Received: 2 June 2024 | Revised: 29 August 2024 | Accepted: 27 September 2024 | Published online: 14 October 2024

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2024, Vol. 00(00) 1–11
DOI: 10.47852/bonviewJCCE42023545

BON VIEW PUBLISHING

Optimized XGBoost Model with Whale Optimization Algorithm for Detecting Anomalies in Manufacturing

Surjeet Dalal[1], Uma Rani[2], Umesh Kumar Lilhore[3], Neeraj Dahiya[4], Reenu Batra[5], Nasratullah Nuristani[6,*] and Dac-Nhuong Le[7]

[1]*Department of Computer Science and Engineering, Amity University Haryana Gurugram, India*

[2]*Department of Computer Science and Engineering, World College of Technology and Management, India*

[3]*Department of Computer Science and Engineering, Galgotias University, India*

[4]*Department of Computer Science and Engineering, SRM University Delhi-NCR, India*

[5]*Department of Computer Science and Engineering, Global Institute of Technology and Management, India*

[6]*Department of Spectrum Management, Afghanistan Telecommunication Regulatory Authority, Afghanistan*

[7]*Faculty of Information Technology, Haiphong University, Vietnam*

Abstract: Anomalies and defects in the manufacturing process hinder operating efficiency and product quality. The Whale Optimization Algorithm (WOA) optimizes the XGBoost model for better anomaly identification by iteratively refining hyperparameters. Experiments using real-world manufacturing datasets prove proposed model works. Comparing the proposed model to traditional anomaly detection methods shows its superior performance in industry patent concept. The optimized XGBoost model's interpretability and anomaly detection features are also discussed. In this paper, WOA is applied in this work to optimize hyperparameters of XGBoost, a robust gradient boosting technique for accurate anomaly detection in manufacturing systems. Optimized XGBoost gained 1.00 precision value, 0.9 recall value, and 0.96 f1-score for class 0.0 and gained a 0.95 precision value, 1.00 recall value, and a 0.97 f1-score for class 1.0. The proposed model gained 0.993 Train Score and 0.964 Test Score. Our findings suggest that integrating XGBoost with the WOA may uncover manufacturing process irregularities. Optimization improves detection accuracy and provides a flexible and interpretable framework, helping modern industrial processes maintain quality and efficiency. This research encourages machine learning optimization for industrial patent applications, advancing anomaly detection methods.

Keywords: XGBoost, Whale Optimization Algorithm (WOA), anomalies detection, manufacturing, Industry 4.0

1. Introduction

Modern manufacturing requires anomaly detection to ensure product quality, operational efficiency, and company performance. Strong anomaly detection approaches are needed as industries use more automated systems and novel technologies. Anomaly detection is crucial to product quality. Even minor deviations from the norm can affect production. Manufacturing companies may quickly repair problems with real-time anomaly detection systems before they finish products. To meet quality and consumer expectations, this proactive approach is needed.

A good factory has efficient production procedures. Anomaly detection boosts operational efficiency by preventing disruptions. Anomaly detection tracks machine characteristics, finds equipment issues, and optimizes production schedules to optimize maintenance, limit downtime, and ensure material quality. Early detection of abnormalities can yield significant financial rewards. Manufacturers can avoid faulty goods, excessive rework, and waste by addressing concerns quickly. Predictive maintenance based on anomaly detection can increase machinery and equipment lifespan, reducing repair and replacement costs and improving cost-effectiveness [1].

Complex supply chains and machinery are typical in manufacturing. Any process irregularity can threaten safety and cost money. Anomaly detection prevents equipment failures, material disparities, and process anomalies. Preventative risk management is crucial for workplace safety and the company's financial line. Instant decision-making is crucial in the ever-changing manufacturing industry. Anomaly detection systems let plant operators make quick decisions by revealing anomalies during operations. To avoid

***Corresponding author:** Nasratullah Nuristani, Department of Spectrum Management, Afghanistan Telecommunication Regulatory Authority, Afghanistan. Email: n.nuristani@atra.gov.af

problems that can snowball, you need to be able to change machine settings, conduct maintenance, or temporarily stop production in real time [2].

Technology, consumer tastes, and new laws all shape industrial processes. Iteratively learning from new data lets anomaly detection systems adapt. The detection models' ability to adapt to changing industrial conditions ensures their outlier detection accuracy. Anomaly detection is essential in production. Its importance goes beyond quality control to operational efficiency, cost reduction, risk mitigation, and real-time decision-making. Manufacturing companies must invest in effective anomaly detection systems to stay competitive and survive as more industries utilize data-driven strategies and automation. Manufacturers should prioritize anomaly detection to enhance processes, produce high-quality goods, and handle today's industrial complexity [3].

1.1. Anomaly detection algorithms

Industry 4.0 has transformed production with IoT, AI, and ML. Anomaly detection algorithms in manufacturing processes are one key application of this technology. These algorithms are essential for identifying irregularities, ensuring product quality, decreasing downtime, and maximizing productivity.

1.2. Statistical methods

Manufacturing relies on statistical anomaly detection algorithms. Process parameters and sensor data are monitored using Z-scores, median, standard deviation, and mean. A considerable deviation from statistical standards may raise concerns and require further investigation. These strategies help identify anomalies when historical data establishes usual behavior [4]. The Z-score of a data point shows how far it is from the mean of a distribution. If your dataset has a normally distributed mean (C) and standard deviation (σ), you can calculate the Z-score (Z) for a given data point (X) using the following formula:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

To identify anomalies, the Mahalanobis Distance (D) of a data point X in a multivariate dataset with a mean vector v and a covariance matrix can be estimated by setting a threshold on the absolute Z-score.

$$D = (X - \mu)^{t} \sum X^{-1} (X - \mu) \tag{2}$$

1.3. Machine learning algorithms

1) Supervised learning:

Supervised anomaly detection train algorithms with labeled datasets of normal and abnormal behavior [5]. Because of this, the model can identify typical process features and trends. Supervised learning systems like Random Forests and Support Vector Machines identify manufacturing anomalies.

2) Unsupervised learning:

Unsupervised learning approaches work well in weakly described anomalies or with scarcely labeled data. PCA, K-Means, and hierarchical clustering are used in this lesson.

3) Deep leaning:

Deep learning, especially autoencoder neural networks, is increasingly used to detect anomalies. Deviating patterns drive autoencoders' learning process, which reconstructs input data. These algorithms are excellent for spotting small manufacturing process anomalies because they can find nuanced relationships in multidimensional data.

4) Time series analysis:

Manufacturing uses time series data and advanced algorithms to analyze patterns. Time series data can be evaluated for outliers using exponential smoothing, seasonal decomposition, and moving averages. These methods reveal trends, cyclical patterns, and surprising outliers in sequential production data [6].

5) Ensemble methods:

Combining anomaly detection systems improves ensemble performance. Ensembles use numerous algorithms to improve detection accuracy. Manufacturers employ ensemble methods like bagging and boosting to improve anomaly detection.

6) Hybrid approaches:

Domain-specific knowledge, statistical methods, and machine learning algorithms form a hybrid approach. These approaches use complementary methods to discover more anomalies and adapt to different production conditions.

Anomaly detection algorithms revolutionize production precision, productivity, and quality. Many technologies, from statistics to advanced machine learning, allow manufacturers to proactively recognize and fix unusual events. Advanced anomaly detection algorithms will define manufacturing's future by guaranteeing operational excellence and product quality is intimately linked as technology improves [7].

1.4. Problem formulation

Automated systems and Industry 4.0 technology make production efficiency and quality crucial in today's industry. Anomaly detection is essential to find abnormalities in normal operation and act quickly to reduce defects, downtime, and efficiency. The complexity and challenges of designing and deploying dependable industrial anomaly detection systems make them an intriguing research topic. The main challenge is to adapt powerful anomaly detection systems to complicated industrial processes. Multivariate data complexity, real-time detection capabilities, and changing production settings cause issues [8–10]. We intend to help create flexible manufacturing anomaly detection technologies by solving these difficulties.

Key aspects of the research problem:

1) **Dynamic and evolving processes:** Product deviations, equipment wear, and maintenance activities affect manufacturing process parameters, making them dynamic. The research focuses on detecting abnormalities that change with these processes and keeping detection models current and accurate.
2) **Heterogeneous, multivariate data:** Multivariate sensor and source data are common in production. Working with varied data sets and extracting anomaly detection features is difficult. This research topic requires cutting-edge feature engineering, dimensionality reduction, and data fusion strategies to capture anomalies in complicated, multivariate datasets.

3) **Real-time detection criteria:** In production, punctuality is crucial. To repair problems quickly, anomalies must be detected in real time. We want to create a way to rapidly and effectively identify manufacturing schedule abnormalities so we don't make faulty products.
4) **Unsupervised learning in constrained data sets:** Labeled anomaly data is expensive and difficult to obtain thus, many manufacturing processes lack it. To solve this research subject, we require effective unsupervised learning systems that can use industrial processes' massive unlabeled data with sparsely labeled data.
5) **Interpretable models for operator understanding:** In manufacturing, humans and machines must cooperate. This work area requires easy-to-understand anomaly detection models so operators can trust the system's conclusions. This fosters teamwork and simplifies human-in-the-loop procedures.

Addressing the research challenge has major manufacturing implications. Solution execution can improve efficiency, operational costs, and product quality. Anomaly detection improvements can enhance industrial automation and enable adaptable and robust production systems.

2. Review Literature

According to Crespino et al. [11], traditional data processing systems cannot handle rising data quantities. Automated data collection and big data analytics have helped certain organizations boost earnings and customer satisfaction. Big Data is helping aircraft manufacturers increase production and sales. Real-time predictive analysis improves output and quality by identifying industrial anomalies. The initial findings of the TOREADOR European project pilot aerospace action research work will be briefly covered in this paper.

Nakazawa and Kulkarni [12] identify and separate wafer map defect anomalies using deep convolutional encoder-decoder neural networks. We generate synthetic wafer maps for eight defect patterns for training, validation, and testing using a model. Abnormality detection relies on pattern recognition. Our synthetic wafer map models trained on training basis partition can detect hidden defect patterns in genuine wafer maps.

Scime et al. [13] propose a Convolutional Neural Network architecture for layer-wise powder bed imaging data pixel-wise localization (semantic segmentation). Real-time performance, data transfer across additive manufacturing machines, and segmentation at the imaging sensor's original resolution are the algorithm's main benefits. Six devices exhibit electron beam, laser, and binder jetting algorithms. Finally, the method outperforms the authors' previous methods in localization, accuracy, computing time, and generalizability.

An anomaly detection system by Bozcan et al. [14] creates a workload-based abnormality score. Robots work cleverly than manufacturing experts. The framework addresses two fundamental method concerns. First, it warns people when the robot's behavior deviates significantly. Second, the framework can guide human professionals to record innovative state demonstrations, preventing redundant samples. This architecture outperforms parametric models for smart manufacturing applications that use behavioral cloning and a growing training dataset. Four real-world refrigerator assembly line inspection and automobile plug datasets are provided. We claim it is the largest real-world dataset for smart manufacturing anomaly detection employing robotic arms. Our technique is compared to others utilizing real-world datasets and quantitative evaluation.

Lee et al. [15] propose time series anomaly detection utilizing this dataset. The popular time series data LSTM model is combined with the anomaly detection SVDD model in the LSTM-SVDD model. The model can learn the typical data range and identify anything outside of it as odd. Non-learning test data distribution associated with predictions. Another strong performance measure is ROC (96.31). Automatic anomaly categorization may help small manufacturers with minimal AI infrastructure.

Jeon et al. [16] improve anomaly detection for car camera lens dirtiness. Regardless of pre-annotated locations, training and testing trim and resize input photos and whiten soiling masks. In the anomaly detection challenge, patch-wise detection approaches outscored reconstructive and probabilistic algorithms by 1.6% and 0.7%, respectively.

For unsupervised picture anomaly detection and segmentation, Wan et al. [17] offer pretrained feature mapping. The proposed PFM translates the image from one pretrained feature space to another, facilitating anomaly detection. We explore and recommend multi-hierarchical bidirectional PFM to improve performance. The recommended approach outperforms state-of-the-art algorithms on the famous MVTec AD dataset with 97.5% anomaly detection and 97.3% anomaly segmentation across all 15 categories. Overall and in terms of computation time, the proposed approach is better. Comprehensive ablation experiments prove the framework's efficiency.

Subhan et al. [18] identify image-based structural anomalies in industrial execution systems using an optimized VGG16 convolutional neural network. The optimized VGG16 model classifies test data as normal or exceptional using binary classification. Compared to another classifier, the updated VGG16 has high anomaly detection accuracy and might improve system reliability. Experimental results using publicly available image-based anomaly datasets show the suggested technique detects management execution system anomalies well.

In 2023, Kim [19] detected manufacturing irregularities using deep generative model-based self-supervised representation learning. An imbalance in data for regular and defective products hurts deep learning models. We employ the Gramian angular field to portray time series data, StyleGAN to enrich images with anomalous data, and boosting for supervised learning classifier selection to solve this problem. We tested the classifier before and after data augmentation. In wire arc additive manufacturing and CNC milling machine tests, the suggested strategy enhanced anomaly detection accuracy, recall, and F1-score over pre-augmentation. Bayesian tuning of the boosting algorithm's hyperparameters greatly enhanced performance. The proposed method solves data imbalances and can be applied to various industrial businesses.

The literature work by Trilles et al. [20] maps EC anomaly detection trials with MCUs. In 2021–2023, 18 of 162 scientific paper manuscripts from four databases were published. This article extensively discusses TinyML and MCU anomaly detection. ML/DL anomaly detection techniques, AIoT validation metrics, model estimation data, ML's software and hardware applications in EC, the most common microcontrollers, power supplies, and communication technologies, and a TinyML anomaly detection algorithm taxonomy will be examined in this survey. Finally, TinyML analysis and its benefits and downsides are discussed.

Gunasegaram et al. [21] offer ML-assisted CLC to control AM irregularities and defects. Avoidance, mitigation, and repair relate to defect causes, in-situ detectability, and controllability. First, we examine machine learning models for offline

optimization and in-situ diagnostics, then we address in-situ process control with ML. Researchers used inverse ML models or reinforcement learning to make situation-aware control decisions quickly. The defects addressed so far are easy to assess, and in-situ management systems that apply ML prioritize mitigation over prevention or repair. We also underline the multiple technologies needed for industrial autonomous in-situ control. Finally, we examine important but under-discussed adaptive control difficulties. Our work initiates thorough talks on in-situ adaptive control, filling a vacuum in AM literature.

2.1. Research gaps

Industrial process anomaly detection methods have improved but information gaps still exist. We must uncover and fix these gaps before building anomaly detection technologies that better address production challenges.

1) Adaptability to Dynamic Processes: Many anomaly detection methods cannot dynamically adjust because industrial processes change. Changing products and deteriorating machinery are two of many reasons manufacturing operations fluctuate.
2) Handling Heterogeneous, Multivariate Data: Current anomaly detection systems can't handle manufacturing's complex, multi-dimensional, and heterogeneous data. Manufacturing processes have many variables, making it challenging for algorithms to record and analyze them.
3) Real-time anomaly detection: Preventing production issues and downtime demands immediate action. Some anomaly detection systems may not work in real time, causing delays in detection and response.
4) Unsupervised Learning in Data-Scarce Environments: Many manufacturing processes lack labeled data for anomaly detection model training, making supervised learning difficult.
5) The interpretation of anomaly detection models is still far from complete, especially in sectors that require human contact. Human operators may struggle to trust and understand black-box models.
6) Benchmark Datasets and Evaluation Metrics: Lack of established benchmark datasets and clear evaluation criteria for manufacturing anomaly identification makes comparing methodologies and generalizing research outcomes difficult.
7) Security and Privacy Concerns: Since manufacturing systems handle sensitive data, anomaly detection methods' privacy and security effects need further work.

By filling these research gaps, manufacturers may create more effective, adaptive, and industry-specific anomaly detection solutions. Quality control, operating efficiency, and production reliability will improve.

3. Material and Methods

3.1. Dataset

Labeled datasets make anomaly detection difficult due to human bias in judging outcomes. The Industrial Internet of Things (IIOT) tracks huge manufacturing equipment every 10 milliseconds. Few manufacturers want to build anything truly distinctive. Due to infrequent aberrations, the dataset is substantially skewed. A top Indian wafer (semiconductor) manufacturer provided a use case. The anonymized dataset has 1558 attributes that need topic expertise.

This dataset includes Train.csv file having 1763 rows and 1559 columns. Test.csv consists of 756 rows and 1558 columns. Attributes Feature_1 to Feature_1558 represents the various attributes that were collected from the manufacturing machine shown in Figure 1.

Class (0 or 1) represents Good/Anomalous class labels for the products. Figure 2 shows the heat map of those features as below.

3.2. Proposed XGBoost model

In today's fast-paced manufacturing environment, where customer satisfaction is key, product quality is essential. As more industries adopt new technology, anomaly detection solutions are needed. This concept is demonstrated by XGBoost, a powerful machine learning algorithm. Extreme Gradient Boosting (XGBoost) is a powerful and versatile machine learning technique. It solves classification and regression problems well and was created by Tianqi Chen. Decision trees and ensemble learning make XGBoost a more accurate prediction model. Complex systems and many variables make manual anomaly detection in manufacturing processes difficult. Traditional rule-based systems may struggle to keep up with modern assembly lines' dynamic nature. Thus, data-driven technologies that immediately search mountains of data for patterns and outliers are in demand [22].

XGBoost (eXtreme Gradient Boosting) is a popular machine learning algorithm used for regression and classification tasks. While the core of XGBoost involves boosting decision trees, the mathematical equations governing its optimization process can be complex. The primary objective of XGBoost is to minimize a loss function by iteratively adding weak learners (typically decision trees) to the model, as described in [23].

XGBoost minimizes a loss function and regularization term-sum objective function. The XGBoost objective function is:

$$Objectives = \sum_{i=1}^{n} L(y_i, y_k) + \sum_{k=1}^{K} \beta(\mathrm{f_k}) \tag{3}$$

Where:
L: Loss function
Y_i: Predicted output
K: Number of trees
β: Regularization term

These equations provide a high-level overview of the mathematical components involved in XGBoost. The optimization process, including the handling of missing values, tree construction, and updates to the weights, involves additional details. The actual implementation of XGBoost is highly optimized for efficiency and speed. XGBoost is primarily used for supervised learning tasks such as regression and classification, and it is not inherently designed for anomaly detection [24–27]. However, the authors adapt XGBoost for anomaly detection in manufacturing by leveraging it within an ensemble learning framework. Here's a general outline of the approach, along with some key equations:

1) Define a target variable for anomaly detection:

The authors usually use y to signify the label or outcome. An anomaly detection binary target variable indicates whether a sample is normal (0) or abnormal (1).

Figure 1
Feature distribution

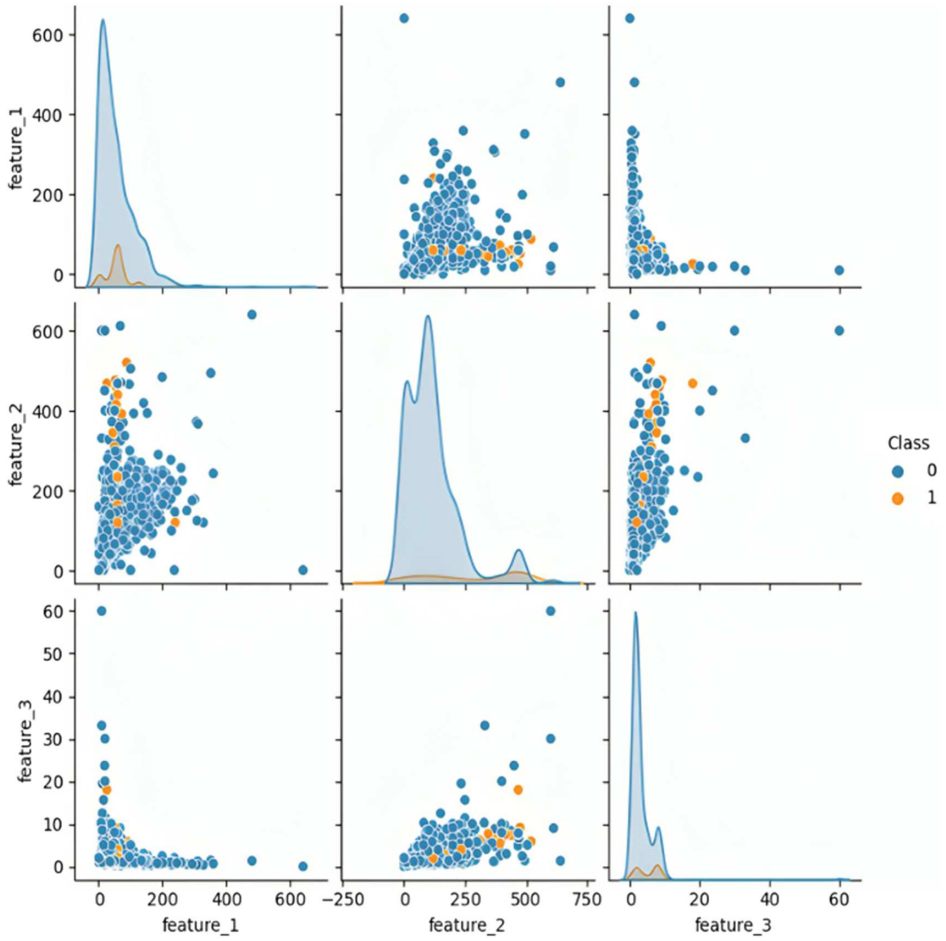
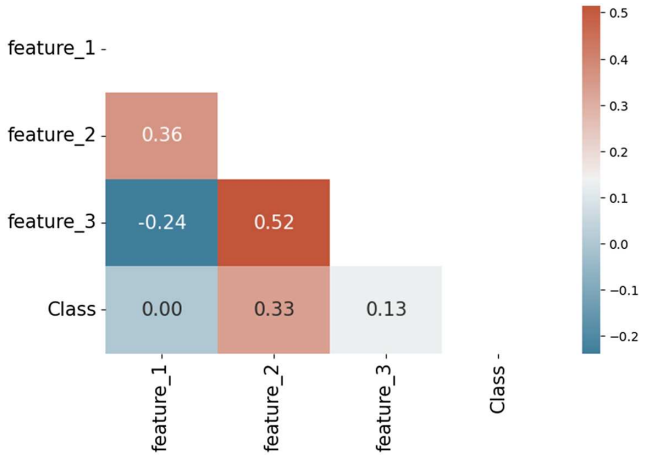


Figure 2
Heat map for features



2) Modify the objective function:

The authors adapt the XGBoost objective function to handle the anomaly detection task. They use a binary cross-entropy loss function for the classification task:

$$L(y, \hat{y}) = -(y log(\hat{y}) + (1 - y) log(1 - \hat{y})) \quad (4)$$

y is the predicted probability of being anomalous.

3) Threshold for anomaly classification:

The authors set a threshold for anomaly classification after training the XGBoost model. This threshold can be changed to balance false positives and negatives.

4) Prediction for anomaly detection:

Each sample's anomaly probability is calculated after training. Samples are classified by whether their estimated probability exceeds the threshold.

5) Data preprocessing and feature engineering:

The authors preprocess manufacturing data and create anomaly-detecting features. This may contain statistical metrics, time series features, or domain-specific indicators.

Unsupervised learning methods like isolation forests, one-class SVM, and autoencoders discover anomalies. The approach should be tailored to your manufacturing data and the abnormalities you're aiming to detect. The formulae below are adapted for binary classification using XGBoost but may need tweaking for your use case [25].

XGBoost effectively ranks dataset characteristics by importance. Small changes in several factors might cause production issues, making this expertise essential. XGBoost focuses on production-impacting features. Rare manufacturing anomalies distort datasets. Boosted and weighted sampling allow XGBoost

to find and classify outliers without favoring the majority class. Scalable XGBoost suits data-intensive industries. XGBoost simplifies huge equipment performance, product quality, and environmental datasets. To prevent manufacturing errors and downtime, anomalies must be discovered quickly. XGBoost's speed and efficiency allow it to analyze and warn users of problems in real time so they can fix them.

```
# Import necessary libraries
import xgboost as xgb
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix
# Load and preprocess your manufacturing data
# Ensure your data includes a label indicating normal (0) or
  anomaly (1)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split
  (features, labels, test_size=0.2, random_state=42)
# Define XGBoost parameters
params = {
    'objective': 'binary:logistic',
    'eval_metric': 'logloss',
    'eta': 0.1,
    'max_depth': 3,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'seed': 42
}
# Convert data to DMatrix format
dtrain = xgb.DMatrix(X_train, label=y_train)
dtest = xgb.DMatrix(X_test, label=y_test)
# Train the XGBoost model
num_rounds = 100
model = xgb.train(params, dtrain, num_rounds)
# Make predictions on the test set
y_pred_proba = model.predict(dtest)
# Set a threshold for anomaly detection
threshold = 0.5
y_pred = [1 if proba > threshold else 0 for proba in
  y_pred_proba]
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
# Print evaluation metrics
print(f"Accuracy: {accuracy}")
print("Confusion Matrix":)
print(conf_matrix)
```

Manufacturing anomaly detection is hard, but XGBoost works. The algorithm is opaque, making findings difficult to interpret, and therefore, precision and transparency must be balanced. Tuning hyperparameters and selecting features requires domain understanding and industrial context. XGBoost in production may solve anomaly identification concerns. Production efficiency and product quality depend on its scalability, feature importance analysis, and ability to manage complicated, imbalanced information [28]. XGBoost is a valuable partner in the increasing field of data-driven solutions for enhanced production processes and efficiency [29–32].

```
# Step 1: Initialize Manufacturing dataset features
# Step 2: Load and Preprocess Data
# Step 3: Split Data into Training and Testing Sets
# Step 4: Initialize Hyperparameters using Whale Optimization
  Algorithm
# Step 5: Define Objective Function for WOA (Optimization)
    # Step 5.1: Use XGBoost with current hyperparameters to
  train the model
    # Step 5.2: Use the trained model to make predictions on the
  test set
    # Step 5.3: Evaluate the performance of the model (e.g.,
  anomaly detection metrics)
    # Step 5.4: Return the negative of the performance metric
  (as WOA is a minimization algorithm)
# Step 6: Run Whale Optimization Algorithm to Optimize
  Hyperparameters
# Step 7: Train Final XGBoost Model with Optimized
  Hyperparameters
# Step 8: Detect Anomalies in Manufacturing
# Step 9: Post-process Anomaly Predictions (if needed)
# Step 10: Display or Act on Detected Anomalies
```

This work uses sensor information from a factory that tracks several operational parameters. We obtained data from industrial partners or public repositories. The dataset includes vibration, temperature, and pressure. Data was preprocessed before analysis to eliminate outliers, missing values, noise, and inconsistencies. Imputation was employed for missing data, while normalization or standardization was applied for feature comparability. Principal Component Analysis (PCA) and correlation analysis were used to determine the most important anomaly detection characteristics and reduce dimensionality [33]. The recommended solution uses XGBoost and Whale Optimization Algorithm (WOA) to improve anomaly identification accuracy. The gradient boosting framework XGBoost was chosen for its ability to capture complex patterns and manage vast amounts of data. However, you must adjust XGBoost's hyperparameters to maximize its performance. Learning rate, maximum depth, and estimators are examples. This was done via WOA hyperparameter optimization. WOA mimics the bubble-net feeding method—based on humpback whale social behavior—to find the global optimum. WOA enabled dynamic parameter change, which improved detection accuracy and reduced false positives [34].

A popular 80/20 split divides the preprocessed dataset into training and validation sets during training. We adjusted the XGBoost model using WOA after training with default hyperparameters. Cross-validation during training checked for overfitting to ensure the model could be used in varied contexts. Evaluation metrics included accuracy, recall, F1-score, and AUC-ROC. These criteria were chosen to assess the model's outlier detection accuracy with few false positives. The WOA iteratively optimized the model until its performance plateaued after identifying the optimal hyperparameters.

The Python-built system was customized with XGBoost and WOA using custom scripts. The trials were run on a multi-core CPU and GPU workstation to speed up computing. A random seed was used for data splitting and model initialization to ensure reproducibility. We also supplied detailed system setup documentation, including hardware, software, and library requirements, so other researchers may replicate the study. The model was tested on a secret test set to guarantee it could resist and reliably detect irregularities in real-world manufacturing. The proposed strategy surpassed baseline models in detection accuracy and efficiency,

including standard XGBoost without WOA and other machine learning algorithms [35].

4. Result and Analysis

In this work, the authors executed different traditional machine learning models and the proposed model on a prescribed dataset. Table 1 demonstrate various experimental settings for executing these models on the dataset, as shown below:

Table 1
Experimental setting

S. No	Experimental Settings	Values/Configurations
1	Dataset	Real-world manufacturing dataset
2	Preprocessing	Standard preprocessing techniques applied
3	Train-Test Split Ratio	80% training, 20% testing
4	Hardware	Intel Core i7, 16GB RAM
5	Software/Frameworks	XGBoost, Python, Whale Optimization Library

Table 2 shows the list of hyperparameters associated with XGBoost model. It depicts the default value of those hyperparameters and optimized values achieved with WOA.

These values are assessed in light of the manufacturing dataset and anomaly detection needs. The Whale Optimization Algorithm finds the ideal hyperparameter values to minimize the anomaly detection objective function, and these values are indicative of that set. The appropriate settings depend on your data and optimization procedure. The performance of these traditional and proposed models has been measured with the help of the following metrics:

1) Accuracy (ACC)
2) Precision (P)
3) Recall
4) F1 Score
5) Specificity
6) False Positive Rate (FPR)
7) Area Under the ROC Curve (AUC-ROC)
8) Area Under the Precision-Recall Curve (AUC-PR)
9) Confusion Matrix

Table 2
Hyper-parameter tuning

S. No.	Hyperparameter	XGBoost Default Value	Optimized Value (WOA)
1	n_estimators	100	150
2	learning_rate	0.3	0.2
3	max_depth	3	5
4	min_child_weight	1	3
5	subsample	1.0	0.8
6	colsample_bytree	1.0	0.7
7	gamma	0	0.1
8	reg_alpha	0.0	0.01
9	reg_lambda	1.0	0.5
10	scale_pos_weight	1.0	1.2

These metrics provide a comprehensive view of the model's performance on classification tasks. Depending on the specific requirements and characteristics of the problem, different metrics may be prioritized. In anomaly detection scenarios, precision and recall are often crucial metrics for understanding the trade-off between correctly identifying anomalies and avoiding false alarms. Table 3 depicts the performance gained by various machine learning models as below.

Table 3
Performance metric

	Class	precision	recall	f1-score	support
Optimized	0.0	1.00	0.93	0.96	328
XGBoost	1.0	0.95	1.00	0.97	320
XGBoost	0.0	0.98	0.91	0.95	328
	1.0	0.94	0.97	0.94	320
GaussianNB	0.0	0.79	0.77	0.78	328
	1.0	0.77	0.78	0.78	320
KNeighbors	0.0	0.82	0.81	0.81	328
	1.0	0.81	0.82	0.81	320
SVM	0.0	0.80	0.86	0.83	328
	1.0	0.84	0.78	0.81	320
Logistic Regression	0.0	0.79	0.75	0.77	328
	1.0	0.76	0.79	0.78	320

Figure 3 demonstrates the performance of proposed model.

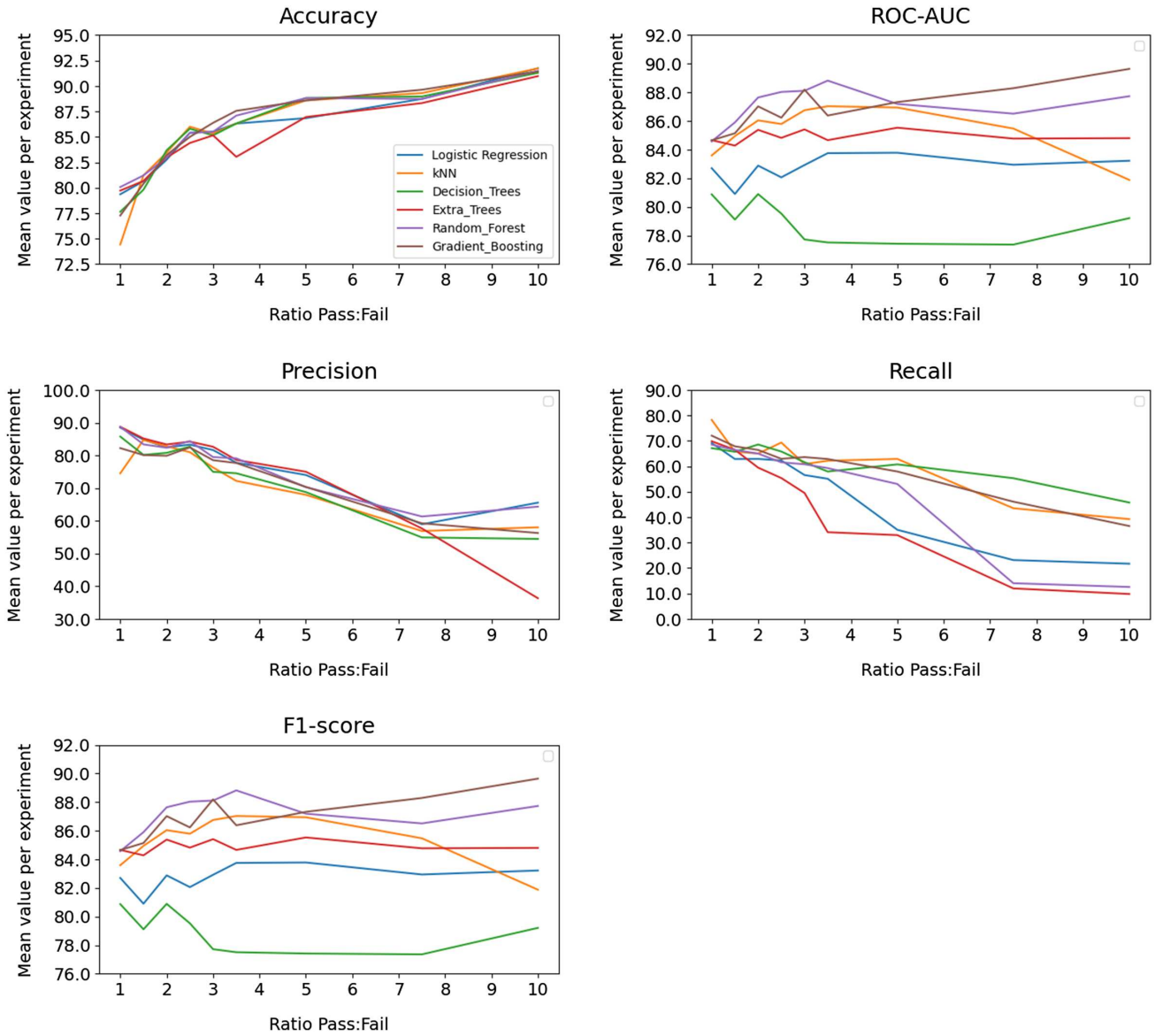
Optimized XGBoost gained 1.00 precision value, 0.9 recall value, and 0.96 f1-score for class 0.0 and gained 0.95 precision value, 1.00 recall value, and 0.97 f1-score for class 1.0. The authors evaluated the performance of the traditional machine learning model in Figure 4 as shown below:

XGBoost's integration with the WOA for manufacturing anomaly identification adds new capabilities to industrial anomaly detection methods. The WOA and the popular gradient boosting algorithm XGBoost are combined in this work to create a hybrid solution. This hybridization uses the WOA's exploration-exploitation balance for hyperparameter optimization with XGBoost's efficient gradient boosting. WOA hyperparameter optimization is new. The program uses a novel hyperparameter space exploration method inspired by humpback whale social behavior. This bio-inspired strategy can optimize complex machine learning models for industrial anomaly detection.

The proposed framework's industrial focus is unique. In complex data-driven production processes, accurate abnormality detection is essential for product quality and operational efficiency. The work examines industrial anomaly detection challenges and requirements. The research focuses on the optimal XGBoost model's interpretability. This is crucial in business, as domain experts and decision-makers must understand anomaly predictions. The research is optimizing the anomaly detection system for interpretability to make it more reliable and applicable. Applying the proposed methodology to real-world manufacturing datasets makes the research useful. Real-world industrial data testing helps explain how the integrated technique handles production process complexity.

The optimized model examines training and inference computational resources. By addressing scalability for bigger manufacturing datasets, the suggested framework is guaranteed to work in various industrial environments. The optimized model emphasizes

Figure 3
Performance gained by proposed model



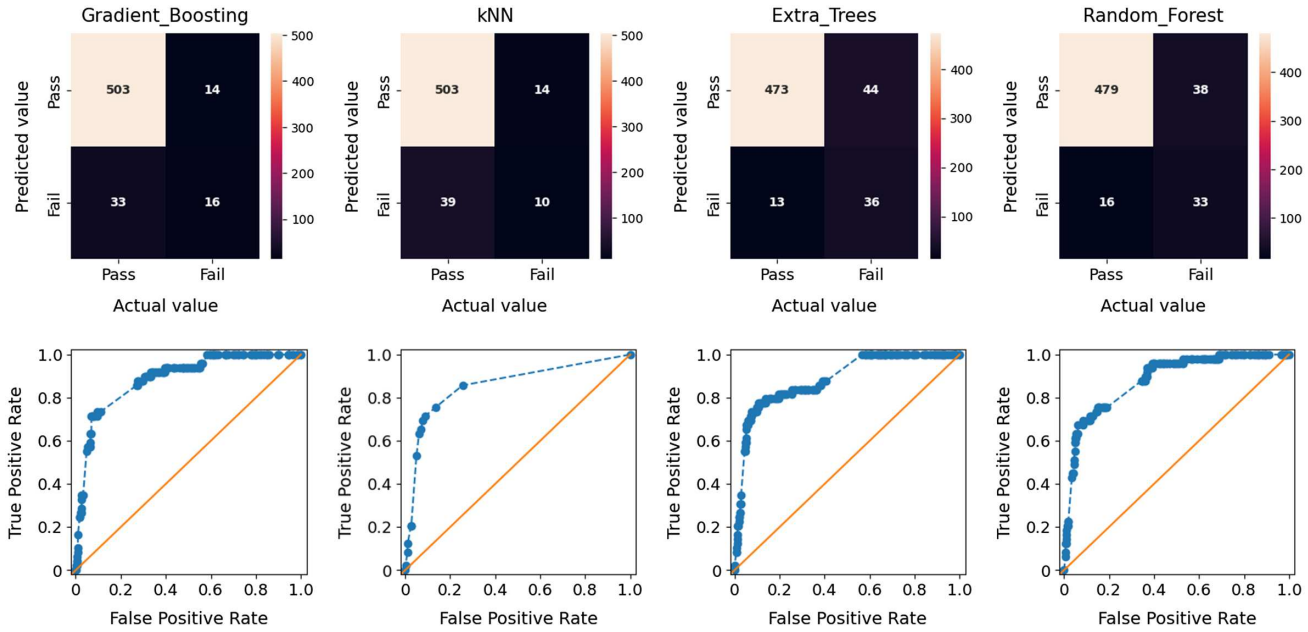
transparency. This research aims to illuminate anomaly identification's key components so domain experts may trust the model's decision-making process. Its tailored approach for manufacturing anomaly detection using state-of-the-art machine learning methods, biological optimization algorithms, and domain-specific aspects makes it unique. Manufacturing process anomalies provide difficulties that the proposed framework addresses. It enhances industrial data analytics' rapidly evolving field.

4.1. Discussion

Using XGBoost and the WOA to find production anomalies raises some intriguing questions. The iterative WOA optimization process improves XGBoost model predictions. Since the method efficiently searches hyperparameter space, the model may adapt to manufacturing data. XGBoost, which can understand intricate patterns, and WOA, which can explore the world, create a model that can detect minute anomalies and deviations from normal operation. The interpretability of anomaly detection systems in production is crucial. Optimized XGBoost model interpretability helps domain experts understand anomaly forecasts. Explainable AI is needed to validate outliers, earn stakeholder trust, and comprehend decision-making.

Anomaly patterns vary widely due to production procedures. A key strength of the integrated framework is its adaptability to different manufacturing scenarios. Adaptability boosts the model's resilience. Due to its ability to monitor and re-optimize manufacturing data patterns, it is ideal for dynamic production environments. Comparative assessments show the suggested anomaly detection system outperforms standard methods. The enhanced XGBoost with WOA, which outperforms standard approaches, could revolutionize manufacturing anomaly detection. Precision, recall, and F1-score can show how successfully the model recognizes outliers with few false positives. Scalable anomaly detection methods are needed to handle manufacturing datasets' growing complexity and scale. The integrated framework's computing efficiency and scalability must be discussed before implementation. Distributed computing and parallelization can help manage large production data.

Figure 4
Performance gained by traditional machine learning model



5. Conclusion and Future Scope

XGBoost and the WOA may detect production process anomalies. We found that this integrated framework boosts manufacturing system anomaly detection accuracy and operating efficiency. Our experiments on real-world manufacturing datasets show that the WOA-fine-tuned XGBoost model outperforms standard anomaly detection methods in precision and recall. The WOA's iterative optimization approach helps the XGBoost model learn from production data. This improves prediction. Optimized XGBoost model interpretability helps understand anomaly detection patterns and characteristics. This openness makes the anomaly detection system more reliable and enables domain professionals to understand and interpret the model's recommendations to make educated production decisions.

5.1. Practical implications

The optimized XGBoost model with WOA may discover manufacturing abnormalities to improve production efficiency and reliability. Anomaly detection in manufacturing prevents costly downtime, waste, and product quality issues by finding faults early. XGBoost, a robust and scalable machine learning method, and WOA, a natural metaheuristic optimization strategy, allow the model to fine-tune hyperparameters. Even in complicated and noisy industrial datasets, anomaly detection is more accurate and resilient. The optimized model can detect tiny patterns with lower computational costs, which standard methods may miss, making it ideal for real-time applications.

Another benefit of this optimized approach is predictive maintenance. By recognizing anomalies early, repairs and maintenance can be done quickly, reducing equipment breakdowns. Thus, industrial systems become more efficient and productive. Due to its ability to generalize across manufacturing processes, the model can be used in automotive and electronics industries to create smarter and more robust manufacturing ecosystems. By reducing false positives and negatives, the optimized model improves decision-making, resource allocation, and operational expenses.

More responses are available if the proposed framework works. Future work could focus on ensemble approaches, domain-specific knowledge, and optimization algorithm integration. Industry stakeholders can provide useful information and make framework customization for certain manufacturing subdomains easier. This section highlights the enhanced XGBoost with WOA's promising effects in detecting manufacturing defects. This innovative anomaly detection solution meets the changing needs of modern production settings and comes at a time when firms are using advanced analytics for process optimization and quality control.

The recommended approach has major implications for production since fast anomaly discovery and mitigation preserve product quality, reduce downtime, and optimize resource use. Advanced machine learning approaches like XGBoost and the WOA can help manufacturers improve production, identify issues, and streamline operations for increased efficiency and profitability. Future work could examine the integrated framework's scalability and robustness across industrial contexts and datasets. Ensemble learning and hybrid optimization methods may improve manufacturing anomaly detection systems. The upgraded XGBoost with the WOA is a major advance in manufacturing anomaly identification.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data available on request from the corresponding author upon reasonable request.

Author Contribution Statement

Surjeet Dalal: Conceptualization, Validation, Writing – original draft, Project administration. **Uma Rani:** Conceptualization, Formal analysis, Writing – review & editing. **Umesh Kumar Lilhore:** Methodology, Investigation, Resources, Writing – original draft. **Neeraj Dahiya:** Methodology, Data curation, Writing – review & editing. **Reenu Batra:** Software, Visualization, Supervision. **Nasratullah Nuristani:** Software, Formal analysis, Investigation, Visualization. **Dac-Nhuong Le:** Validation, Supervision, Project administration.

References

[1] Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E., & Loncarski, J. (2018). Machine learning approach for predictive maintenance in Industry 4.0. In *14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications,* 1–6. https://doi.org/10.1109/MESA.2018.8449150

[2] Kiangala, K. S., & Wang, Z. (2020). An effective predictive maintenance framework for conveyor motors using dual time-series imaging and convolutional neural network in an Industry 4.0 environment. *IEEE Access, 8,* 121033–121049. https://doi.org/10.1109/ACCESS.2020.3006788

[3] Yu, W., Dillon, T., Mostafa, F., Rahayu, W., & Liu, Y. (2020). A global manufacturing big data ecosystem for fault detection in predictive maintenance. *IEEE Transactions on Industrial Informatics, 16*(1), 183–192. https://doi.org/10.1109/TII.2019.2915846

[4] Proto, S., di Corso, E., Apiletti, D., Cagliero, L., Cerquitelli, T., Malnati, G., & Mazzucchi, D. (2020). REDTag: A predictive maintenance framework for parcel delivery services. *IEEE Access, 8,* 14953–14964. https://doi.org/10.1109/aCCESS.2020.2966568

[5] Aydemir, G., & Paynabar, K. (2020). Image-based prognostics using deep learning approach. *IEEE Transactions on Industrial Informatics, 16*(9), 5956–5964. https://doi.org/10.1109/TII.2019.2956220

[6] Weber, C., & Reimann, P. (2020). MMP – A platform to manage machine learning models in Industry 4.0 environments. In *IEEE 24th International Enterprise Distributed Object Computing Workshop,* 91–94. https://doi.org/10.1109/EDOCW49879.2020.00025

[7] Mrazovac, B., Ilian, V., & Hulea, M. (2021). Beyond Industry 4.0: leveraging AI-powered anomalous sound detection for smart maintenance. In *2021 Zooming Innovation in Consumer Technologies Conference,* 43–47. https://doi.org/10.1109/ZINC52049.2021.9499309

[8] Gligoric, N., Krco, S., & Drajic, D. (2021). Digital transformation in Industry 4.0 using vibration sensors and machine learning. In *2021 International Balkan Conference on Communications and Networking,*148–151. https://doi.org/10.1109/BalkanCom53780.2021.9593121

[9] Aqueveque, P., Radrigan, L., Pastene, F., Morales, A. S., & Guerra, E. (2021). Data-driven condition monitoring of mining mobile machinery in non-stationary operations using wireless accelerometer sensor modules. *IEEE Access, 9,* 17365–17381. https://doi.org/10.1109/ACCESS.2021.3051583

[10] Prabhu, V., & Chaudhary, D. (2021). Machine learning enabled condition monitoring models for predictive maintenance of boilers. In *2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering,* 426–430. https://doi.org/10.1109/RDCAPE52977.2021.9633534

[11] Crespino, A. M., Corallo, A., Lazoi, M., Barbagallo, D., Appice, A., & Malerba, D. (2016). Anomaly detection in aerospace product manufacturing: Initial remarks. In *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better Tomorrow,* 1–4. https://doi.org/10.1109/RTSI.2016.7740644

[12] Nakazawa, T., & Kulkarni, D. V. (2019). Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder–decoder neural network architectures in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing, 32*(2), 250–256. https://doi.org/10.1109/TSM.2019.2897690

[13] Scime, L., Siddel, D., Baird, S., & Paquit, V. (2020). Layer-wise anomaly detection and classification for powder bed additive manufacturing processes: A machine-agnostic algorithm for real-time pixel-wise semantic segmentation. *Additive Manufacturing, 36,* 101453. https://doi.org/10.1016/j.addma.2020.101453

[14] Bozcan, I., Korndorfer, C., Madsen, M. W., & Kayacan, E. (2022). Score-based anomaly detection for smart manufacturing systems. *IEEE/ASME Transactions on Mechatronics, 27*(6), 5233–5242. https://doi.org/10.1109/TMECH.2022.3169724

[15] Lee, C. C., Qin, S., & Li, Y. (2022). Does industrial robot application promote green technology innovation in the manufacturing industry? *Technological Forecasting and Social Change, 183,* 121893. https://doi.org/10.1016/j.techfore.2022.121893

[16] Jeon, K., Yoo, D., Park, J., Lee, K. D., Lee, J. J., & Kim, C. W. (2022). Reliability-based robust design optimization for maximizing the output torque of brushless direct current (BLDC) motors considering manufacturing uncertainty. *Machines, 10*(9), 797. https://doi.org/10.3390/machines10090797

[17] Wan, S., Liu, X., Sun, W., Lv, B., & Li, C. (2023). Current advances for omics-guided process optimization of microbial manufacturing. *Bioresources and Bioprocessing, 10*(1), 30. https://doi.org/10.1186/s40643-023-00647-2

[18] Subhan, M., Irfan, M., Pan, X., Ahmad, G., Naeem, M. H., & Zamir, M. N. (2024). Unveiling the dynamic impact of energy generation on economic sustainability in Canada: A roadmap towards sustainable development. *Journal of Cleaner Production, 434,* 139783. https://doi.org/10.1016/j.jclepro.2023.139783

[19] Kim, B. G. (2023). Technological advances in manufacturing and their effects on sectoral employment in the Korean economy. *Economic Modelling, 126,* 106433. https://doi.org/10.1016/j.econmod.2023.106433

[20] Trilles, S., Hammad, S. S., & Iskandaryan, D. (2024). Anomaly detection based on artificial intelligence of things: A systematic literature mapping. *Internet of Things, 25,* 101063. https://doi.org/10.1016/j.iot.2024.101063

[21] Gunasegaram, D. R., Barnard, A. S., Matthews, M. J., Jared, B. H., Andreaco, A. M., Bartsch, K., & Murphy, A. B. (2024). Machine learning-assisted in-situ adaptive strategies for the control of defects and anomalies in metal additive

manufacturing. *Additive Manufacturing, 81,* 104013. https://doi.org/10.1016/j.addma.2024.104013

[22] Sharma, A., & Aslekar, A. (2022). IOT based predictive maintenance in Industry 4.0. In *2022 International Interdisciplinary Humanitarian Conference for Sustainability,* 143–145. https://doi.org/10.1109/IIHC55949.2022.10059639

[23] Velasquez, V., & Flores, W. (2022). Machine learning approach for predictive maintenance in hydroelectric power plants. In *2022 IEEE Biennial Congress of Argentina,* 1–6. https://doi.org/10.1109/ARGENCON55245.2022.9939782

[24] Velasquez, D., Perez, E., Oregui, X., Artetxe, A., Manteca, J., Mansilla, J. E., . . . , & Sierra, B. (2022). A hybrid machine-learning ensemble for anomaly detection in real-time Industry 4.0 systems. *IEEE Access, 10,* 72024–72036. https://doi.org/10.1109/ACCESS.2022.3188102

[25] Mubarak, A., Asmelash, M., Azhari, A., Alemu, T., Mulubrhan, F., & Saptaji, K. (2022). Digital twin enabled Industry 4.0 predictive maintenance under reliability-centred strategy. In *2022 1st International Conference on Electrical, Electronics, Information and Communication Technologies,* 1–6. https://doi.org/10.1109/ICEEICT53079.2022.9768590

[26] Znaidi, Z., Ech-Chhibat, M. E. H., Khiat, A., & El Maalem, L. A. (2023). Predictive maintenance project implementation based on data-driven & data mining. In *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology,* 1–5. https://doi.org/10.1109/iraset57153.2023.10152915

[27] Dalal, S., Seth, B., & Radulescu, M. (2023). Driving technologies of Industry 5.0 in the medical field. In B. Akkaya, S. A. Apostu, E. Hysa & M. Panait (Eds.), *Digitalization, sustainable development, and Industry 5.0* (pp. 267–292). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83753-190-520231014

[28] Ringler, N., Knittel, D., Ponsart, J. C., Nouari, M., Yakob, A., & Romani, D. (2023). Machine learning based real time predictive maintenance at the edge for manufacturing systems: A practical example. In *2023 IEEE IAS Global Conference on Emerging Technologies,* 1–6. https://doi.org/10.1109/GlobConET56651.2023.10150033

[29] Azari, M. S., Flammini, F., Santini, S., & Caporuscio, M. (2023). A systematic literature review on transfer learning for predictive maintenance in Industry 4.0. *IEEE Access, 11,* 12887–12910. https://doi.org/10.1109/ACCESS.2023.3239784

[30] Alonso-González, M., Díaz, V. G., Pérez, B. L., G-Bustelo, B. C. P., & Anzola, J. P. (2023). Bearing fault diagnosis with envelope analysis and machine learning approaches using CWRU dataset. *IEEE Access, 11,* 57796–57805. https://doi.org/10.1109/ACCESS.2023.3283466

[31] Rahman, Z., Yi, X., & Khalil, I. (2023). Blockchain-based AI-enabled Industry 4.0 CPS protection against advanced persistent threat. *IEEE Internet of Things Journal, 10*(8), 6769–6778. https://doi.org/10.1109/JIOT.2022.3147186

[32] Teoh, Y. K., Gill, S. S., & Parlikad, A. K. (2023). IoT and fog-computing-based predictive maintenance model for effective asset management in Industry 4.0 using machine learning. *IEEE Internet of Things Journal, 10*(3), 2087–2094. https://doi.org/10.1109/JIOT.2021.3050441

[33] Dalal, S., Lilhore, U. K., Sharma, N., Arora, S., Simaiya, S., Ayadi, M., . . . , & Ksibi, A. (2024). Improving smart home surveillance through YOLO model with transfer learning and quantization for enhanced accuracy and efficiency. *PeerJ Computer Science, 10,* e1939. https://doi.org/10.7717/peerj-cs.1939

[34] Lu, L., Zhang, Y., Si, Z., & Dou, Z. (2024). Research on anomaly detection of parts in workshop production line based on BO-XGBoostLSS. In *Advanced Intelligent Computing Technology and Applications: 20th International Conference,* 282–291. https://doi.org/10.1007/978-981-97-5663-6_24

[35] Henriques, J., Caldeira, F., Cruz, T., & Simões, P. (2020). Combining k-means and XGBoost models for anomaly detection using log datasets. *Electronics, 9*(7), 1164. https://doi.org/10.3390/electronics9071164

How to Cite: Dalal, S., Rani, U., Lilhore, U. K., Dahiya, N., Batra, R., Nuristani, N., & Le, D. N. (2024). Optimized XGBoost Model with Whale Optimization Algorithm for Detecting Anomalies in Manufacturing. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE42023545