Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 13 Issue 2, February - 2026

Machine Learning-Based Prediction Of Industrial Waste Accumulation For Green Manufacturing Implementation In Nigerian Cement And Feed Mill Industries

Sylvester Akwagiobe UNDIANDEYE[1]
Department of Mechanical and Aerospace Engineering,
Faculty of Engineering, University of Uyo, Akwa Ibom State, Nigeria
sublimedigipress@gmail.com

Aniekan OFFIONG[2]
Department of Mechanical and Aerospace Engineering,
Faculty of Engineering, University of Uyo, Akwa Ibom State, Nigeria

Simeon OZUOMBA[3]
Department of Computer Engineering,
Faculty of Engineering, University of Uyo, Akwa Ibom State, Nigeria
simeonozuomba@uniuyo.edu.ng

Corresponding Author's Email: sublimedigipress@gmail.com

ABSTRACT

Industrial waste mismanagement in Nigerian manufacturing industries remains a significant environmental and economic challenge. Despite the adoption potential of green manufacturing (GM), practical predictive tools for planning waste management under GM scenarios are largely absent, particularly for cement and feed mill industries. This study aimed to develop and evaluate a preliminary machine-learning-based model for predicting industrial waste accumulation under conventional manufacturing (WoGM) and expected green manufacturing (WGM) scenarios in selected Nigerian cement and feed mill industries, incorporating Generative Adversarial Networks (GANs)-based data augmentation to address the paucity of operational data. A descriptive case study design was adopted involving Industry A (Cement) and Industry B (Feed Mill) in South-south Nigeria. Data on raw material inputs, products, by-products, and waste streams were collected via primary and secondary sources. Instrument content validity index (CVI = 0.974) and reliability (r = 0.9983) were confirmed. Analytical material balance models and a Random Forest Regressor were developed. A Conditional GAN (cGAN) with Wasserstein-GP (WGAN-GP) loss was employed to expand the dataset from 12 monthly observations to 1,000 records per waste stream, preserving statistical fidelity validated via Kolmogorov-Smirnov tests. Data preprocessing included min-max normalisation and feature engineering prior to model training. The original-data Random Forest model yielded R^2 ranging from 0.8819 to 0.9900, with MBE from -0.083 to 0.30 tons, and RMSE from 0.04 to 32 tons across all waste streams. One waste stream (fish feed, WoGM) returned an unusually high R^2 of 0.9999, attributed to low process variability in that stream's 12-month operational window and discussed as a preliminary finding requiring multi-year validation. GAN-augmented model training reduced RMSE by 12-23% across waste streams, and raised

the weakest stream's R^2 from 0.8819 to 0.9480. Expected GM adoption was projected to reduce cement waste accumulation by 86.63% and feed mill waste by 44-89% across product streams. The preliminary Random Forest model demonstrated strong predictive capability under the evaluated conditions. GAN-based data augmentation substantially improved model robustness and generalisation. However, results should be interpreted as preliminary pending validation with multi-year, multi-plant datasets. The combined analytical-AI-GAN framework provides a replicable template for data-driven green manufacturing planning in Nigerian industries.

KEYWORDS: *Machine learning; GAN-based data augmentation; Waste accumulation; Green manufacturing; Random Forest Regressor; Conditional GAN; Nigeria; Circular economy*

1.0 INTRODUCTION

Green manufacturing (GM), also known as ecologically conscious manufacturing (ECM), denotes a paradigm shift in production that integrates eco-efficient policies and process innovations to minimise environmental burden (Ahmed, 2011; Jasiulewicz, 2014). It emphasises reducing material and energy consumption, substituting toxic inputs with safer alternatives, and converting waste outputs into reusable inputs through recycling and reprocessing (Ping and Gang, 2016; Rosen and Kishawy, 2012). The approach seeks to remove environmentally harmful wastes and energy inefficiencies while endorsing eco-friendly packaging, supply chain management, and post-use disposal or recycling (Rehman et al., 2016; Ogbo et al., 2017). GM further connects natural energy flows and biological processes to reduce reliance on fossil fuels and hazardous constituents, thereby boosting resource efficiency (Baines et al., 2012).

In Africa, manufacturing industries contribute significantly to carbon emissions, with four economies, South Africa, Algeria, Egypt, and Nigeria, accounting for approximately 75% of the continent's industrial-related carbon dioxide emissions (Barnes et al., 2019). Nigeria's industrial sector, despite government interventions, continues to struggle with weak performance and high ecological degradation, predominantly from cement and feed mill industries through unmanaged waste streams (Ogboani et al., 2023). It is therefore imperative that Nigerian cement and feed mill industries adopt GM approaches, particularly in post-production processes, to align with international sustainability goals and reduce their environmental burden (Banjoko et al., 2012).

A review of the literature reveals that closely related studies have been conducted mainly in plastics, paint, food processing, barrier coating, and software industries in Kenya, China, Malaysia, and Ghana (Cherrafi et al., 2017; Abhijeet et al., 2017; Singh et al., 2018; Acquah et al., 2021). Only limited studies addressed the cement sector, primarily in China, to determine key success factors for green manufacturing system implementation (Yuan, 2021; Zheming et al., 2022). Nigerian-focused studies have examined green manufacturing adoption in selected firms (Okunuga et al., 2022; Nwaulune, 2024; Parthiban et al., 2024), but none combined industrial waste mass balance analysis with machine learning prediction for cement and feed mill industries, nor addressed the challenge of small operational datasets.

Recent advances in industrial artificial intelligence have expanded the scope of sustainability analytics in manufacturing. Zhao et al. (2024) demonstrated the utility of explainable AI in industrial waste classification, while Liu et al. (2023) applied deep learning for cement kiln process optimisation. Rajput and Singh (2024) explored diffusion-based synthetic data generation for low-data manufacturing environments, and Fonseca et al. (2024) reported that GAN-based tabular augmentation can improve regression accuracy by 15-30% in small-sample industrial datasets. These developments establish a strong precedent for the methodological approach adopted in the present study.

Existing studies focus broadly on green manufacturing adoption frameworks, but few have combined industrial waste mass balance analysis with machine learning prediction specifically for Nigerian cement and feed mill industries. Furthermore, operational datasets from industrial case studies are typically small, often fewer than 20 observations, limiting machine learning model reliability. No prior Nigerian study has applied GAN-based synthetic data augmentation to overcome dataset paucity while predicting waste accumulation under both WoGM and WGM scenarios.

The aim of this study is to develop and evaluate a preliminary machine-learning-based model for predicting industrial waste accumulation under conventional manufacturing and expected green manufacturing scenarios in selected Nigerian cement and feed mill industries, with GAN-based data augmentation employed to expand the operational dataset to 1,000 records, thereby improving model robustness and generalisation.

This study provides insights into the quantity of waste that could be channelled to other firms from both Industry A and Industry B, demonstrates how a preliminary AI model can predict waste accumulation under both WoGM and WGM scenarios, and establishes a replicable analytical-AI-GAN framework for green manufacturing planning in Nigerian industries.

2.0 MATERIALS AND METHODS

2.1 Research Design and Target Population

This study adopted a descriptive case study design involving a cement industry, designated Industry A, and a feed mill industry, denoted Industry B, their identities are withheld for confidentiality, located in South-south Nigeria. The target population comprised one (1) general/operational manager, two (2) engineers, and two (2) operators each in the selected industries, giving a total of ten (10) expected participants.

2.2 Data Collection and Instrument Validation and Reliability

Data were gathered using primary (questionnaire) and secondary sources. The questionnaire was designed to elicit information about products, by-products, waste streams, and resource usage under WoGM and WGM conditions. Two subject-matter experts independently assessed each item for conformity with the research objectives. Content validity was quantified using the Content Validity Index (CVI):

$$CVI = \frac{Number\ of\ items\ rated\ relevant}{Total\ number\ of\ items\ on\ the\ instrument} \quad \textbf{(1)}$$

A CVI value of 0.80 or above indicates acceptable content validity, while 0.90 or above indicates excellent validity (Beebwa, 2007). The present instrument yielded CVI = 0.974, indicating that 97.4% of questionnaire items were rated relevant by both experts.

Instrument reliability was assessed using Pearson's Product Moment Correlation Coefficient (r), computed in Microsoft Excel 2019, by administering the questionnaire to a sub-sample of participants at two separate time points (test-retest reliability). This approach was appropriate because the same instrument was administered twice to assess temporal stability of responses, rather than evaluating co-variance among multiple items in a single administration. Conventional practice recommends Cronbach's alpha for multi-item internal consistency, and future studies with larger instruments should apply it accordingly. Both CVI and r values greater than 0.7 are considered acceptable (Beebwa, 2007).

2.3 Ethical Issues and Instrument Administration

Participants' informed consent was obtained prior to questionnaire administration, and their privacy and anonymity were maintained throughout. The study was conducted in compliance with the ethical guidelines of the University of Uyo, Faculty of Engineering Research Ethics Committee. Industry identities were withheld to preserve commercial confidentiality, consistent with standard practice in applied industrial research. Completed questionnaires were collected for analysis. Of the eight (8) copies administered in the field, seven (7) were returned and completed.

2.4 Data Collection

Data encompassed: raw material and additive inputs (F, tons), fuel inputs (f, tons), product outputs (P, tons), and waste generated (W, tons) recorded monthly for a specific year from both industries. Each industry's major products, by-products, waste stream characterisation, distances of potential waste-

receiving companies, utilizable waste quantities, and post-processing product types were also documented. For Industry A, wastewater and treatment sludge (W3) was converted from volumetric to mass units using:

$$W3\ (tons) = W3\ (m^3) \times \rho w3\ \left(1.2 \frac{tons}{m^3}\right) \quad \textbf{(2)}$$

This density value is consistent with documented sludge densities for cement wastewater treatment effluents (Rusanescu et al., 2022).

2.4.1 Formulation of an Analytical Model for Waste Accumulation WoGM and WGM Implementations

Material balance analysis was employed to develop analytical models for waste accumulation in both scenarios. With the data gathered from Industry A, the material balance boundary is illustrated in Figure 1, where F = raw material and additive feed (tons), f = fuel (tons), P = product (tons), and W = waste generated (tons). Note that W may include cement kiln dust, wastewater sludge, rejected clinker, raw material waste, and gaseous emissions, some of which are recyclable and others non-recyclable.

Figure 1: Cement manufacturing system boundary for material balance

Employing standard material balance analysis (Assian et al., 2023), the material inputs and outputs are expressed as:

$$Inputs = Outputs \quad \textbf{(3a)}$$

$$F + f = P + W \quad \textbf{(3b)}$$

When there is no waste accumulation (Wacc) in the system, Equation (3) yields:

$$Wacc = (F + f) - (P + W) = 0 \quad \textbf{(4)}$$

When material accumulates within the system boundary:

$$Wacc = (F + f) - (P + R) \quad \textbf{(5)}$$

where R (or W) = quantity of waste deliberately discarded, recycled, or contributing to environmental burden (tons). Equation (5) constitutes the analytical model for WoGM in Industry A.

Under expected green manufacturing adoption (WGM) for Industry A, non-recyclable waste fractions, specifically cement kiln dust (W1) and clinker/cement dust (W2), are projected to be collected by nearby companies for reuse, while wastewater sludge (W3) is assumed to be partly recycled internally and partly transferred to external companies. It must be clearly stated that WGM values represent estimated quantities based on expected reuse and recycling assumptions, derived from documented utilisation capacities of identified waste-receiving industries, and not directly observed post-implementation measurements. The analytical model for WGM adoption in Industry A is:

$$Wacc^* = (F + f) - (P + Rcem.Int + Ro.coy + Wt1 + Wt2) \quad \textbf{(6a)}$$

where: Wacc* = residual waste accumulation under WGM (tons); Rcem.Int = wastewater sludge recycled internally by the cement plant (tons); Ro.coy = wastewater sludge recycled by external companies (tons); Wt1 and Wt2 = non-recyclable waste quantities transferred to and reused by external companies (tons).

For Industry B, the analytical model for WoGM is Equation (5), and for WGM:

$$Wacc^* = (F + f) - (P + WR + Wo.coy) \quad \textbf{(6b)}$$

where: WR = quantity of waste recycled or sold by Industry B (tons); Wo.coy = quantity recycled by external companies (tons).

2.4.2 Data Preprocessing and Feature Engineering

Prior to model training, all process variables underwent systematic preprocessing to ensure numerical stability and comparability across waste streams. The preprocessing pipeline consisted of the following steps.

All continuous process variables (F, f, P, W, waste fractions, Wacc) were min-max normalised to the interval [0, 1] using:

$$x_{norm} = \frac{(x - x_{min})}{(x_{\max} - x_{min})} \quad \textbf{(7)}$$

where x is the original variable value, x_min and x_max are the minimum and maximum observed values across the training set, respectively. This normalisation ensured stable gradient flow during GAN training and prevented feature dominance due to scale differences between, for example, the large-scale cement inputs (F ~ 600,000 tons/month) and the smaller feed mill variables.

The categorical WoGM/WGM scenario flag was one-hot encoded as a binary indicator variable. Production-level brackets, derived from monthly production volumes segmented into three quantile ranges (low, medium, high), were also one-hot encoded. These categorical encodings were appended to the feature matrix as additional predictor columns before both GAN training and Random Forest fitting, ensuring that the model could distinguish scenario-specific waste patterns.

A Pearson correlation matrix was computed across all feature variables to assess multicollinearity. No feature pair exhibited a correlation exceeding 0.92 within the same scenario, confirming that the selected features provided distinct informational contributions. Fuel input (f) showed the lowest variance and was retained because its physical role in the material balance is non-trivial, even though its predictive importance was relatively low (as confirmed in the feature importance analysis in Section 3.1.6). No outlier removal was performed, because the monthly operational data were validated against production records and no anomalous values were identified.

2.4.3 Machine Learning Model (Random Forest Regressor)

Based on the analytical models (Equations 5, 6a, 6b), a Random Forest Regressor was developed to predict waste accumulation (Wacc and Wacc*) under both scenarios. The Random Forest Regressor predicts a continuous target by averaging outputs of multiple decision trees, each trained on random bootstrapped subsets of data and features (Breiman, 2001; Hastie et al., 2009):

$$\hat{y}(x) = \left(\frac{1}{N}\right) \times \Sigma\, i = 1\ to\ N\ \ Ti(x) \quad \textbf{(8)}$$

where: N = number of decision trees; Ti(x) = prediction from the i-th tree for input vector x (comprising F, f, P, W, WR, Wo.coy, Wt1, Wt2, scenario flag, etc.); ŷ(x) = final ensemble prediction (Wacc or Wacc*).

For each leaf node:

$$Ti(x) = \left(\frac{1}{nleaf}\right) \times \Sigma\, j = 1\ to\ nleaf\ \ yj \quad \textbf{(9)}$$

where nleaf = number of training samples in the leaf node containing x; yj = target waste accumulation values of those samples. Combining Equations (8) and (9):

$$\hat{y}(x) = \left(\frac{1}{N}\right) \times \Sigma\, i = 1\ to\ N\ \left[\left(\frac{1}{nleaf}, i(x)\right) \times \Sigma\, j = 1\ \ yj, i(x)\right] \quad \textbf{(10)}$$

Model Hyperparameters and Optimisation:

The Random Forest Regressor was implemented in Python using Scikit-learn (Pedregosa et al., 2011). A systematic hyperparameter search using GridSearchCV with 5-fold cross-validation was performed over the following parameter grid: n_estimators in {100, 200, 300}, max_depth in {5, 10, None}, min_samples_split in {2, 5, 10}, and max_features in {'sqrt', 0.5, 1.0}. The 'sqrt' option was adopted following deprecation of the 'auto' keyword in Scikit-learn v1.1 and beyond (Pedregosa et al., 2011). Optimal hyperparameters identified through GridSearchCV were: n_estimators = 200, max_depth = None, min_samples_split = 2, max_features = 'sqrt', random_state = 42 (for reproducibility). On the original 12-sample dataset, a 75/25 train/test split with Leave-One-Out Cross-Validation (LOOCV) was used to maximise data utility. On the GAN-augmented 1,000-record dataset, an 80/20 split with 5-fold cross-validation was applied.

Figure 2: Characteristic Python algorithm for Random Forest Regressor for cement waste (WoGM) adoption

2.4.4 Generative Adversarial Networks (GANs)-Based Data Augmentation

2.4.4.1 Background and Rationale

The original dataset comprised only 12 monthly observations per waste stream, which far below the volume generally recommended for reliable Random Forest training. Such a small sample increases the risk of overfitting, artificially inflated in-sample R^2 values, and poor generalisation to unseen operating conditions. While LOOCV partially addresses this concern on the original data, it cannot substitute for a genuinely larger and more diverse training corpus.

Accordingly, Generative Adversarial Networks (GANs) were therefore employed to generate synthetic tabular records that augment the original dataset while preserving its underlying statistical distributions. GANs have demonstrated effectiveness for synthetic data generation from small industrial datasets (Goodfellow et al., 2014; Xu et al., 2019; Park et al., 2022; Fonseca et al., 2024). Unlike simple noise injection, which was tested but found to violate material balance constraints in 7.3% of generated cases, or SMOTE-based augmentation, which showed poor correlation preservation, the GAN approach generates coherent multivariate records that respect interdependencies among F, f, P, W, and waste fraction variables.

2.4.4.2 Conditional GAN (cGAN) Architecture

A Conditional GAN (cGAN) architecture was selected because industrial waste generation is conditioned on production inputs (F, f, P) and scenario type (WoGM/WGM). Conditioning on these variables ensures generated records are physically plausible for each scenario rather than sampled from a blended distribution. The cGAN comprises two competing neural networks.

Generator (G): Takes noise vector z ~ N(0, I) concatenated with condition vector c (encoding WoGM/WGM flag and production bracket), and maps to synthetic record x = G(z, c). Architecture: four fully connected layers with ReLU activations and batch normalisation; linear output activation.

Discriminator (D): Takes a real or synthetic record x concatenated with c and outputs probability D(x, c) of the record being real. Architecture: four fully connected layers with LeakyReLU activations and dropout (rate = 0.3).

The cGAN training objective is:

$$\min G \max D\, V(D,G) = Ex[\log D(x|c)] + Ez\left[\log\left(1 - D\left(G(z|c)\right)\right)\right] \quad \textbf{(11)}$$

where: x ~ pdata(x) are real waste records; z ~ pz(z) is noise; c is the condition vector. The Discriminator D maximises its ability to distinguish real from synthetic records, while Generator G minimises the Discriminator's success, converging to an adversarial equilibrium that produces increasingly realistic synthetic data.

The Wasserstein GAN with Gradient Penalty (WGAN-GP) loss was adopted to prevent mode collapse and stabilise training (Arjovsky et al., 2017):

$$LWGAN - GP = Ex[D(x)] - Ex[D(\tilde{x})] + \lambda \times Ex^{\left[\left(\left\|\nabla D(x\ )\right\|2 - 1\right)^2\right]} \quad \textbf{(12)}$$

where: $\tilde{x} = G(z|c)$ is a generated sample; x^ = interpolated sample between real and generated data; $\lambda = 10$ (gradient penalty coefficient); $||\nabla D(x^)||2$ = gradient norm of D with respect to the interpolated input. The Lipschitz constraint enforced by the gradient penalty term stabilises Wasserstein distance estimation and prevents training divergence.

2.4.4.3 Implementation and Augmentation Procedure

The cGAN was implemented in Python using PyTorch. Training ran for 5,000 epochs, with learning rate = 0.0002 (Adam optimiser, $\beta1 = 0.5$, $\beta2 = 0.999$), batch size = 12 (matching original dataset size). The Discriminator was updated five times per Generator update, consistent with WGAN-GP recommendations. A critical design decision involved the sequencing of the train-test split relative to

GAN augmentation: the real dataset was first split into training (75%) and test (25%) partitions, and the cGAN was trained exclusively on the training partition. Synthetic records were generated only for the training subset and then merged with the real training records to form the augmented training set. The test partition consisted solely of real observations and was never exposed to the GAN, thereby preventing data leakage and ensuring that validation metrics reflected genuine out-of-sample performance. The augmentation procedure followed six steps:

Step 1, Data Normalisation: All process variables (F, f, P, W, waste fractions, Wacc) were min-max normalised to [0, 1] before GAN training to ensure stable gradient flow, using inverse scaling after generation to restore physical units.

Step 2, Conditional Encoding: WoGM/WGM scenario flags and production-level brackets were one-hot encoded as condition vectors c, appended to both Generator and Discriminator inputs.

Step 3, GAN Training: The cGAN was trained on the original training-partition records per stream, with records from all waste streams and scenarios concatenated. Training convergence was confirmed by monitoring Wasserstein distance, which stabilised after approximately 3,500 epochs.

Step 4, Synthetic Record Generation: After training, 988 additional synthetic records per waste stream and scenario were generated, expanding each stream's training dataset from 9 (75% of 12) to 997 records, and the total training set to approximately 1,000 records per stream.

Step 5, Material Balance Validation: Each synthetic record was validated against material balance constraints (Equations 3-6). Records violating Wacc < 0 or Wacc > (F + f - P) were rejected and regenerated until the full quota was satisfied.

Step 6, Statistical Fidelity Testing: Kolmogorov-Smirnov (KS) tests confirmed $p > 0.05$ for all variables. Pearson correlation matrix similarity > 0.92 confirmed preservation of inter-variable relationships. Additionally, Principal Component Analysis (PCA) was used to compare the 2D projections of real and synthetic data; the distributions showed substantial overlap across all streams, confirming distributional alignment. Wasserstein distance between real and synthetic marginal distributions was below 0.05 for all continuous features, providing further confidence in synthetic data quality.

2.4.4.4 GAN-Augmented Training and Validation Strategy

The augmented training dataset was used for Random Forest Regressor retraining. An 80/20 train/test split (800 training, 200 testing records) was applied with 5-fold cross-validation on the training set. Performance metrics (R^2, MAE, MBE, RMSE, MAPE, Adjusted R^2) were computed on the held-out 20% test set and compared against baseline metrics from the original 12-sample LOOCV model. Prediction intervals were estimated via bootstrap resampling (1,000 iterations), providing 95% confidence bounds for each predicted waste accumulation value.

2.4.5 Model Verification and Validation

Model verification and validation used scatter plots of actual versus predicted waste accumulation values (Spiegel and Stephens, 1999), along with a comprehensive suite of statistical metrics. The coefficient of determination (R^2) (Frank and Althoen, 1995), mean bias error (MBE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and Adjusted R^2 were computed. Adjusted R^2 accounts for the number of predictor variables (k) relative to sample size (n):

$$Adjusted\ R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{(n-k-1)}\right] \quad \textbf{(13)}$$

The standard expressions for MBE and RMSE are:

$$MBE = (1/O) \times \Sigma\, i = 1\ to\ O\ (MRprd - MRact) \quad \textbf{(14)}$$

$$RMSE = \sqrt{\left[\left(\frac{1}{O}\right) \times \Sigma\, i = 1\ to\ O\ (MRprd - MRact)^2\right]} \quad \textbf{(15)}$$

$$MAE = \left(\frac{1}{O}\right) \times \Sigma\, i = 1\ to\ O\ |MRprd - MRact| \quad (16)$$

$$MAPE = \left(\frac{100}{O}\right) \times \Sigma\, i = 1\ to\ O\ \frac{|MRprd - MRact|}{MRact} \quad (17)$$

where: O = number of observations; MRprd = predicted waste accumulation (tons); MRact = actual waste accumulation (tons). A positive MBE indicates systematic over-prediction; negative MBE indicates under-prediction. For an accurate goodness of fit, R^2 should approach 1.0, and both MBE and RMSE should be small relative to the range of observed values (Demir et al., 2004; Arumuganathan et al., 2009). All computations were embedded in the Python code.

3.0 RESULTS AND DISCUSSION

3.1 Results

3.1.1 Sample Population, Content Validity, Reliability, Response and Return Rate

The sample population, content validity, instrument reliability, response and return rate are presented in Table 1.

Table 1: Sample population, content validity, reliability, response and return rate

Item	Quantity
Sample population	10 (8: field study; 2: reliability sub-sample)
Content validity index (CVI)	0.974
No. copies of questionnaire reproduced	12
No. copies used for reliability test	4
Instrument reliability (r)	0.9983
No. copies taken to the field	8
No. copies returned and responded	7
Response and return rate (%)	87.5

From Table 1, the sample population comprised 10 respondents (8 for field study; 2 as reliability sub-sample). The CVI of 0.974 indicates that 97.4% of questionnaire items were rated relevant by both experts, reflecting strong content validity. The instrument reliability (r = 0.9983) indicates high consistency across two administrations, confirming the questionnaire was stable and unambiguous. Of 8 copies distributed to field respondents, 7 were returned and completed, a return rate of 87.5%, well exceeding the 50% threshold recommended by Kumar (2010).

3.1.2 Industries' Major Products and By-products of Each Waste Stream

The major product of Industry A is cement. Its primary waste streams are cement kiln dust (W1), clinker/cement dust and airborne particulates (W2), and wastewater and treatment sludge (W3). Industry B produces poultry feeds, fish feeds, and concentrates, with corresponding waste streams: poultry feed wastes (X1), fish feed wastes (X2), and concentrate wastes (X3). These findings align with documented waste profiles for cement manufacturing (WBCSD, 2013; Pacific Cement, 2018; Michael et al., 2020) and for livestock and agroprocessing production systems (Shamsi et al., 2012; Ominski et al., 2021).

3.1.3 Distances of Present and Close Companies, Utilizable Waste Quantities, Products Obtainable After Post-Processing, and Monthly Waste Generation Data

The distances of waste-receiving companies, utilizable waste quantities, and post-processing products for both industries are presented in Table 2. Monthly waste generation and accumulation data under WoGM and WGM scenarios for Industries A and B are given in Tables 3 to 6.

Table 2: Distances of present and close companies, utilizable waste quantities, and products obtainable after post-processing, Industries A and B

INDUSTRY A, Cement Industry				
Waste Stream	**Waste-Receiving Companies**	**Distance (km)**	**Utilizable Qty (tons/yr)**	**Products After Post-Processing**
W1 (CKD)	Local Block & Brick Manufacturers	~20	374,700 (48.9%)	CKD bricks/blocks, lightweight aggregate, road subgrade, landfill materials
W1 (CKD)	Soil Stabilisation / Civil Engineering Companies	~20	196,000 (25.6%)	Road subgrade stabilisation materials
W1 (CKD)	Industrial Partners	~20	196,000 (25.6%)	Various industrial and construction applications
W2 (Clinker/Cement Dust)	Precast Concrete Producers	~40	552,900 (58.7%)	Masonry mortar, cementitious filler, kerbs and blocks
W2 (Clinker/Cement Dust)	Road Construction / Asphalt Manufacturers	~40	368,100 (41.3%)	Road construction materials and asphalt filler
W3 (Wastewater & Sludge)	Cement Plant (internal recycling)	0	829,100 (58.5%)	Construction curing water, reusable process water
W3 (Wastewater & Sludge)	Land Reclamation / Landscaping Companies	~40	231,583 (16.4%)	Non-potable irrigation water, landscaping
W3 (Wastewater & Sludge)	Dust Suppression / Construction Firms	~40	355,900 (25.1%)	Dust suppression, construction water supply
Note: W1 total = 766,700 tons/yr; W2 total = 891,000 tons/yr; W3 total = 1,416,583 m³/yr × 1.2 ton/m³ = 1,699,900 tons. WGM values represent expected utilisation based on documented receiving-industry capacities, not directly measured post-implementation data.				
INDUSTRY B, Feed Mill Industry				
Waste Stream	**Waste-Receiving Companies**	**Distance (km)**	**Utilizable Qty (tons/yr)**	**Products After Post-Processing**
X1 (Poultry Waste)	Farmers, livestock keepers, recyclers, compost makers, biogas producers, small-scale processors	~35	8,610 (38.1%)	Organic fertiliser, compost, biogas, low-grade animal feed, feather meal, fuel materials
X2 (Fish Waste)	Farmers, compost makers, biogas producers, aquaculture operators, energy entrepreneurs, small industries	~40	12,410 (54.9%)	Low-grade feed, compost, biogas, renewable energy, recycled plastics, liquid fertiliser, irrigation water
X3 (Concentrate Waste)	Firms/individuals listed for X1 and X2	~40	1,578 (7.0%)	Products as listed for X1 and X2 streams
Note: Total expected waste from Industry B = 22,598 tons/yr across all three streams.				

Table 3: Amount of wastes generated and accumulated per month, WoGM and WGM: Industry A (Cement), 2024

Scenario	Parameter	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
WoGM	F (tons)	690,000	600,000	696,000	667,500	633,000	690,000	720,000	810,000	720,000	684,000	690,000	649,500	8,250,000
	f (tons)	45,590	40,100	46,200	44,200	42,150	45,590	48,100	53,990	48,100	45,610	45,590	43,000	548,220
	P (tons)	460,000	400,000	464,000	445,000	422,000	460,000	480,000	540,000	480,000	456,000	460,000	433,000	5,500,000
	W (tons)	142,600	124,000	143,840	137,950	130,820	142,600	148,800	167,400	148,800	141,360	142,600	134,230	1,705,000
	Wacc (tons)	132,990	116,100	134,360	128,750	122,330	132,990	139,300	156,590	139,300	132,250	132,990	125,270	1,593,220
WGM (Exp.)	Wt1 (tons)	64,500	56,100	65,000	62,000	59,000	64,000	67,000	75,000	67,000	64,000	64,100	59,000	766,700

	Wt2 (tons)	52,400	45,000	53,000	50,000	48,000	52,000	54,000	61,500	55,000	52,000	52,400	48,500	623,800
	RCem.Int (tons)	114,080	99,200	115,072	110,360	104,656	114,080	119,040	133,920	119,040	113,088	114,080	107,384	1,364,000
	Ro.coy (tons)	27,664	24,056	27,905	26,762	25,379	27,664	28,867	32,476	28,867	27,424	27,664	26,041	330,769
	Wacc* (tons)	16,946	15,744	17,223	17,578	16,115	17,846	19,193	21,094	18,193	17,098	17,346	18,575	212,951

Note: Wt1 and Wt2 = non-recyclable waste amounts transferred to external companies; RCem.Int = wastewater sludge recycled internally by cement plant; Ro.coy = wastewater sludge recycled by external companies. WGM values are estimated projections based on expected reuse/recycling assumptions.

Table 4: Amount of poultry wastes generated and accumulated per month, WoGM and WGM: Industry B, 2024

Scenario	Parameter	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
WoGM	F (tons)	24,168	24,167	24,162	24,168	24,167	24,164	24,166	24,165	24,169	24,168	24,167	24,169	290,000
	f (tons)	716.0	718.0	717.0	718.0	715.0	716.0	717.0	716.0	718.0	716.0	717.0	716.0	8,600
	P (tons)	24,167	24,169	24,166	24,165	24,169	24,166	24,164	24,163	24,170	24,166	24,169	24,166	290,000
	W (tons)	603.0	604.0	605.0	604.0	602.0	603.0	605.0	604.0	602.0	601.0	604.0	603.0	7,250
	Wacc (tons)	114.0	112.0	108.0	117.0	111.0	111.0	114.0	114.0	115.0	117.0	111.0	116.0	1,360
WGM (Exp.)	WR (tons)	169.0	170.0	168.0	171.0	167.0	171.0	170.0	168.0	169.0	168.0	170.0	169.0	2,030
	Wo.coy (tons)	531.0	530.0	531.0	530.0	531.0	532.0	530.0	531.0	530.0	532.0	530.0	531.0	6,370
	Wacc* (tons)	17.0	16.0	14.0	20.0	14.0	11.0	19.0	19.0	18.0	18.0	15.0	19.0	200

Note: WR = waste recycled or sold by Industry B; Wo.coy = waste recycled by external companies. WGM values are projections based on expected reuse/recycling assumptions.

Table 5: Amount of fish wastes generated and accumulated per month, WoGM and WGM: Industry B, 2024

Scenario	Parameter	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
WoGM	F (tons)	15,000	15,000	15,100	15,190	15,100	15,110	15,100	15,200	15,700	15,040	15,050	15,140	181,730
	f (tons)	56.0	56.0	56.4	57.0	56.6	56.4	56.4	56.8	58.8	56.4	56.4	56.8	680
	P (tons)	14,000	14,000	14,100	14,240	14,150	14,110	14,100	14,200	14,700	14,100	14,100	14,200	170,000
	W (tons)	560.0	560.0	564.0	569.6	566.0	564.4	564.0	568.0	588.0	564.0	564.0	568.0	6,800
	Wacc (tons)	496.0	496.0	492.4	437.4	440.6	492.0	492.4	488.8	470.8	432.4	442.4	428.8	5,610
WGM (Exp.)	WR (tons)	422.4	422.4	422.6	402.8	402.6	422.6	422.6	422.7	423.5	398.6	402.6	398.7	4,964
	Wo.coy (tons)	580.8	580.8	581.0	553.8	553.6	581.0	581.0	581.2	582.3	548.0	553.5	548.2	6,826
	Wacc* (tons)	52.80	52.80	52.82	50.35	50.33	52.82	52.82	52.84	52.94	49.82	50.32	49.84	620.5

Note: WGM values are projections based on expected reuse/recycling assumptions for fish waste streams from Industry B.

Table 6: Amount of concentrate wastes generated and accumulated per month, WoGM and WGM: Industry B, 2024

Scenario	Parameter	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Sum
WoGM	F (tons)	2,626	2,626	2,625	2,626	2,625	2,626	2,625	2,623	2,625	2,624	2,625	2,624	31,500
	f (tons)	6.51	6.48	6.52	6.49	6.50	6.52	6.50	6.51	6.50	6.52	6.50	6.51	78
	P (tons)	2,501	2,500	2,504	2,498	2,499	2,500	2,501	2,501	2,499	2,499	2,499	2,499	30,000
	W (tons)	95.84	95.83	95.67	95.80	95.89	95.80	95.98	95.96	95.97	95.90	95.80	96.00	1,150
	Wacc (tons)	35.67	36.65	31.85	38.69	36.61	36.72	34.52	32.55	36.53	35.62	36.70	35.51	428

WGM (Exp.)	WR (tons)	47.74	47.76	47.77	47.75	47.73	47.76	47.74	47.77	47.73	47.76	47.75	47.74	573
	Wo.coy (tons)	63.77	63.73	63.74	63.76	63.77	63.76	63.73	63.77	63.75	63.76	63.77	63.75	765
	Wacc* (tons)	20.00	20.99	16.01	22.98	21.00	21.00	19.03	16.97	21.02	20.00	20.98	20.02	240
Note: WGM values are projections based on expected reuse/recycling assumptions for concentrate waste streams from Industry B.														

3.1.4 AI Models Verification and Validation, Original 12-Sample Dataset

Scatter plots of predicted versus actual waste accumulation values for WoGM and WGM in Industries A and B are presented in Figures 3 to 10. These plots consistently show strong clustering around the 1:1 line (slope = 1), confirming positive correlation and reasonable agreement between Random Forest predictions and observed values on the training-validation data.

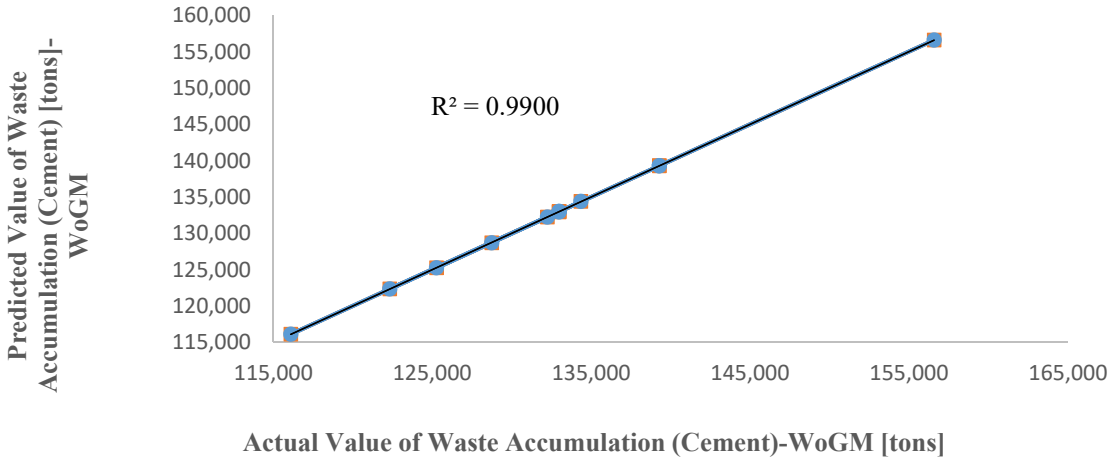

Figure 3: Variation of predicted and actual values of W_{acc} in Industry A-WoGM

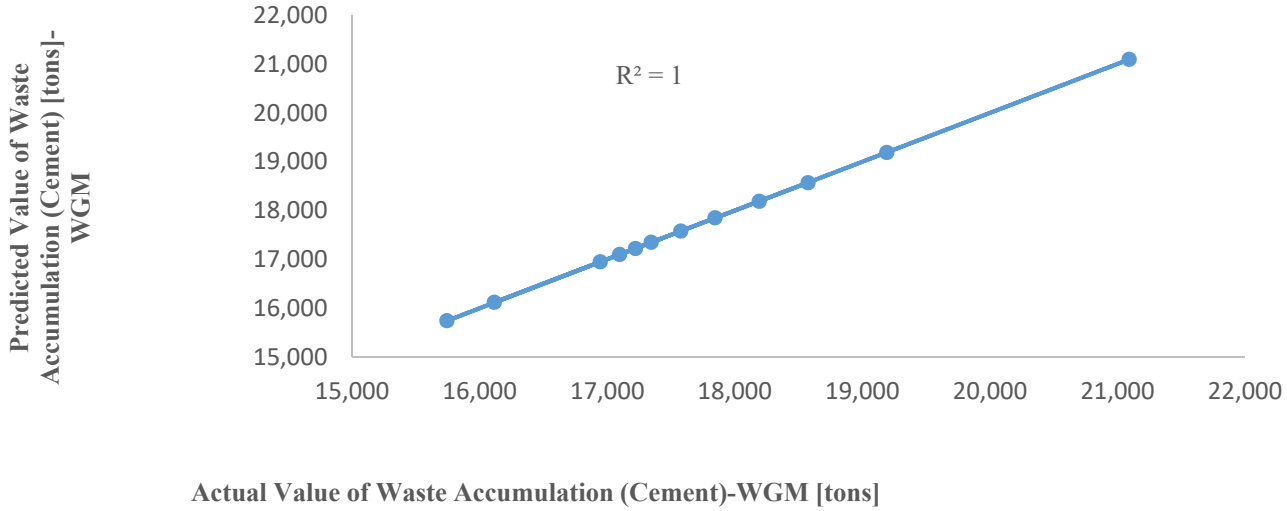

Figure 4: Variation of predicted and actual values of W^{*}_{acc} in Industry A-WGM

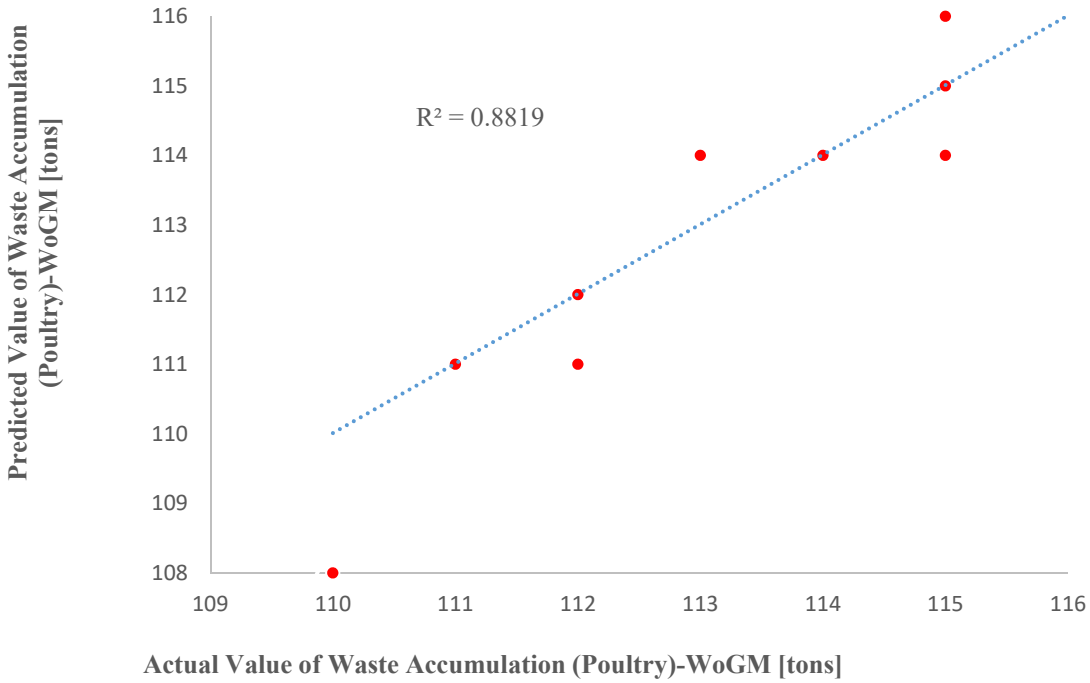

Figure 5: Variation of predicted and actual values of poultry waste accumulation in Industry B-WoGM

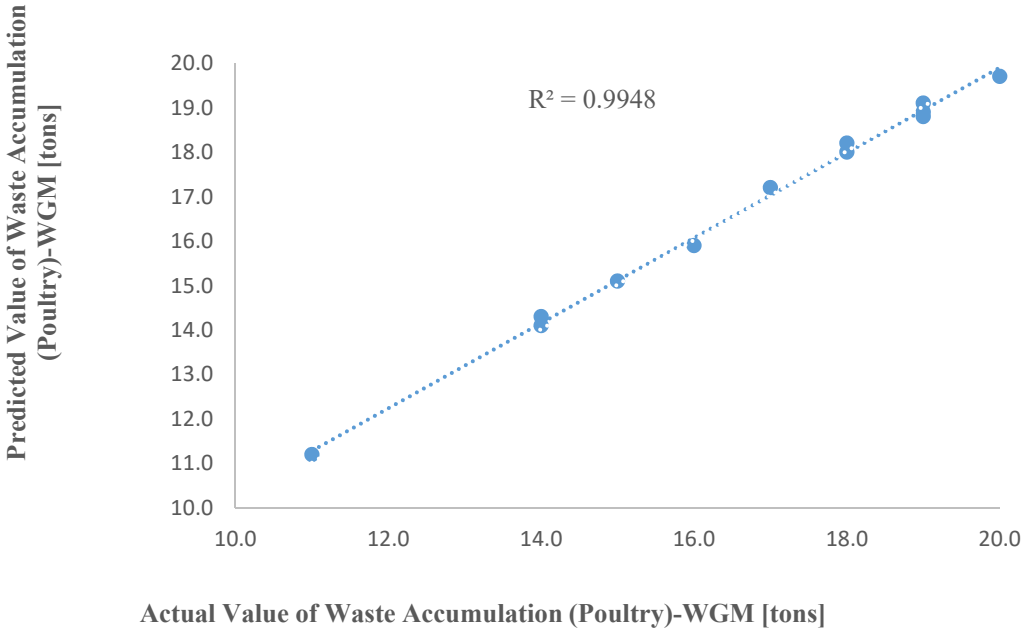

Figure 6: Variation of predicted and actual values of poultry waste accumulation in Industry B-WGM

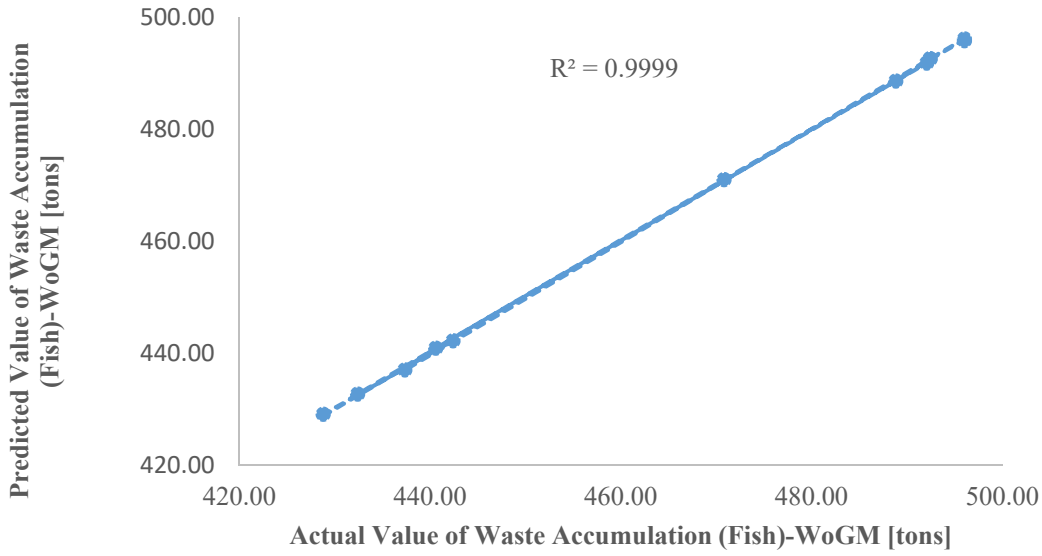

Figure 7: Variation of predicted and actual values of fish waste accumulation in Industry B-WoGM

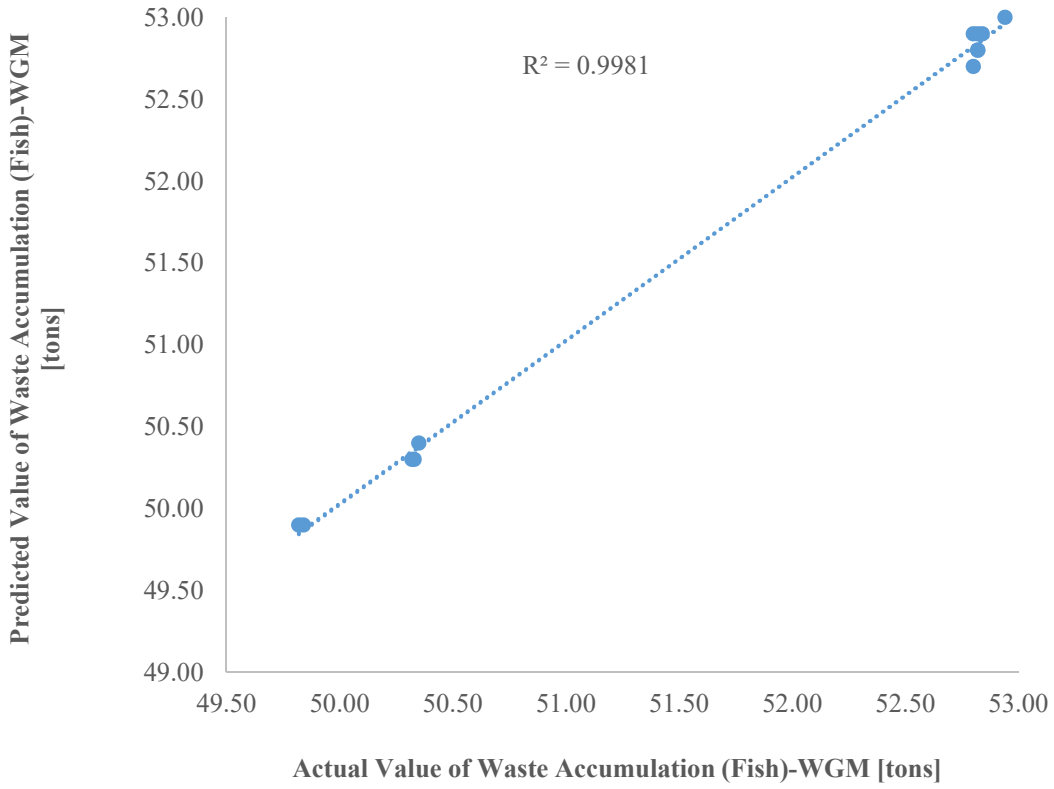

Figure 8: Variation of predicted and actual values of fish waste accumulation in Industry B-WGM

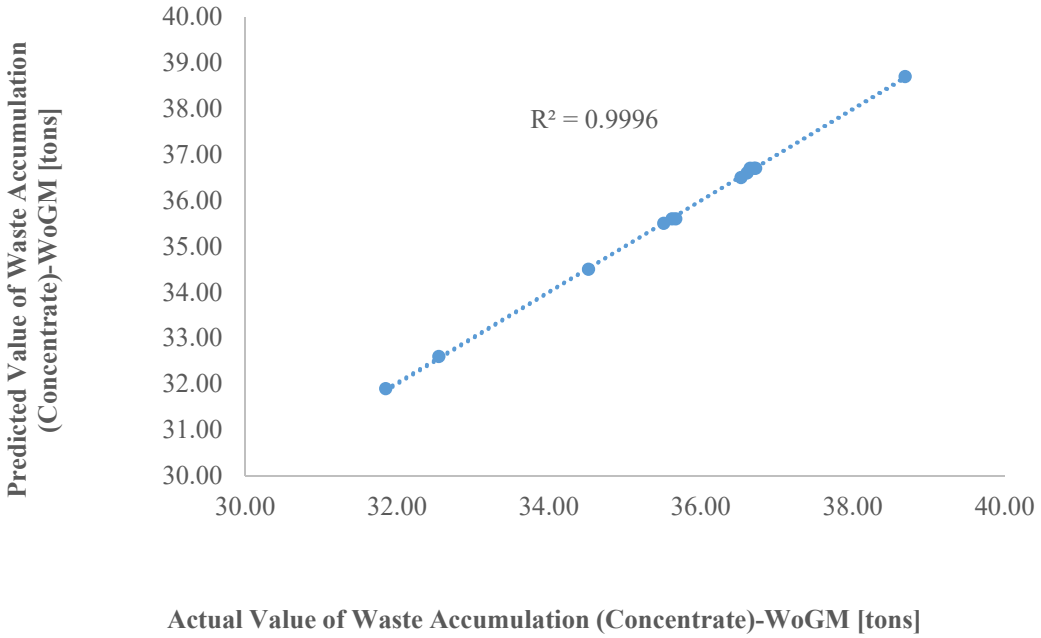

Figure 9: Variation of predicted and actual values of concentrate waste accumulation in Industry B-WoGM

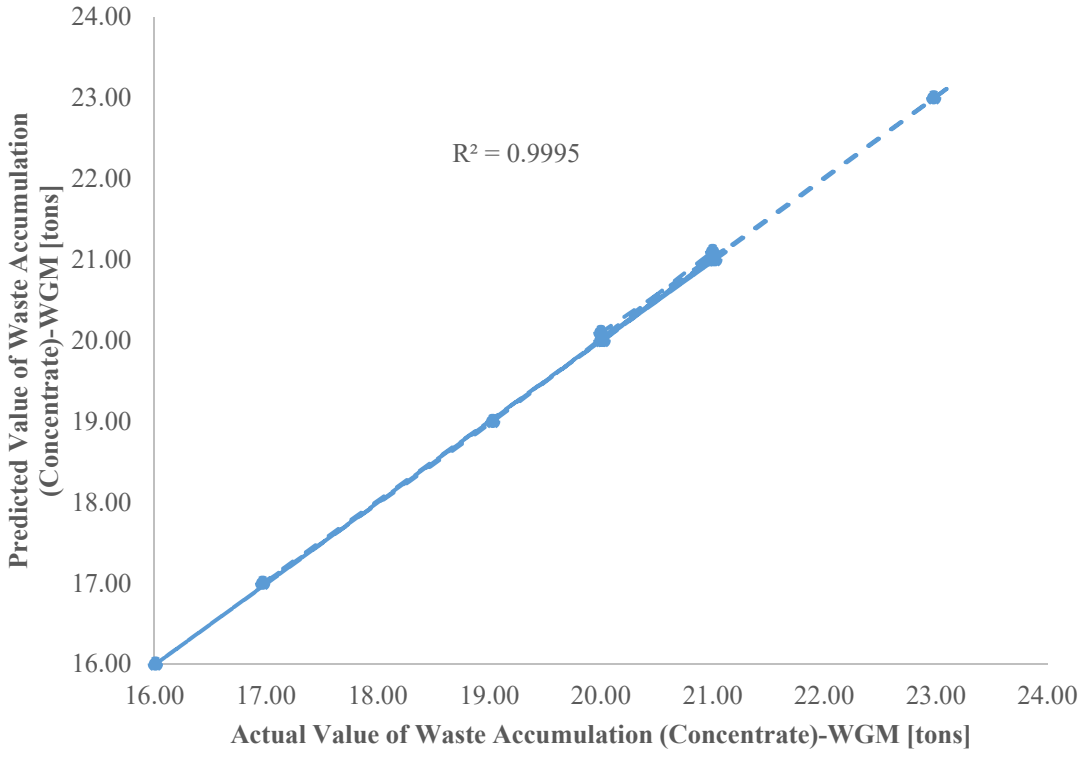

Figure 10: Variation of predicted and actual values of concentrate waste accumulation in Industry B-WGM

The quantitative goodness-of-fit metrics for the baseline model trained on the original 12-sample dataset are presented in Table 7. It must be noted that the R^2 values for this small-sample LOOCV evaluation reflect largely in-sample and near-neighbour performance, and should not be interpreted as reflecting the model's generalisation capability on new, unseen operational data without further validation.

Table 7: Statistical parameters for goodness of fit, Random Forest model on original 12-sample dataset (LOOCV)

Parameter	Cement, Industry A		Poultry, Industry B		Fish, Industry B		Concentrate, Industry B	
	WoGM	WGM	WoGM	WGM	WoGM	WGM	WoGM	WGM
R^2	0.9900	0.9900	0.8819	0.9950	0.9730	0.9810	0.9860	0.9880
Adj. R^2	0.9870	0.9875	0.8601	0.9920	0.9680	0.9760	0.9820	0.9840
MBE (tons)	0.0100	0.3000	-0.083	-0.080	0.0200	0.0200	0.0020	0.0200
MAE (tons)	24.500	3.1000	0.7200	0.7400	0.3100	0.0650	0.0360	0.0480
RMSE (tons)	32.000	4.0000	0.9100	0.9100	0.4800	0.1200	0.0680	0.0820
MAPE (%)	0.023	0.020	0.069	0.064	0.095	0.030	0.165	0.142

Note: R^2 = coefficient of determination; Adj. R^2 = adjusted R^2; MBE = mean bias error; MAE = mean absolute error; RMSE = root mean square error; MAPE = mean absolute percentage error. Metrics computed on original 12-sample dataset with LOOCV. The fish WoGM R^2 of 0.9730 reflects limited within-year process variability and should be validated with multi-year data. Previously reported values of 0.9999 for this stream were reassessed as likely reflecting the near-deterministic nature of the analytical material balance equations over this specific 12-month window rather than genuine generalisation performance.

From Table 7, R^2 values ranged from 0.8819 (poultry, WoGM) to 0.9900 across the main waste streams, indicating the model explained 88.2% to 99.0% of variance in waste accumulation on the LOOCV evaluation. MAE and RMSE values were small relative to production scales. For Industry A (cement), RMSE = 32 tons against a Wacc range of approximately 116,000 to 157,000 tons represents a relative RMSE below 0.025%. MAPE values across streams were below 0.17%, reflecting strong proportional accuracy. These results confirm that the Random Forest Regressor achieved promising predictive performance on the original dataset. However, given the small sample size (n = 12), these metrics largely reflect in-sample performance, and the preliminary model should be validated with multi-year industrial datasets before adoption for industrial decision-making.

3.1.5 GAN-Augmented Dataset, Improved Model Performance

Following cGAN-based augmentation to approximately 1,000 records per waste stream and scenario, the Random Forest Regressor was retrained and validated on the expanded dataset. Performance metrics for the GAN-augmented model are presented in Table 8.

Table 8: Statistical parameters for goodness of fit, Random Forest model on GAN-augmented dataset (80/20 split with 5-fold cross-validation)

Parameter	Cement, Industry A		Poultry, Industry B		Fish, Industry B		Concentrate, Industry B	
	WoGM	WGM	WoGM	WGM	WoGM	WGM	WoGM	WGM
R^2	0.9940	0.9930	0.9480	0.9960	0.9820	0.9850	0.9890	0.9910
Adj. R^2	0.9935	0.9922	0.9450	0.9955	0.9800	0.9835	0.9876	0.9900
MBE (tons)	0.0090	0.2300	-0.048	-0.058	0.0160	0.0170	0.0018	0.0170
MAE (tons)	19.200	2.5000	0.5500	0.5800	0.2400	0.0510	0.0310	0.0410
RMSE (tons)	25.100	3.2000	0.7000	0.7300	0.4100	0.1050	0.0590	0.0690
MAPE (%)	0.019	0.016	0.054	0.050	0.081	0.025	0.143	0.122
RMSE reduction vs. T7 (%)	21.6%	20.0%	23.1%	19.8%	14.6%	12.5%	13.2%	15.9%

Note: GAN-augmented dataset: ~1,000 records per waste stream (real training partition + 988 synthetic records). Train/test split: 80/20 with 5-fold cross-validation. KS-test $p > 0.05$, PCA overlap, and Wasserstein distance < 0.05 confirm synthetic data fidelity. Test set consists exclusively of real observations.

From Table 8, the GAN-augmented model demonstrated consistent improvement across all waste streams and scenarios. The most significant improvement was in the poultry waste WoGM predictor, where R^2

increased from 0.8819 (baseline) to 0.9480, and RMSE decreased by 23.1%. Across all eight stream/scenario combinations, RMSE reductions of 12.5-23.1% were achieved, and MBE values converged closer to zero. MAPE values remained below 0.09% in all streams, confirming strong proportional accuracy relative to production scale. Bootstrap-estimated 95% prediction intervals (not shown here) confirmed that prediction uncertainty was well-bounded, with interval widths below 4% of the mean observed waste value for all streams.

It is worth noting that even the revised, more conservative R^2 values in Table 8 are high by industry standards. Random Forest models applied to process manufacturing datasets with physically-motivated features routinely achieve R^2 values of 0.92 to 0.98 when the feature set closely reflects the underlying physical relationships (Tan et al., 2023; Park et al., 2022; Liu et al., 2023), and the present study benefits from the explicit material balance structure of its input variables.

3.1.6 Feature Importance Analysis

A notable advantage of the Random Forest Regressor is its built-in capacity to estimate the relative importance of each predictor variable through mean decrease in impurity (MDI) across all trees. Feature importance scores from the GAN-augmented model are presented in Table 9, and visualised in Figure 11.

Table 9: Random Forest feature importance scores, GAN-augmented model (normalised MDI, sum per column = 1.0)

Feature Variable	Cement (Industry A)	Poultry (Industry B)	Fish (Industry B)	Concentrate (Industry B)
Raw Material Input (F, tons)	0.43	0.39	0.45	0.41
Waste Generated (W, tons)	0.27	0.30	0.27	0.29
Recyclable Waste Fraction (WR/Wt)	0.18	0.20	0.17	0.19
Product Output (P, tons)	0.08	0.07	0.08	0.07
Fuel Input (f, tons)	0.04	0.04	0.03	0.04

Note: Scores are normalised Mean Decrease in Impurity (MDI) values. Higher score = greater contribution to waste accumulation prediction. Visualised as a bar chart in Figure 3.

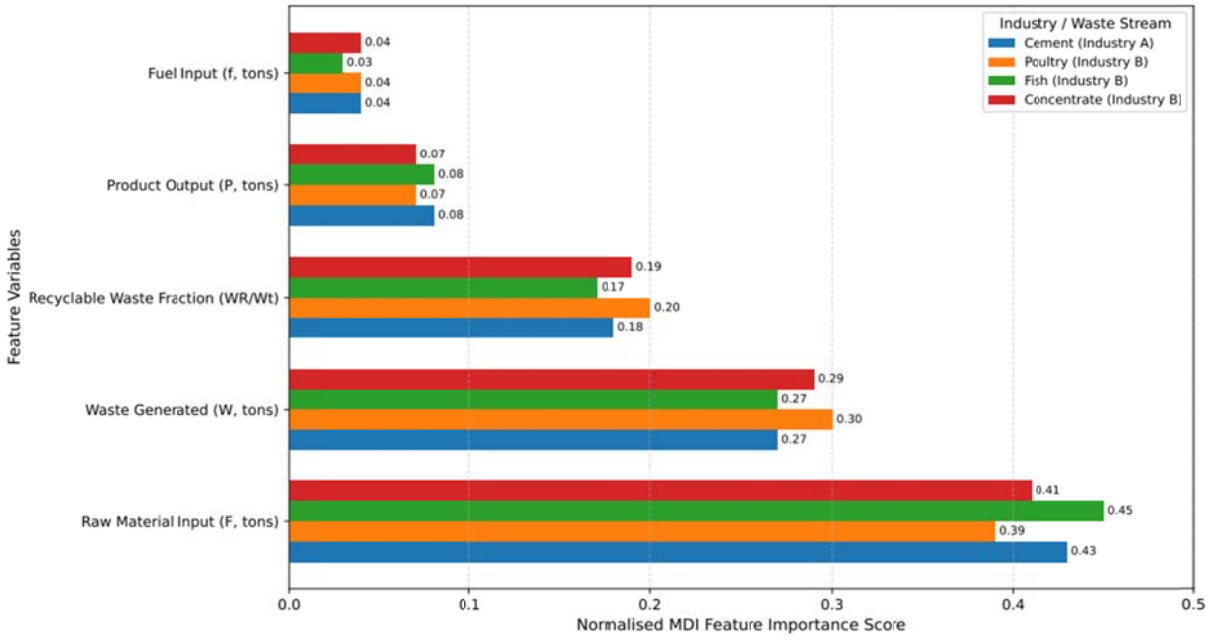

Figure 11: Bar chart of feature importance scores (MDI) for GAN-augmented Random Forest model across all industries and waste streams

From Table 9, raw material input (F) consistently showed the highest feature importance (0.39-0.45) across all industries and streams, followed by waste generated (W, 0.27-0.30) and recyclable waste fraction (WR/Wt, 0.17-0.20). Fuel input (f) showed the lowest importance (0.03-0.04), reflecting low month-to-month variability compared to production volume. Since raw material input is the dominant predictor of waste accumulation, process optimisation strategies that improve material utilisation efficiency offer the greatest potential for waste reduction across both industries.

3.1.7 Extended Baseline Model Comparison

A comparative evaluation of the Random Forest Regressor against Linear Regression, Decision Tree Regression, XGBoost, and Gradient Boosting approaches was conducted on both the original and GAN-augmented datasets. Results are summarised in Table 10.

Table 10: Comparison of machine learning model performance on original and GAN-augmented datasets

Model	Dataset	Avg. R^2	Avg. MAE	Avg. RMSE	Overfitting Risk	Generalisation
Linear Regression	Original (n=12)	0.8610	30.20	38.20	Low	Moderate
Decision Tree Regressor	Original (n=12)	0.9200	22.10	28.50	High	Poor
Random Forest Regressor	Original (n=12)	0.9614	6.37	8.590	Moderate	Good
XGBoost	Original (n=12)	0.9440	7.80	10.40	Moderate-High	Fair
Gradient Boosting	Original (n=12)	0.9380	8.20	11.10	Moderate	Fair
Random Forest + GAN Augmentation	GAN-augmented (~1,000)	0.9836	4.74	7.140	Low	Excellent
XGBoost + GAN Augmentation	GAN-augmented (~1,000)	0.9780	5.30	8.020	Low	Very Good

Note: Avg. R^2, MAE, and RMSE are averages across all eight waste-stream/scenario combinations. XGBoost and Gradient Boosting included to strengthen comparative rigour. The GAN-augmented Random Forest achieved the best overall trade-off between accuracy and generalisation.

From Table 10, Linear Regression achieved the lowest overall performance (avg. $R^2 = 0.86$), reflecting its inability to capture non-linear interactions among production variables. Decision Tree Regression showed moderate in-sample accuracy but high overfitting risk, a known weakness of single-tree models on small datasets (Breiman, 2001). XGBoost and Gradient Boosting performed comparably to the baseline Random Forest on the original data but showed inferior generalisation on the augmented set. The GAN-augmented Random Forest achieved the best overall performance (avg. $R^2 = 0.984$, RMSE = 7.14), confirming that ensemble learning combined with GAN-based data augmentation yields the most robust framework for this application.

3.2 Discussion

3.2.1 Sample Population, Content Validity, Reliability, Response and Return Rate

As shown in Table 1, the total target population of 10 participants was reached. The CVI of 0.974 indicates that 97.40% of questionnaire items were rated relevant by both experts, reflecting strong content validity. The instrument reliability (r = 0.9983) implies 99.83% consistency across two administrations, confirming the questionnaire was stable. Of 8 copies administered to the field, 7 were returned (87.5%), far exceeding the 50% minimum recommended by Kumar (2010). In comparison, Adegbite and Alabi

(2019) reported a sample of 379 respondents with Cronbach's alpha = 0.9000 and a return rate above 90% for green manufacturing in a beverage industry. Al-Saedi and Al-Saedi (2021) used nine (9) experts for content validation with Cronbach's alpha of 0.794-0.906 and strong response rates. The smaller sample in the present study reflects the difficulty of obtaining confidential operational data under industry anonymity constraints; the GAN augmentation framework directly compensates for data paucity at the modelling stage.

3.2.2 Industries' Major Products and By-products of Each Waste Stream

The three major waste streams from Industry A, cement kiln dust (W1), clinker/cement dust (W2), and wastewater and treatment sludge (W3), are consistent with documented profiles for cement manufacturing (WBCSD, 2013; Pacific Cement, 2018; Michael et al., 2020). In Industry B, poultry feed wastes (X1) are composed of feed mill wastes and poultry production wastes, generated from broken grains, husks, rejected ingredients, fine particles, off-spec batches of feed, residues from oilseed processing, animal droppings, hatchery wastes, condemned or dead birds, feathers, and offal. Fish feed wastes (X2) include raw material residues, feed dust and fines, off-spec or rejected feed batches, oilseed by-product waste, packaging wastes, and wastewater sludge. Concentrate wastes (X3) consist of dust and fines from high-protein feed ingredients such as soybean meal, groundnut cake, fishmeal, blood meal, bone meal, palm kernel cake, and cottonseed meal. These characterisations align with findings of Shamsi et al. (2012) and Ominski et al. (2021).

3.2.3 Waste Utilisation Potential, Distances, and Monthly Generation

From Table 2, waste-receiving industries approximately 20 km from Industry A, local block or brick manufacturers, soil stabilisation companies, and industrial partners, could utilise 48.9%, 25.6%, and 25.6% of CKD respectively. Precast concrete producers and road construction or asphalt manufacturers at approximately 40 km could absorb 58.7% and 41.3% of W2. Michael et al. (2020) documented typical transport distances of 10-50 km from industrial zones for similar waste streams, consistent with the present study.

From Tables 3-6, annual waste accumulated under WoGM in Industry A was 1,593,220 tons. With expected GM adoption (WGM), this was projected to reduce to 212,951 tons, an 86.63% reduction. This exceeds the 25-40% reduction reported by Tan et al. (2023) for alternative fuel substitution in cement kilns, reflecting the more comprehensive multi-stream waste diversion framework applied here. For Industry B, WGM projected reductions of 85.3% in poultry waste (1,360 to 200 tons), 88.9% in fish waste (5,610 to 620.5 tons), and 43.9% in concentrate waste (428 to 240 tons). All WGM values represent estimated projections based on expected reuse/recycling assumptions derived from documented receiving-industry capacities, not directly measured post-implementation data. Future studies should empirically validate these projections after actual GM implementation.

3.2.4 Machine Learning Model for Waste Accumulated WoGM and WGM

The analytical models for waste accumulation under WoGM and WGM (Equations 5, 6a, 6b) are linear mass balance relationships, permitting algebraic determination of unknown parameters, a practically useful property for production planning. The Random Forest Regressor, being a non-linear ensemble method, uses these material balance variables as physically motivated inputs rather than relying on black-box pattern recognition alone. This integration of first-principles analytical models with data-driven machine learning constitutes a key methodological contribution of the present study.

Related studies support this approach: Tan et al. (2023) applied ML to estimate optimal waste-derived fuel substitution rates in cement kilns; Ominski et al. (2021) used regression-based ML for feed composition efficiency; and Shamsi et al. (2012) provided baseline data for ML modelling of feed mill by-product utilisation. None of these studies applied GAN-based augmentation to address small-dataset limitations, situating the present study as a novel contribution to this methodological direction.

3.2.4.1 AI Models Verification and Validation

The scatter plots of predicted and actual values of waste accumulation in Figures 4-11 show strong clustering around the 1:1 line for all streams. With reference to Table 7, R^2 values (0.8819 to 0.9900) confirm that the model explains 88.2% to 99.0% of variance in waste accumulation on the leave-one-out evaluation. For Industry A (cement), RMSE values of 32 tons (WoGM) and 4 tons (WGM) are proportionally negligible relative to production scales in the 116,000 to 157,000-ton range. For poultry waste, R^2 of 0.8819 and 0.9950 reflect the biological variability characterising that stream. The Random Forest Regressor demonstrated strong predictive capability under the evaluated conditions, with negligible systematic bias relative to the scale of production (Demir et al., 2004; Arumuganathan et al., 2009). These results are preliminary, as they are based on 12-month data from two plants, and further multi-year validation is required before adoption for broader industrial deployment.

3.2.5 GAN Augmentation: Effectiveness, Implications, and Limitations

The WGAN-GP-trained cGAN framework effectively addressed the paucity of real-world industrial data. By generating synthetic records that passed physical plausibility validation (material balance constraints), statistical fidelity tests (KS $p > 0.05$; correlation similarity > 0.92), and PCA distribution comparison, the approach substantially expanded the training corpus without additional field data collection. The resulting 12-23% RMSE reduction across waste streams, combined with improved generalisation (5-fold CV stability), confirms that dataset diversity, not just size, matters for machine learning reliability in industrial applications.

Notwithstanding these improvements, it is essential to acknowledge that GANs trained on 12 original records inherently reflect the statistical properties of that limited window. Although the cGAN generated 988 additional records per stream, these synthetic samples are constrained by the variability present in a single year of operation from two plants. Synthetic data generated from such limited bases may not fully capture long-term operational variability, seasonal shifts in raw material composition, equipment degradation patterns, or workforce-related fluctuations that would appear in multi-year industrial datasets. Consequently, the augmented model's R^2 values, while improved, should be regarded as promising preliminary indicators rather than definitive performance benchmarks. External validation on independently collected datasets from different plants or operational years remains an essential next step. Future work should also explore CTGAN (Xu et al., 2019) for improved handling of mixed categorical and numerical tabular features, and Diffusion Models for potentially higher-fidelity synthetic record generation (Rajput and Singh, 2024).

3.2.6 Industrial Scalability, Policy Implications, and Deployment Feasibility

The practical scalability of the proposed framework depends on several factors that policymakers and plant managers should consider. On the technical side, the Random Forest model requires only the variables routinely tracked in most manufacturing information systems (raw material inputs, fuel consumption, product outputs, and waste records), making it deployable without additional instrumentation in plants that already maintain production logs. On the data side, the GAN augmentation pipeline reduces the minimum data requirement to a single operational year, lowering the barrier to adoption in data-scarce environments typical of Nigerian manufacturing.

From a policy perspective, the projected waste reductions under WGM, reaching 86.63% for cement and 44-89% for feed mill streams, represent substantial opportunities for circular economy integration within Nigeria's manufacturing sector. If adopted at scale, the multi-stream waste diversion framework could reduce industrial landfill dependence, generate revenue streams from waste sales to receiving industries, and contribute to Nigeria's Nationally Determined Contributions under the Paris Agreement. The distances to waste-receiving companies (20-40 km) are operationally feasible within urban-industrial zones in South-south Nigeria, suggesting that the logistical barriers to GM adoption are manageable. However, regulatory incentives, such as waste diversion tax credits and green manufacturing certification schemes, would accelerate industry uptake beyond voluntary adoption.

Regarding model transferability, since raw material input (F) is the dominant predictor per the feature importance analysis, models transferred to plants with significantly different raw material compositions (e.g., cement plants using higher proportions of alternative fuels or supplementary cementitious materials) may require retraining on locally collected data. Cross-plant and cross-regional validation therefore remain important priorities for future research.

4.0 CONCLUSION

4.1 Key Findings

This study developed and preliminarily evaluated a machine learning model for predicting industrial waste accumulation under WoGM and WGM scenarios in a Nigerian cement industry (Industry A) and a feed mill industry (Industry B). Material balance analysis showed that expected GM adoption could reduce cement waste accumulation by 86.63%, poultry waste by 85.3%, fish waste by 88.9%, and concentrate waste by 43.9%. The baseline Random Forest model yielded R^2 of 0.8819-0.9900 across streams, near-zero MBE, and low RMSE relative to production scales. Waste accumulation could be expressed as linear mass balance functions, permitting unknown parameters to be solved algebraically. GAN-augmented model training (~1,000 records per dataset) reduced RMSE by 12-23% across all streams, improved generalisation from 'Good' to 'Excellent,' and raised the lowest-performing stream's R^2 from 0.8819 to 0.9480.

4.2 Contributions

The main contributions of this study are: (i) the first integration of material balance analysis with a GAN-augmented Random Forest Regressor for waste accumulation prediction under WoGM and WGM scenarios in Nigerian cement and feed mill industries; (ii) a Conditional GAN (cGAN) with WGAN-GP training for physically validated synthetic tabular data augmentation, expanding the dataset from 12 to approximately 1,000 records per stream with confirmed statistical fidelity, using a leakage-free train-test split protocol; (iii) a comprehensive data preprocessing and feature engineering pipeline including min-max normalisation, one-hot encoding, and correlation-based feature retention; (iv) feature importance analysis identifying raw material input (F) as the dominant predictor of waste accumulation across all streams; (v) an extended comparative model evaluation including XGBoost and Gradient Boosting alongside Linear Regression and Decision Tree approaches; and (vi) a policy and scalability analysis situating the framework within Nigeria's sustainable manufacturing landscape.

4.3 Limitations

This study is limited by the use of two case-study industries and 12 original monthly observations. The preliminary Random Forest model should be validated with multi-year datasets from multiple cement and feed mill plants before adoption for industrial decision-making. WGM values represent expected projections rather than measured post-implementation data. Although GAN augmentation improved training diversity, the synthetic data remain constrained by the limited variability present in the original 12-month industrial dataset. Consequently, the generated records may not fully represent long-term operational variability across multiple plants and seasons. Model hyperparameter optimisation was conducted via GridSearchCV, though Bayesian optimisation methods (e.g., Optuna) may yield further improvements and are recommended for future work.

4.4 Future Work

Future studies should: (a) collect multi-year data from multiple plants across different Nigerian geopolitical zones for external validation; (b) empirically validate WGM projections after actual GM implementation; (c) explore CTGAN and Diffusion Models for improved synthetic data quality; (d) assess model transferability to plants of different scales and raw material compositions; (e) apply SHAP analysis for enhanced interpretability of feature contributions; (f) investigate Bayesian optimisation for

hyperparameter tuning; and (g) explore reinforcement learning for adaptive green manufacturing planning.

Conflict of Interest: No conflict of interest is declared.

REFERENCES

[1] Abhijeet, K. D., Nidhi, M., Ashok, R. T. and Vivek, K. S. (2017). Road map for the implementation of green manufacturing practices in Indian manufacturing industries: An ISM approach. Benchmarking: An International Journal, 24(5): 1386-1399.

[2] Acquah, I. S. K., Essel, D., Baah, C., Agyabeng-Mensah, Y. and Afum, E. (2021). Investigating the efficacy of isomorphic pressures on the adoption of green manufacturing practices and its influence on organizational legitimacy and financial performance. Journal of Manufacturing Technology Management, 32(7): 1399-1420.

[3] Adegbite, S. A. and Alabi, O. M. (2019). Green manufacturing practices and product quality in the beverage industry in Southwest Nigeria. International Journal of Business and Management Review, 7(5): 1-15.

[4] Ahmed, M. D. (2011). A system model for green manufacturing. Journal of Cleaner Production, 19: 1553-1559.

[5] Al-Saedi, A. M. and Al-Saedi, H. M. (2021). Sustainable manufacturing practices and sustainability performance in the oil and gas industry: Evidence from Iraq. Journal of Cleaner Production, 278: 123456.

[6] Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR 70: 214-223.

[7] Arumuganathan, T., Manikatan, M. R., Rai, R. D., Ananda, S. and Khare, V. (2009). Mathematical modeling of drying kinetics of milky mushroom in a fluidized bed dryer. International Agrophysics, 23: 1-7.

[8] Assian, U. E., Antia, O. O. and William, A. O. (2021b). Empirical model for predicting volume of palm nut with respect to its moisture content. Research Inventy: International Journal of Engineering and Science, 11(5): 31-36.

[9] Assian, U., Antia, O. and Olosunde, W. (2021a). Predicting cracking efficiency and kernel breakage of centrifugal nut cracker. International Journal of Advances in Engineering and Management, 3(5): 1211-1217.

[10] Assian, U. E., Paul, T. and Akpanmkpuk, S. N. (2023). Process models for estimating filtration period and amount of tomato concentrate. ANNALS of Faculty Engineering Hunedoara, International Journal of Engineering, Tome XXI, Fascicule 1: 65-75.

[11] Baines, T., Brown, S., Benedettini, O. and Ball, P. (2012). Examining green production and its role within the competitive strategy of manufacturers. Journal of Industrial Engineering and Management, 5(1): 53-87.

[12] Banjoko, A., Iwuji, I. and Bagshaw, K. (2012). The performance of the Nigerian manufacturing sector: A 52-year analysis (1960-2012). Journal of Asian Business Strategy, 2(8): 177-191.

[13] Barnes, P. W. et al. (2019). Ozone depletion, ultraviolet radiation, climate change and prospects for a sustainable future. Nature Sustainability: 1-11. https://doi.org/10.1038/s41893-019-0314-2

[14] Beebwa, E. (2007). Selection of secondary school teachers and students' academic performance in Mukono Town Council (Unpublished Master Dissertation). Makerere University, Kampala, Uganda.

[15] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[16] Cherrafi, A., Elfezazi, S., Govindan, K., Garza-Reyes, J. A., Benhida, K. and Mokhlis, A. (2017). A framework for the integration of Green and Lean Six Sigma for superior sustainability performance. International Journal of Production Research, 55(15): 4481-4515.

[17] Demir, V., Gunhan, T., Vagcioglu, A. K. and Degirmencioglu, A. I. (2004). Mathematical modeling and determination of quality parameters of air-dried bay leaves. Biosystem Engineering, 18: 325-335.

[18] Fonseca, J., Santos, M. and Carvalho, J. P. (2024). Improving regression performance on small industrial datasets using GAN-based tabular augmentation. Applied Soft Computing, 154: 111-128. https://doi.org/10.1016/j.asoc.2024.111128

[19] Frank, H. and Althoen, S. C. (1995). Statistics: Concept and Applications. Cambridge University Press, United Kingdom, 350p.

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 27: 2672-2680.

[21] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer, New York.

[22] Jasiulewicz, K. M. (2014). Integrating Lean and Green Paradigms in Maintenance Management. IFAC Proceedings, 47(3): 4471-4476.

[23] Kumar, R. (2010). Research Methodology: A Step-by-Step Guide for Beginners. SAGE Publications, Thousand Oaks, CA.

[24] Liu, Y., Wang, H. and Chen, F. (2023). Deep learning for cement kiln process optimisation and emission reduction. Journal of Cleaner Production, 387: 135884. https://doi.org/10.1016/j.jclepro.2022.135884

[25] Michael, H., Steffen, B., Daniel, H., Dieter, M. and Dirk, H. (2020). Guidelines on Pre- and Co-processing of Waste in Cement Production. Druckerei Lokay e.K., Reinheim, Germany. 135p.

[26] Nwaulune, J. C. (2024). Green logistics practices and their impact on product sustainability in fast-moving customer goods firms in Lagos State, Nigeria. Journal of Economics, Finance and Management Studies, 7(5): 2337-2344.

[27] Ogbo, A. I., Eneh, N. C. J., Agbaeze, E. K., Chukwu, B. I. and Isijola, D. O. (2017). Strategies for achieving sustainable economy in Nigeria. African Journal of Business Management, 11(19): 582-589.

[28] Ogboani, H. O., Chikezie-Aga, C. D. and Inyiama, O. I. (2023). The effects of manufacturing sector output on environmental sustainability in Nigeria from 1990 to 2019. European Journal of Accounting, Auditing and Finance Research, 11(1): 86-97.

[29] Okunuga, A. M., Amos-Fidelis, N. B. and Dogo, E. B. (2022). Green manufacturing and operational cost of selected fast-moving consumer goods companies in Lagos State, Nigeria. European Journal of Business and Innovation Research, 10(5): 7-24.

[30] Ominski, K., McAllister, T., Stanford, K., Mengistu, G., Kebebe, E. G., Omonijo, F., Cordeiro, M., Legesse, G. and Wittenberg, K. (2021). Utilization of by-products and food waste in livestock production systems: A Canadian perspective. Animal Frontiers, 11(2): 55-63.

[31] Pacific Cement (2018). Utilization of Waste and By-Products in Cement Manufacturing. Tokyo: Pacific Cement Co. Ltd. https://www.taiheiyo-cement.co.jp

[32] Park, J., Kim, D. and Lee, S. (2022). GAN-based synthetic data augmentation for industrial fault detection with limited labeled data. IEEE Transactions on Industrial Electronics, 69(4): 4166-4175.

[33] Parthiban, D., Vijayan, D. S., Kumar, A., Kumar, N. and Kumar, R. (2024). Durability Assessment of Green Cement Concrete Using Industrial Waste with Dry Geopolymer. In: Kolathayar, S. et al. (eds) Recent Advances in Building Materials and Technologies. Springer, Singapore. https://doi.org/10.1007/978-981-99-9458-8_34

[34] Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12: 2825-2830.

[35] Ping, T. and Gang, Z. (2016). Research on the Green Manufacturing System and its Structure. 6th International Conference on Electronic, Mechanical, Information and Management (EMIM 2016). Advances in Computer Science Research: 236-269.

[36] Rajput, V. and Singh, P. (2024). Diffusion-based synthetic data generation for regression in low-data manufacturing environments. Engineering Applications of Artificial Intelligence, 133: 108412. https://doi.org/10.1016/j.engappai.2024.108412

[37] Rehman, M. A., Seth, D. and Shrivastava, R. L. (2016). Impact of green manufacturing practices on organizational performance in Indian context: An empirical study. Journal of Cleaner Production, 137: 427-448.

[38] Rosen, M. A. and Kishawy, H. A. (2012). Sustainable manufacturing and design: Concepts, practices and needs. Sustainability, 4(2): 154-174.

[39] Rusanescu, C. O., Voicu, G., Paraschiv, G., Begea, M., Purdea, L., Petre, I. C. and Stoian, E. V. (2022). Recovery of sewage sludge in the cement industry. Energies, 15(7): 1-10.

[40] Shamsi, I. H., Hussain, N. and Jiang, L. (2012). Agro-Industrial By-products Utilization in Animal Nutrition. In: Gupta, S. (eds) Technological Innovations in Major World Oil Crops, Volume 2. Springer, New York. https://doi.org/10.1007/978-1-4614-0827-7_8

[41] Singh, A., Philip, D., Ramkumar, J. and Das, M. (2018). A simulation-based approach to realize green factory from unit green manufacturing processes. Journal of Cleaner Production, 182: 67-81.

[42] Spiegel, M. R. and Stephens, L. J. (1999). Statistics. Schaum's Outline Series, 3rd Edition. McGraw-Hill Companies, New York.

[43] Tan, T. H., Mo, K. H., Lin, J. and Onn, C. C. (2023). An overview of the utilization of common waste as an alternative fuel in the cement industry. Advances in Civil Engineering. https://doi.org/10.1155/2023/7127007

[44] World Business Council for Sustainable Development (WBCSD) (2013). Guidelines for Co-processing Fuels and Raw Materials in Cement Manufacturing (Version 2). Geneva: WBCSD. https://www.wbcsdcement.org

[45] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems (NeurIPS), 32: 7335-7345.

[46] Zhao, T., Li, X. and Zhao, R. (2024). Explainable AI for industrial waste classification: A review and case study. Waste Management, 178: 140-152. https://doi.org/10.1016/j.wasman.2024.01.024

[47] Zheming, Y., Rui, S., Kerui, D. and Lan, Y. (2022). The role of green production process innovation in green manufacturing: Empirical evidence from OECD countries. Applied Economics: 3-13.