraling to form circular or "9"-shaped bubbles. The WOA algorithm mimics this technique by offering a 50% probability of choosing between the shrinking encirclement mechanism or the spiral pattern to optimize the whale's position during the optimization process. The mathematical details of these mechanisms will be presented in the following section [37]. The algorithm assumes that the optimal location at any given moment is either the location of the prey itself or the location of the nearest whale. Based on this, all the whales then adjust their positions accordingly, moving towards the optimal location they have found [7,12]. Simulate the search and targeting process by mathematically modeling using the following equations.

$$\vec{D} = |\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)| \quad (9)$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D} \quad (10)$$

where t represents the current iteration, $\vec{A}$, $\vec{D}$ are coefficient $\vec{X}_p$ is the position vector of the prey, and $\vec{X}$ denotes the position vector of a whale. The vectors $\vec{A} \rightarrow$ and $\vec{A} \rightarrow$ are calculated as follows:

$$\begin{aligned} \vec{A} &= 2\vec{a} \cdot \vec{r}_1 - \vec{a} \\ \vec{C} &= 2 \cdot \vec{r}_2 \end{aligned} \quad (11)$$

where components of $\vec{a}$ are linearly decreased from 2 to 0 over the course of iterations, and r_1, r_2 are random vectors in $[0,1]$ [35].

Second: Bubble Net Attack Mechanism:

This strategy is simulated through two main methods[35]:

- a) **Reducing the Encircling Radius:** The search agents gradually move towards the prey by shrinking the search area.
- b) **Spiral Position Update:** The spiral path taken by the whale towards the prey is simulated using a logarithmic spiral mathematical function.

2.2.1 Spiral Bubble

The WOA is based on the "bubble-netting" technique used by humpback whales during their hunt. It consists of two stages, namely exploration for a wide spectrum of solutions and exploitation to narrow the range of solutions down to the best. This algorithm uses a spiral movement similar to how a whale moves towards its prey (as shown in Figure 1). The simplicity and limited number of parameters required for the algorithm are advantageous for its easy implementation and effectiveness in solving many problems [38].

Net Attack Patterns: Its mathematical representation of the spiral bubble encircling prey strategy used by whales in WOA is as follows:

$$\begin{aligned} X(\vec{t}+1) &= \vec{D}^T \cdot \exp^{ul} \cdot \cos(2\pi l) + X^*(t) \\ \vec{D}^T &= |\vec{X}^*(t) - X(\vec{t})| \end{aligned} \quad (12)$$

The model takes two parameters, bb – which affects the logarithmic spiral, and tt – a random integer between -1 and 1 – and then computes the distance of the i th whale from the target – the best solution found so far. Humpback whales use two main feeding methods: one is to 'follow the circle' in a spiral motion that decreases over time and the other is the 'spiral' method. At the optimization process, the whale position updates by random selection of one of these two paths (with 50% probability to choose the spiral path). The purpose of this method is to simulate different hunting actions of these whales in a proper manner. Here we briefly describe the mathematical model used to represent this strategy [35][38].

$$X(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A}\vec{D}p < 0.5 \\ \vec{D}^T e^{bt} \cos(2\pi t) + \vec{X}^*(t)p < 0.5 \end{cases} \quad (13)$$

where p is a number between 0 and 1 that is selected at random.

2.2.2 Search for Prey

The process of exploration as described here is the random search for the prey, which is accomplished by changing the value of the vector $\vec{A}$ [35]. This random search for prey according to where the sperm whale is, is how it hunts for food. The value of vector $\vec{A}$ is randomly changed between -1 and 1 during the exploration stage. The goal is to get the search agents to spread themselves out from the current position of the reference whale and therefore improve the diversity of the search process. At this stage, a search agent is picked up randomly to change its position according to the specified update mechanism. The reason for this is that it makes the algorithm more efficient in terms of covering the whole search space, even if $|\vec{A}| > 1$. This model can be mathematically described as follows [35]:

$$\vec{D} = |\vec{C}\vec{X}_{rand} - \vec{X}| \quad (14)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{A}\vec{D} \quad (15)$$

Where the $\vec{X}_{rand}$ is the position of a random vector, which is a position representing a whale randomly. The WOA has many similarities with other evolutionary computing techniques, including the Genetic Algorithm (GA). Both are trying to discover an optimum solution, and in both cases, the search for the solution is environmentally inspired, in this case by the animals looking for food. Algorithms based on these methods start with a random set of solutions and iteratively improve them through successive generations [35]. The Whale Optimization Algorithm (WOA) is also inspired by the natural behavior of whales, using the best solutions to steer the search process towards the optimal solution. Whale Optimization Algorithm is easier to implement than the Genetic Algorithm. It is similar to the Genetic Algorithm, but without evolutionary operations such as crossover and mutation. In order to address this drawback, an attempt is being made to improve the Whale Optimization Algorithm by adding a crossover mechanism to improve its performances [35]. The so called 1X method is described in the following equations:

$$Y = w_1x_1 + w_2x_2 + \dots + w_ix_i \quad (16)$$

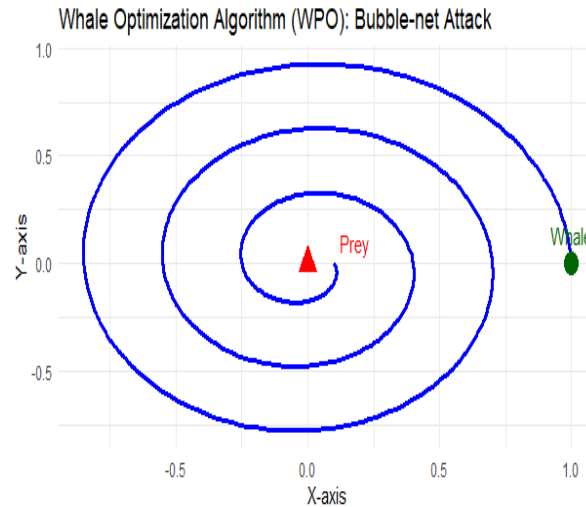


Figure 1: Visualizing the search and exploitation process within the WPO using a spiral bubble model (Drawing in R software).

2.3 Hybrid Bio-inspired Optimization Algorithms (HBOA)

Hybrid optimization algorithms (HBOA) are those which are inspired by a combination of more than one optimization algorithm and are simulated from a biological and natural phenomenon. These hybrid algorithms are designed to address the drawbacks of using single algorithms; they incorporate characteristics from biological evolutionary algorithms and characteristics from other approaches inspired by swarm of insects like PSO-GA, PSO-GWO and GWO. Together, this can help to achieve greater accuracy, quicker convergence, and greater robustness in solving complex problems. The scientific literature presents the development of hybrid algorithms such as PSO-GA, PSO-GWO, HHO-WOA and others multi-optimization algorithms that have shown to outperform the stand-alone algorithms.[39]:

1. **Nature + Nature:** Such as combining PSO and GA or WOA and GWO.
2. **Nature + Traditional Mathematical Techniques:** Such as integrating PSO with gradient descent.
3. **Nature and Machine Learning:** For example, using evolutionary algorithms to optimize neural network weights or select relevant features.

However, in this research, a simplified form of HBOA was used to optimise the power parameter (p) in the Tweedie regression model. The algorithm was developed to use three mechanisms: selection, crossover and mutation, in an iterative process to minimize the root mean squared error (RMSE). The integration of HBOA with the tweedie model is useful as an example of how hybrid algorithms can improve the accuracy of statistical models while keeping them easy to implement [40, 41, 42].

An evolutionary approach which takes advantage of the individual strengths of different evolutionary mechanisms and optimizes their balanced performance. It is based on the following components [43]:

A genetic algorithm based hybridization to improve the solutions.

2. A set of mutations that add random Gaussian noise to data diversity.
3. Selection and replacement plans based on performance to guarantee best results.

Although it is very simple, it is a good method to dynamically balance exploration and exploitation and has been used to effectively and flexibly optimize parameters for regression models [43] and [44].

The HBOA algorithm is a hybrid optimization algorithm that is inspired by the natural and biological phenomenon. It uses a number of techniques to modulate the solution space exploration and exploitation of the best solution, with the goal of achieving the best possible results in a complex problem. It is mostly used for parameter optimization, improving the estimation of parameters, predictive modelling and resource planning.

Fitness Function [45] and [46]:

$$F(x) = \sum_{i=1}^n w_i * fi(x) \quad (20)$$

where $F(x)$ is the fitness value of solution x , $fi(x)$ is the evaluation of criterion i , and w_i denotes the weight assigned to each criterion [47] and [48]:

$$X_{t+1} = X_t + \alpha * (X_{best} - X_t) + \beta * R_t \quad (21)$$

where X_t is the current position, X_{best} refers to the best solution found so far, α , β are learning and exploration parameters, and R_t is a random variable used to promote exploration.

Hybrid Strategy:

$$X_{t+1} = \gamma * X_{local} + (1 - \gamma) * X_{global} \quad (22)$$

where X_{local} is the local solution, X_{global} is the global solution, and γ is a parameter that balances exploitation and exploration.

2.4 Extreme Gradient Boosting algorithm (XGBoost)

The technique of gradient boosting is one of the most impactful in machine learning because it allows to train multiple models, such as decision trees, and fuse them into a single model that achieves a higher accuracy of prediction and lower bias and variance. One of the stand-out developments was the introduction of an innovative distributed weighted quantile sketch algorithm by Chen and Guestrin (2016) in [22] which allows to process weighted data with a provable theoretical guarantee, yielding a novel scalable and efficient algorithm, XGBoost. XGBoost has also been implemented in numerous programming languages, such as R, Julia and Python.[22].

The XGBoost is based on the following principle: "gradient boosting." It is a combination of gradient descent mechanisms and boosting methods and used within the framework of an algorithm called Gradient Boosting Machines (GBM). This method integrates multiple models and assigns different weights to the models in accordance with the errors and performance of the models during training [4]. Data points that happen to be misclassified are given a greater weight, whereas the weights of points

that are correctly classified are lowered; this will help to gradually improve the distribution of training data. In addition, GBM uses second order statistics of the gradients to improve model accuracy. XGBoost's systematic approach involves building a preliminary decision tree to capture the patterns in the data and correct the errors in the initial predictions. It then sequentially adds more decision trees using boosting techniques to get even more accurate [33]. Finally, all the decision trees are aggregated into a single decision tree with very high accuracy by using appropriately computed weights. The algorithm's advanced technical features and its high performance are suitable for complex classification and prediction tasks with high reliability and efficiency [34, 35].

The function of the prediction stage can be computed as follows [49]and[50]:

$$P_i^{(s)} = \sum_{k=1}^s P_k(y_i) = P_i^{(s-1)} + P_s(y_i) \quad (23)$$

where $P_s(y_i)$ refers to the step t in the learner. The original function can be modified to prevent overfitting[51]:

$$\begin{aligned} O^{(s)} &= \sum_{k=1}^n l(y - y_i) + \sum_{k=1}^s \theta(f_i) \\ \theta(f_i) &= \varphi T + 0.5\lambda\|w\|^2 \end{aligned} \quad (24)$$

The boosting approach uses small decision trees which are used to gradually improve the model's predictions, making the prediction accuracy better, especially on weak points of the model. Shrinkage parameter (λ) is very important because it determines how much each tree is going to contribute towards the improvement. The more trees in the collection, and the more diverse they are, the more they will help to correct errors. The cumulative nature of boosting, in contrast to bagging (where each new tree is the enhancement of the previous tree), enhances the organization and sequential nature of the model-building process. There are three important parameters that affect this process; how much is improved and what accuracy can be reached in the final model performance [52].

1. The number of trees (B): Boosting can overfit if B is too large, but this does not happen with bagging or random forests. But this overfitting generally happens gradually and usually does not happen at all. Optimization of B is done with the help of the cross-validation.
2. 3. The shrinkage parameter (λ): This is a small positive number which reflects the learning rate of the boosting process. It makes sure that the algorithm will not make sudden steps.
3. 3. Number of splits per tree (d): This is a parameter that determines the complexity of the boosted ensemble. A frequently successful option is $d = 1$, which means that each tree has only one split (is a stump) [52].

Boosting algorithm in regression:

1. Set

$\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.

2. For $b = 1, 2, \dots, B$, repeat: (a) Fit a tree

$\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r) .

(b) Update

$\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \quad (25)$$

2.5 Data Collection and Preparation:

Average monthly rainfall data over the Kirkuk Governorate, Iraq for the period 2008 to 2014 have been taken from the Kirkuk meteorological department. This data was daily rainfall readings and was the dependent variable. Further, a group of independent influencing variables was also taken into account, such as the level of humidity, wind speed, solar radiation intensity and the temperature. All data were standardized so that it was consistent and minimized the influence of different units of measurement. The entire data set was then split into two data sets: 70% to train the predictive models, and 30% to test the performance and accuracy of the predictive models.

3 EVALUATION METRICS

The prediction accuracy of the models was measured using a set of metrics, including the coefficient of determination (R^2) [53]. R^2 is used to measure how well a model explains the variability in the data. Its mathematical formula is:

$$R^2 = 1 - (SSR/SST) \quad (26)$$

Where:

$SSR = \sum (y_i - \hat{y}_i)^2$ is the residual sum of squares.

$SST = \sum (y_i - \bar{y})^2$ is the total sum of squares.

y_i : The actual values.

$\hat{y}_i$: The predicted values from the model.

$\bar{y}$: The mean of the actual values.

If R^2 equals 1, this means the model accounts for all variability, providing a perfect prediction. If $R^2 = 0$, it indicates that the model does not improve upon the explanatory power of the mean. R^2 can be less than 0 if the model performs worse than relying solely on the mean.

Mean absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (27)$$

Root Mean Squared Error (RMSE)

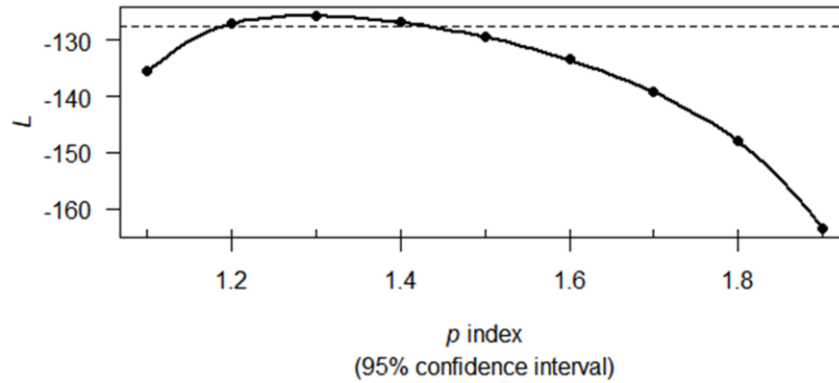
RMSE calculates the squared difference between the predicted values ($\hat{y}$) and the actual values (y).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (28)$$

These metrics provide a comprehensive view of each model's explanatory power and prediction error [54].

4 RESULTS AND DISCUSSION

The probability profile analysis of the Tweedy coefficient p is presented in figure 2. The maximum likelihood (MLE) value for p is $p = 1.2632$. This is between $p=1$ and $p=2$, which is the range of the composite Poisson-Gamma distribution. This shows that the distribution of the respondent variable has a non-zero mode and that there is a cluster of continuous positive values. Pr



Figur 2:Likelihood Profile for the Tweedie p Index with 95% Confidence Interval,p=1.2632.

The data were analyzed using the R program, with 70% used for training and 30% for testing; the results are listed in Table 2.

Table 2: Performance comparison table of prediction algorithms using R^2 , MAE, and RMSE metrics

Model	p	phi	RMSE	MAE	R2
tweedie	1.2632	5	11.97198	7.675406	0.596
WPO	-	-	10.5837	8.7248	0.684
XGBoost	-	-	15.17548	10.85029	0.35
HBOA	-	-	11.59302	7.713321	0.621

Figure 3 shows that the WPO model outperform the other models studied with accuracy, with lowest RMSE value of 10.5837, which means it has the best representation of the data and it is the most accurate model. The HBOA model was then tried, which had a slightly lower RMSE than the Tweedie model and represented average performance for both models, HBOA having slightly better. While the predictions from the XGBoost model was the lowest one with an RMSE of 15.17548, this model was the least accurate of all models studied, suggesting it may not match the data or may need to be improved. As a whole, the results show a clear difference in the efficiency of the models, with the WPO model outperforming the other models.

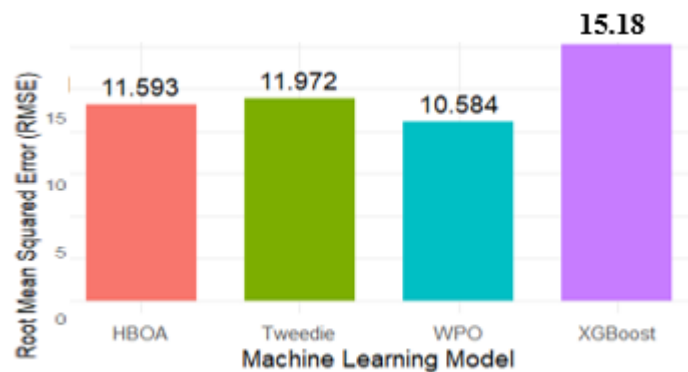


Figure 3: Performance of prediction algorithms and the Tweedie model using RMSE criteria.

The mean absolute errors (MAEs) of four predictive models are compared in figure 4. The Mean Absolute Error is the average of the absolute error between the observed and the predicted values, which is used as a measure of the accuracy of model. The Tweedie model had the lowest MAE (7.675), indicating that it was the best model for minimizing absolute prediction errors. The second model was the HBOA model, which had a very close value (7.713), but the results were similar to the first model, and it was very effective. The WPO model had the lowest accuracy of the two previous models, with the value of 8.7248 within acceptable error limits. The XGBoost model, on the other hand, possessed the highest MAE (10.85), suggesting the lowest accuracy among the four models, and showing greater deviation from the actual values.

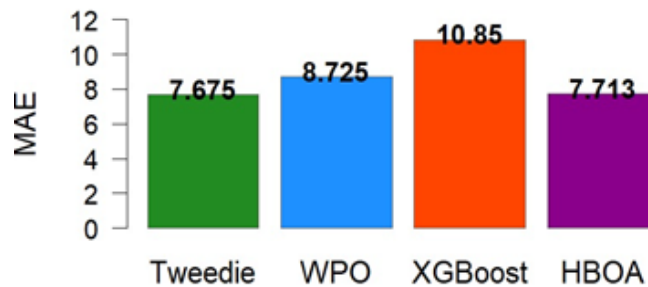


Figure 4 shows the performance of prediction algorithms and the Tweedie model based on the MAE Criteria.

The WPO model explains the maximum proportion of the variance in the data and is the best performing model ($R^2 = 0.684$) (see Figure 5). The model is similar to WPO ($R^2 = 0.621$) and is a good and efficient model especially for complex data. The Tweedie model has a moderate fit ($R^2 = 0.596$) and it should be used when there are excessive dispersion or zero values in data. The XGBoost model is the least successful ($R^2 = 0.35$), which indicates that it requires fine-tuning of its parameters or enhancements of the data.

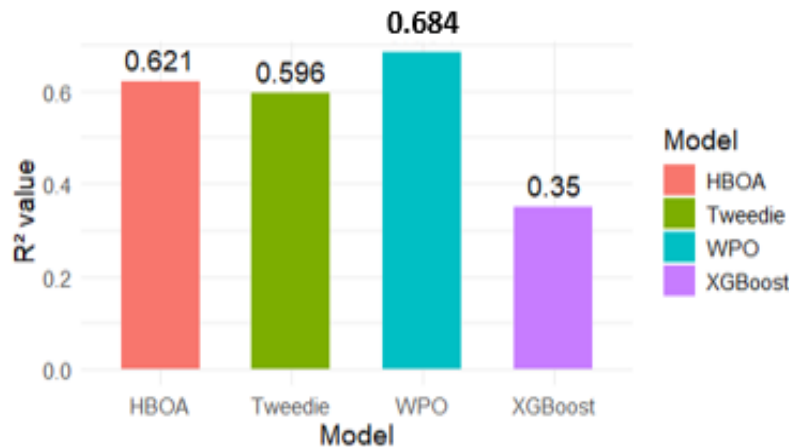


Figure 5: Performance of prediction algorithms and the Tweedie model using R^2 criteria

Figure 6 indicates that the algorithms have different performances. The highest coefficient of determination ($R^2 = 0.684$) was obtained with the WPO algorithm, which is best able to explain the variance in the data. Besides this it was the lowest root mean square error ($RMSE = 10.58$) among all which means that it predicts accurately and efficiently. The mean absolute error ($MAE = 8.72$) was not the best but the coefficient of determination ($R^2 = 91.28\%$) and root mean square error ($RMSE = 10.35$) were considered. So, the WPO is considered a good option for continuous data.

The Tweedie model proved to be stable and reliable, with a coefficient of determination (R^2) of 0.596, a mean absolute error (MAE) of 7.67, and a root mean square error (RMSE) of 11.97. While these results are less accurate than those of the WPO algorithm, the model illustrates its capacity to provide balanced and stable predictions which is acceptable for applications where reliability and stability is more important than maximum accuracy. The average performance of the HBOA algorithm was acceptable with the R^2

of 0.621, MAE of 7.71, and RMSE of 11.59. Despite not being able to achieve better results than WPO, it was able to achieve a balance between accuracy and error and performed better than the XGBoost model. With some modifications, like the use of other techniques or the adjustment of parameters, HBOA may be made competitive. In contrast, the XGBoost algorithm showed moderate performance in this study with the lowest coefficient of determination ($R^2 = 0.35$), and the highest error values (MAE = 10.85 and RMSE = 15.17).

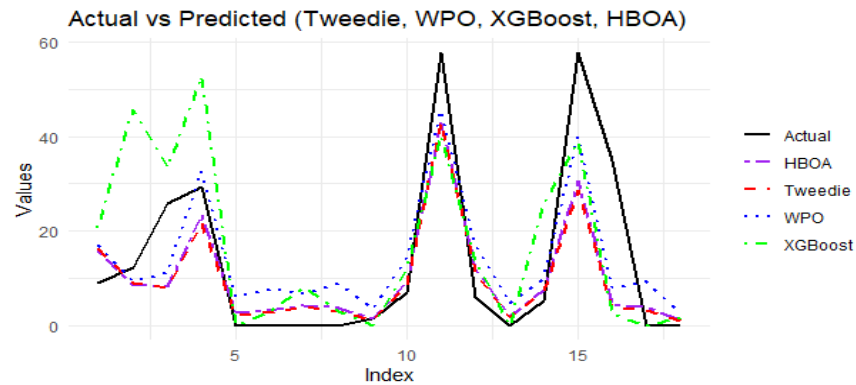


Figure 6: Actual values compared to predictions (Tweedie, WPO, XGBoost, and HBOA).

Lastly, the obtained results suggest that XGBoost has its weaknesses on the nature of this dataset and more research is needed to enhance its performance by tuning the parameters or raising the training dataset size.

5 CONCLUSIONS AND RECOMMENDATIONS

In this study, four predictive algorithms, including the Tweedie algorithm, the WPO algorithm, XGBoost, and HBOA, are assessed to see how well they can predict rainfall for agricultural water management. The results proved that WPO has better performance and high accuracy for processing complex and continuous data. Moreover, the Tweedie algorithm has been tested to perform well, reliably and with stability, which is ideal for long term sustainability. But HBOA had average results, which can be improved by fine tuning or by combining with other solutions. Also, although the performance of the XGBoost was also not desirable on this data, there is a need for a significant change and/or size increase of the input data.

Moreover, this study pointed to the possibility of utilizing the accuracy of WPO and the reliability of Tweedie as a solution to enhance the system efficiency and resilience. Using the WPO algorithm is a suitable option for applications requiring high accuracy, especially when handling complex and diverse data. It is an interesting fact that the Tweedie model is preferred when stability and reliability is important to success.

Given that it is an open area for research, the performance of the HBOA algorithm must be improved by tuning the parameters of the algorithm or by combining it with other more advanced technologies to boost its effectiveness. Furthermore, it is crucial that a combined solution based on the flexibility of the Tweedie model and the precision of the WPO algorithm is developed to create more efficient, integrated models.

REFERENCES

- [1] S. Attri, S. Gupta, and S. Singh, "Risk Forecasting Using Artificial Intelligence and Machine Learning," vol. 16, no. 1, pp. 187–196, 2025, doi: 10.4018/979-8-3373-1200-2.ch009.
- [2] R. Kumar et al., "Enhancing climate forecasting with AI: Current state and future prospect," F1000Research, vol. 13, p. 1094, 2024, doi: 10.12688/f1000research.154498.1.
- [3] N. C. Dzupire, P. Ngare, and L. Odongo, "A Poisson-Gamma Model for Zero Inflated Rainfall Data," J. Probab. Stat., vol. 2018, 2018, doi: 10.1155/2018/1012647.
- [4] J. Suhaila, "Tweedie models for Malaysia rainfall simulations with seasonal variabilities," vol. 14, no. 10, pp. 3648–3670, 2023, doi: 10.2166/wcc.2023.275.
- [5] Peter K. Dunn, "Occurrence and quantity of precipitation can be modelled simultaneously," Int. J. Climatol. <https://doi.org/10.1002/joc.1063>, vol. 24, no. 10, pp. 1231–1239, 2004.
- [6] P. K. D. Md Masud Hasan, "Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia," Int. J. Climatol. <https://doi.org/10.1002/joc.2162>, vol. 31, no. 9, pp. 1389–1397, 2011.
- [7] S. Dalal et al., "Optimized XGBoost Model with Whale Optimization Algorithm for Detecting Anomalies in Manufacturing," J. Comput. Cogn. Eng., vol. 00, no. August, pp. 1–11, 2024, doi: 10.47852/bonviewjccce42023545.
- [8] O. L. Aubert D. Loumagne C., "Sequential assimilation of soil moisture and stream flow data in a conceptual rainfall runoff model," J. Hydrol., vol. 280, pp. 145–161, 2003.

- [9] M. A. Toth E, Brath A, "Comparison of short-term rainfall prediction models for real-time flood forecasting," J. Hydrol., vol. 239, pp. 132–147., 2000.
- [10] T. Swan, "Generalized estimating equations when the response variable has a Tweedie distribution: an application for multi-site rainfall modelling," Dissertation, p. 168, 2006, [Online]. Available: <http://eprints.usq.edu.au/3388>
- [11] M. Hasan and P. K. Dunn, "Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia," vol. 1397, no. June 2010, pp. 1389–1397, 2011, doi: 10.1002/joc.2162.
- [12] M. Hasan and P. K. Dunn, "Seasonal rainfall totals of Australian stations can be modelled with distributions from the Tweedie family," vol. 3101, no. November 2014, pp. 3093–3101, 2015, doi: 10.1002/joc.4194.
- [13] M. Hasan and P. K. Dunn, "Understanding the effect of climatology on monthly rainfall amounts in Australia using Tweedie GLMs," vol. 1017, no. April 2011, pp. 1006–1017, 2012, doi: 10.1002/joc.2332.
- [14] M. Hayati and R. Permatasari, "Rainfall Prediction in Statistical Downscaling Using Tweedie Compound Response and Lasso Penalty," vol. 4099, pp. 537–546, 2023.
- [15] J. Rydén, "Statistical analysis of monthly precipitation in Sweden using the Tweedie distribution," 2025.
- [16] S. Mirjalili, A. Lewis, "The whale optimization algorithm," Adv. Eng. Softw. 95, pp. 51–67, 2016.
- [17] S. Mirjalili and A. Lewis, "Advances in Engineering Software The Whale Optimization Algorithm," vol. 95, pp. 51–67, 2016, doi: 10.1016/j.advengsoft.2016.01.008.
- [18] X. Yang, Nature-Inspired Optimization Algorithms. 2014.
- [19] T. Cai et al., "Dynamic Niches-Based Hybrid Breeding Optimization Algorithm for Solving Multi-Modal Optimization Problem," Mathematics, vol. 12, no. 17, 2024, doi: 10.3390/math12172779.
- [20] G. Kponhinto, S. Khemiri-Kallel, A. A. Abba Ari, A. M. Gueroui, O. Thiare, and J. Hwang, A Hybrid Biologically-Inspired Optimization Algorithm for Data Gathering in IoT Sensor Networks, vol. 33, no. 2. Springer US, 2025. doi: 10.1007/s10922-025-09899-z.
- [21] P. Trojovský and M. Dehghani, A new bio-inspired metaheuristic algorithm for solving optimization problems based on walrus behavior, vol. 13, no. 1. Nature Publishing Group UK, 2023. doi: 10.1038/s41598-023-35863-5.
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Aug, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
- [23] F. Ghafarian, R. Wieland, D. Lüttschwager, and C. Nendel, "Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data," Environ. Model. Softw., vol. 156, no. February, p. 105466, 2022, doi: 10.1016/j.envsoft.2022.105466.
- [24] C. N. Obiora, A. Ali, and A. N. Hasan, "Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance," 2021 IEEE PES/IAS PowerAfrica, PowerAfrica 2021, no. August, 2021, doi: 10.1109/PowerAfrica52236.2021.9543159.
- [25] A. Ibrahim Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," Ain Shams Eng. J., vol. 12, no. 2, pp. 1545–1556, 2021, doi: 10.1016/j.asej.2020.11.011.
- [26] B. So and E. A. Valdez, "Zero-inflated Tweedie boosted trees with CatBoost for Insurance loss analytics," Appl. Soft Comput., vol. 177, no. July, 2025, doi: 10.1016/j.asoc.2025.113226.
- [27] H. Zhou, W. Qian, and Y. Yang, "Tweedie gradient boosting for extremely unbalanced zero-inflated data," Commun. Stat. Simul. Comput., pp. 1–23, 2020, doi: 10.1080/03610918.2020.1772302.
- [28] G. M. Cordeiro, R. Labouriau, and D. A. Botter, "An introduction to Bent Jørgensen's ideas," Brazilian J. Probab. Stat., vol. 35, no. 1, pp. 2–20, 2021, doi: 10.1214/19-BJPS458.
- [29] P. K. Dunn and G. K. Smyth, Evaluation of Tweedie exponential dispersion model densities by Fourier inversion, vol. 18, no. 1. 2008. doi: 10.1007/s11222-007-9039-6.
- [30] P. Dunn and G. Smyth, "Tweedie family densities: methods of evaluation," Proc. 16th Int. Work. ..., no. December 2014, pp. 1–8, 2001.
- [31] C. F. Kurz, "Tweedie distributions for fitting semicontinuous health care utilization cost data," BMC Med. Res. Methodol., vol. 17, no. 1, pp. 1–9, 2017, doi: 10.1186/s12874-017-0445-y.
- [32] W. A. Hadba, "Tweedie regularization model and application," Commun. Stat. - Theory Methods, vol. 54, no. 6, pp. 1729–1746, 2025, doi: 10.1080/03610926.2024.2349698.
- [33] Jørgensen B., "The Theory of Dispersion Models," 1 Ed. Chapman Hall/CRC, 1997.
- [34] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," Adv. Eng. Softw., vol. 95, pp. 51–67, 2016, doi: 10.1016/j.advengsoft.2016.01.008.
- [35] Z. Adil Madloul, "Develop Whale Optimization Algorithm(WOA) In Genetic Method To Predict the Optimal Treatment for Diseases," J. Al-Qadisiyah Comput. Sci. Math., vol. 16, no. 4, 2024, doi: 10.29304/jqscm.2024.16.41792.
- [36] N. Khodadadi, E. S. M. El-kenawy, S. Faridmarandi, M. S. Ghahfarokhi, A. Ibrahim, and S. Mirjalili, "A novel version of whale optimization algorithm for solving optimization problems," Handb. Whale Optim. Algorithm Var. Hybrids, Improv. Appl., no. November, pp. 591–600, 2023, doi: 10.1016/B978-0-32-395365-8.00048-8.
- [37] A. Kaveh, M.I. Ghazaan, "Enhanced whale optimization algorithm for sizing optimization of skeletal structures," Mech. Based Des. Struct. Mach., vol. 45, no. 3, pp. 345–362, 2017.
- [38] S. Mahmood, N. Z. Bawany, and M. R. Tanweer, "A comprehensive survey of whale optimization algorithm: modifications and classification," Indones. J. Electr. Eng. Comput. Sci., vol. 29, no. 2, pp. 899–910, 2023, doi: 10.11591/ijeecs.v29.i2.pp899-910.
- [39] I. Rahman and Junita Mohamad-Saleh, "Hybrid bio-Inspired computational intelligence techniques for solving power system optimization problems: A comprehensive survey," Appl. Soft Comput., vol. 69, pp. 72–130, 2018.
- [40] Amit Malik, Amita Rani, "Hormesis-Based Optimization (HBO) Algorithm: A Biologically Inspired Computational Approach," J. Inf. Syst. Eng. Manag., vol. 10, no. 49s, pp. 1229–1254, 2025, doi: 10.52783/jisem.v10i49s.10114.
- [41] and M. B. M. Del Gallo, G. Mazzuto, F. E. Ciarapica, "Artificial Intelligence to Solve Production Scheduling Problems in Real Industrial Settings: Systematic Literature Review," Electron. vol. 12, Nov. 2023., vol. 12 Nov, 2023.
- [42] H. Bouzary and F. F. Chen, "A classification-based approach for integrated service matching and composition in cloud manufacturing,"

-
- Robot. Comput. Manuf. vol. 66, Artic. , 2020, vol. 66, no. ID 101989, 2020.
- [43] M. Mitchell, "An Introduction to Genetic Algorithms," MIT Press., 1998.
- [44] T. et al. Cai, "Dynamic niche technology based hybrid breeding optimization algorithm for multimodal feature selection.," Sci. Reports., 2025.
- [45] D. Nguyen, T., & Le, "Hybrid Bio-inspired Optimization Algorithms," A Rev. Appl. Soft Comput., vol. 110, 10764, 2021.
- [46] X. Zhang, Y., & Li, "HBOA for Parameter Optimization in Predictive Models.," Expert Syst. with Appl. ., vol. 159, 11359, 2020.
- [47] I. Hanif, "Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction," 2020, doi: 10.4108/eai.2-8-2019.2290338.
- [48] P. Vilash and D. M. A. Hameed, "Improvement Of Extreme Gradient Boosting (XGBoost) Algorithm Using Neural Decision Trees on Miscarriage Dataset," SSRN Electron. J., 2025, doi: 10.2139/ssrn.5190287.
- [49] N. S. Alkurdi, A. A., Asaad, R. R., Almufti, S. M., & Ahmed, "Evaluating the impact of point-biserial correlation-based feature selection on machine learning classifiers," a Credit card Fraud Detect. case study. Rev. Gestão Tecnol., vol. 24(Special, pp. 166–196, 2024.
- [50] H. B. Abdulrahman, S. M., Hani, A. A., Zeebaree, S. R., Asaad, R. R., Majeed, D. A., Sallow, A. B., & Ahmad, "Intelligent home iot devices: an exploration of machine learning- based networked traffic investigation.," J. Ilm. Ilmu Terap. Univ. Jambi, vol. 8, no. 1, pp. 1–10, 2024.
- [51] A. S. Aziz and H. R. Fadlallah, "Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) models for Hepatitis C Prediction," Int. J. Adv. Appl. Comput. Intell., vol. 3, no. 2, pp. 18–28, 2023, doi: 10.54216/ijaaci.030202.
- [52] "statistical_machine_learning RR مهم.pdf."
- [53] J. Gao, "R-Squared (R^2) – How much variation is explained? , " Res. Methods Med. Heal. Sci., vol. 5, no. 4, pp. 104–109, 2024, doi: 10.1177/26320843231186398.
- [54] S. M. Palacio, "Machine-Learning Applied Methods PhD in Economics Machine-Learning Applied Methods Sebastián Mauricio Palacio".