CHEMICAL ENGINEERING TRANSACTIONS

VOL. 122, 2025

A publication of AIDIC

The Italian Association of Chemical Engineering
Online at www.cetjournal.it

Guest Editors: Jeng Shiun Lim, Bohong Wang, Guo Ren Mong, Petar S. Varbanov

ISBN 979-12-81206-23-6; **ISSN** 2283-9216

DOI: 10.3303/CET25122004

A Soft Clustering and Machine Learning-Based Framework for Latent Phase Mapping in Poultry Litter Composting

Malvin Ma[a], Xiao Qi Ng[a], Yew Woh Hui[b], Norsharina Md Saad[c], Jian Ding Tan[d], Rui Zhi Yap[e], Jian Ping Tan[f,*]

[a]School of Energy and Chemical Engineering, Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor Darul Ehsan, Malaysia.
[b]China-ASEAN College of marine Sciences, Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor Darul Ehsan, Malaysia.
[c]Agro-Biotechnology Institute (ABI), National Institutes of Biotechnology Malaysia (NIBM), Jalan Bioteknologi, 43400 Serdang, Selangor.
[d]School of Robotics and Automation Engineering, Xiamen University Malaysia, Jalan Sunsuria, Bandar Sunsuria, 43900 Sepang, Selangor Darul Ehsan, Malaysia.
[e]Edafos Agro PLT, Jalan TPJ 5, Taman Perindustrian Jaya 2, 85000 Segamat, Johor, Malaysia.
[f]Center of Excellence for Sustainable Technologies for Agricultural Advancement and a Zero-waste Economy (STARGAZE), Xiamen University Malaysia, 43900 Sepang, Selangor, Malaysia.
jianping.tan@xmu.edu.my

Poultry litter composting is a complex, multi-phase biological process that remains challenging to monitor and control using traditional methods. This study presents a machine learning framework that combines Gaussian Mixture Models (GMMs) for unsupervised soft clustering with supervised regression models for real-time phase inference. Using high-resolution sensor data from two full-scale poultry litter composting experiments, GMMs successfully identified four latent composting phases: pre-composting, mesophilic, thermophilic, and maturation, capturing gradual transitions and overlapping states that conventional temperature thresholds often fail to detect. A supervised learning module was then developed to map real-time sensor snapshots to the GMM-derived phase probabilities. Models trained via H2O AutoML achieved high predictive performance across all clusters ($R^2 > 0.99$, RMSE < 0.032), with ensemble tree methods outperforming other algorithms. The proposed approach enables low-latency, label-free compost phase monitoring using only instantaneous temperature and humidity data, making it well-suited for embedded control systems and real-time decision support. This framework provides a generalizable and interpretable tool for intelligent composting, with applications in process optimization, anomaly detection, and phase-aware control strategies in agro-industrial waste management.

1. Introduction

Poultry litter, comprising bedding material, feathers, droppings, feed residues, and other bio-organic residue from the poultry production process, is a major agro-industrial waste from the global poultry sector. In light of the global poultry population reaching 29.15 billion in 2023, with China accounting for 21.85 % and Southeast Asia accounting for 18.52 % (Food and Agriculture Organization of the United Nations, 2025), the effective management of poultry litter has become a pressing environmental and engineering challenge. In developing countries, this waste is often processed into bio-organic fertilizer via composting, due to its rich macronutrient content, primarily nitrogen, potassium, phosphorus). If not properly managed, this process can lead to nitrate leaching, greenhouse gas emissions, and pathogen outbreaks (Alarefee et al., 2023). Recent studies on integrated valorisation strategies using ex-situ microbial inoculation and black soldier fly larvae (BSFL) composting show the potential of biologically enhanced composting approaches for nutrient recovery and microbial stabilization (Yeow et al., 2024). Although thermal and biochemical valorisation is rising in developed regions, composting remains dominant for its operational simplicity and low cost. (Duarah et al., 2022).

Paper Received: 29 May 2025; Revised: 21 September 2025; Accepted: 29 October 2025
Please cite this article as: Ma M., Ng X.Q., Hui Y.W., Md Saad N., Tan J.D., Yap R.Z., Tan J.P., 2025, A Soft Clustering and Machine Learning-Based Framework for Latent Phase Mapping in Poultry Litter Composting, Chemical Engineering Transactions, 122, 19-24 DOI:10.3303/CET25122004

Composting serves to stabilize organic matter, reduce volume, and produce a valuable soil amendment. Despite its long-standing use and widespread adoption, it remains inherently nonlinear, multi-phase, and difficult to monitor and control in real time (Huang et al., 2024). Poultry litter composting typically proceeds through mesophilic, thermophilic, cooling, and maturation phases, driven by nutrient, microbial, aeration, moisture, and temperature interactions (Walling and Vaneeckhaute, 2021). These transitions are highly sensitive to process conditions, and mismanagement can lead to incomplete stabilization, odour generation, and increased time to maturity. Yet, current monitoring practices rely heavily on manual sampling or rule-based heuristics, which are incapable of detecting latent transitions or process inefficiencies (Huang et al., 2024). The need for more resilient, low-emission composting strategies has become evident in broader efforts to align compost-based bio-organic fertilizer applications with national greenhouse gas (GHG) mitigation goals (Mulya et al., 2025).

With the rise of Industry 4.0 and precision agriculture, there is growing interest in deploying data-driven and AI-powered methods to monitor, model, and optimize composting processes in a real-time and phase-aware manner. While existing studies have applied soft clustering techniques in isolation, such as Gustafson-Kessel (Giusti and Marsili-Libelli, 2010), or supervised learning on manually labelled dataset (Ding et al., 2022), there is a lack of integrated frameworks that jointly leverage unsupervised soft phase identification and supervised predictive modelling to map latent composting behaviour in a data-driven and interpretable manner. In particular, no prior work has implemented such an approach specifically for poultry litter composting, where latent phase transitions are difficult to detect using conventional rule-based monitoring.

This study proposes a two-stage machine learning framework that integrates soft clustering via Gaussian Mixture Models (GMMs) with snapshot-based supervised learning to perform latent composting phase identification and probabilistic phase mapping using real-time process sensor data in poultry litter composting. GMMs are used to infer composting phases in an unsupervised manner, yielding probabilistic phase assignments that capture uncertainty and gradual transitions—features absent in hard clustering or rule-based methods. These soft labels are then used to train supervised machine learning models that map current process states to phase probability vectors, enabling real-time phase monitoring without requiring manual labels. The framework is validated using data from two full-scale poultry litter composting experiments, with cluster structures and supervised phase mappers evaluated independently. This approach aims to demonstrate an interpretable and scalable solution for intelligent monitoring and analysis of composting processes in agro-industrial settings.

2. Methodology

This study proposes a two-stage machine learning (ML) framework for intelligent monitoring and phase prediction in poultry litter composting processes. The methodology integrates unsupervised soft clustering and snapshot-based supervised learning to predict latent composting phase dynamics using real-time sensor data.

2.1 Data Description and Preprocessing

Data for this study was obtained from a previously published experimental investigation on poultry litter composting using a forced aeration system installed at the San Miguel Research and Training Center of De La Salle University (Patiño-Forero et al., 2023). The system operated in a high-altitude environment (>2700 m above sea level), under mountainous temperature conditions, and incorporated automated monitoring and control components connected via IoT interfaces. The experiments used closed-loop reactors with poultry manure amended with sawdust as a bulking agent and moisture regulator. The corresponding dataset has been made publicly available for reuse (Patiño-Forero, 2024).

Two full-scale composting experiments were carried out under controlled conditions. Experiment 1 generated a dataset consisting of 4,302 time-stamped records, while Experiment 2 consisted of 2,786 records. Both experiments recorded data at 5-minute intervals, capturing high-resolution process dynamics over several days. The monitored process variables included compost pile temperature (Temp. Biopila), relative humidity (Humidity). Although oxygen concentration was recorded in Experiment 2, it was excluded from the supervised learning phase to ensure consistent input structures across both experiments. Actuator signals, such as heater temperature and binary controls for air and heat injection, were also excluded to maintain the focus on intrinsic, passive process dynamics and to avoid potential confounding from system-specific control logic. Records were indexed by cumulative time (cum_minutes) to preserve temporal structure. The retained process variables were standardized using z-score normalization. This ensured that features contributed equally to distance metrics used in clustering and that model training was not biased by scale differences across variables.

2.2 Soft Clustering for Phase Identification

Gaussian Mixture Models (GMMs) were applied to identify latent composting phases based on process data. Unlike hard clustering, GMMs provide probabilistic assignments, allowing smooth representation of phase transitions. Cluster number was selected based on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), with four clusters identified in each experiment. The resulting phase distributions reflected key composting stages: early mesophilic, thermophilic, cooling, and maturation. Clustering was performed separately for each experiment to account for possible structural differences in process dynamics owing to sensor placement and ambient condition variability.

2.3 Supervised Learning for Phase Mapping

To predict composting phase in real time, a snapshot-based supervised learning model was developed. Using only current process sensor readings as inputs, four regression models were trained to estimate each of the soft GMM cluster probabilities. Each model was trained using H2O AutoML, which tested multiple algorithms (Generalized Linear Models, Gradient Boosting Machines, XGBoost, DeepLearning, Distributed Random Forest, and ensemble meta models that combine predictions from multiple aforementioned base models). The best performing model was selected based on RMSE. Models were trained using an 80:20 train-test split and outputs were renormalized to ensure valid probability distributions. This mapping enables real-time estimation of composting phase without temporal data or manual labels, supporting data-driven monitoring and control strategies. A summary flowchart of the two-stage machine learning framework is given in Figure 1.

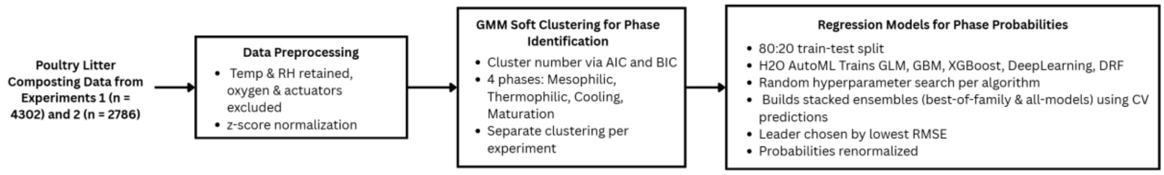

Figure 1: Flowchart of the two-stage machine learning framework for phase identification and prediction in poultry litter composting

3. Results and Discussion

3.1 Soft Clustering and Phase Structure

Gaussian Mixture Models (GMMs) were applied to the preprocessed process data for each experiment to identify latent composting phases. Based on Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), a 4-cluster model was selected for both experiments. The resulting clusters were mapped to canonical composting phases: pre-composting (Cluster 0), mesophilic (Cluster 1), thermophilic (Cluster 2), and maturation (Cluster 3). GMMs captured soft phase boundaries and transitional behaviour typical of biological processes like composting, unlike hard segmentation.

The GMM clustering results are visualized in Figure 2, which presents compost pile temperature over time color-coded by dominant cluster assignment (top row) and the soft cluster probabilities as stacked area plots (bottom row). Both experiments displayed a clear progression through composting stages, with notable overlaps between phases, especially at the onset of mesophilic and late thermophilic phases.

Experiment 1 exhibited relatively balanced phase durations: pre-composting dominated from 0–5,500 minutes, mesophilic from 4,350–9,850 min, thermophilic from 9,500–13,750 min, followed by a brief mesophilic overlap (12,800–14,575 min), and finally maturation from 14,550–21,515 min. This suggests a robust and active composting process with minimal phase reversion. In contrast, Experiment 2 showed a prolonged pre-composting phase, persisting from 0–7,700 min. During this period, soft overlaps with the mesophilic cluster emerged around 6,400–6,950 min, indicating early microbial activity. The mesophilic phase was dominant from 6,950–10,100 min, with thermophilic activity peaking between 9,850–11,950 min. A brief maturation onset appeared between 11,350–11,600 min before thermophilic conditions re-emerged transiently. Full maturation was established from 11,800 min onward. This delayed transition is likely due to the high-altitude (~2,700 m) location and low ambient temperatures, which may have suppressed initial microbial respiration and self-heating.

The GMM soft probability outputs enabled a flexible, data-driven interpretation of composting dynamics, revealing nonlinear phase boundaries and phase co-existence that conventional rule-based methods would fail to detect. This was particularly evident in periods of oscillation or reversion, such as the short-lived return to mesophilic conditions in Experiment 1 after the thermophilic phase. Interestingly, the observed temperature ranges for each phase did not fully align with textbook definitions (e.g., mesophilic: 25–42 °C; thermophilic: >45 °C) (Finore et al., 2023). GMMs identified latent states by capturing multivariate patterns (temperature, humidity, oxygen) instead of fixed thresholds. This is especially important in non-standard composting

environments, such as high altitudes or extreme climates, where absolute temperature-based heuristics may misrepresent compost maturity.
These results confirm that poultry litter composting is a multistage, nonlinear process governed by gradual microbial and thermodynamic transitions. GMM-based soft clustering offers a generalizable and interpretable framework for modelling these dynamics, and forms the basis for the next section, where real-time supervised learning is applied to map phase probabilities from sensor data.

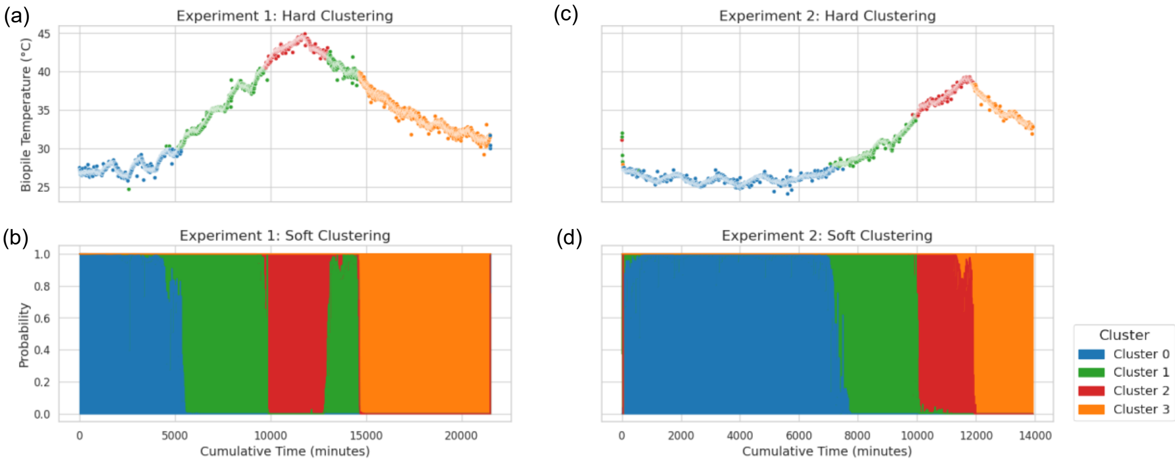

Figure 2: Latent composting phase identification using Gaussian Mixture Models (GMMs): (a) raw temperature profiles (°C) for Experiment 1, color-coded by dominant phase; (b) normalized soft GMM cluster probabilities for Experiment 1; (c) raw temperature profiles (°C) for Experiment 2, color-coded by dominant phase; (d) normalized soft GMM cluster probabilities for Experiment 2

3.2 Supervised Learning for Real-Time Phase Mapping

To enable real-time, label-free inference of composting phases from raw sensor data, a snapshot-based supervised learning module was developed to predict the GMM-derived cluster probabilities directly from instantaneous process variables. This approach facilitates the reconstruction of the soft composting phase structure in real time, supporting intelligent, phase-aware monitoring in automated composting systems without requiring time-series modeling or manual phase labels.
For each experiment, four regression models were trained: one for each latent phase (Clusters 0 to 3, corresponding to pre-composting, mesophilic, thermophilic, and maturation). These models were built using H2O AutoML, which systematically evaluated a suite of algorithms, including XGBoost, Distributed Random Forest (DRF), Gradient Boosting Machines (GBM), Generalized Linear Models (GLM), and Deep Learning networks. Models were ranked based on root mean squared error (RMSE) on a held-out test set (20 % of the dataset), and the best-performing model was selected for each cluster probability target.
Table 1 summarizes the model performance across all targets and experiments. All models demonstrated high predictive accuracy, with R^2 values exceeding 0.99 and RMSE values below 0.032. These results confirm that composting phase probabilities—derived from unsupervised clustering—can be reliably predicted using only real-time temperature and humidity inputs. This reinforces the value of snapshot-based modeling in settings where storage, latency, or hardware constraints limit the feasibility of full sequence-based approaches.

Table 1: Supervised learning performance for predicting GMM-derived composting phase probabilities from sensor data

Dataset	Cluster (Phase)	Best Model	RMSE	MAE	R^2
Experiment 1	Cluster 0 (Pre)	XGBoost	0.03176	0.00670	0.99465
Experiment 1	Cluster 1 (Meso)	DRF	0.00547	0.00084	0.99986
Experiment 1	Cluster 2 (Thermo)	DRF	0.03141	0.00355	0.99437
Experiment 1	Cluster 3 (Mat.)	GBM	0.01319	0.00333	0.99854
Experiment 2	Cluster 0 (Pre)	GBM	0.00412	0.00107	0.99993
Experiment 2	Cluster 1 (Mes.)	GBM	0.00612	0.00110	0.99968
Experiment 2	Cluster 2 (Thermo)	GBM	0.01016	0.00263	0.99928
Experiment 2	Cluster 3 (Mat.)	GBM	0.01106	0.00241	0.99892

An analysis of the H2O AutoML leaderboards provides further insight into model behavior. For Experiment 1, tree-based ensemble models consistently outperformed other architectures: XGBoost for Cluster 0 (pre-composting), DRF for Clusters 1–2 (mesophilic and thermophilic), and GBM for Cluster 3 (maturation). The dominance of ensemble tree models suggests composting phase probabilities are driven by nonlinear interactions (e.g., temperature–humidity), which tree methods capture effectively. In contrast, GLMs performed poorly across all targets, with RMSE values exceeding 0.4 in some cases, highlighting their limitations in modelling nonlinear, high-variance biological data. Similarly, deep learning models underperformed, which is likely due to the low dimensionality of the input data (only two to three features per snapshot) and the absence of temporal dependencies, which neural networks typically exploit.
Figure 3 illustrates the predicted versus actual GMM probabilities for all clusters, confirming close alignment and high predictive accuracy across both experiments. The predicted values track closely along the 1:1 diagonal, indicating minimal bias or variance. Among the four phases, all models achieved similarly strong predictive performance, with only marginal differences in RMSE and MAE across clusters. While the mesophilic cluster (Cluster 1) yielded slightly lower errors in some cases, this variation was minor and not considered substantively meaningful. Overall, the consistency in performance across all clusters indicates that the proposed models are equally capable of estimating all composting phases from snapshot sensor inputs.

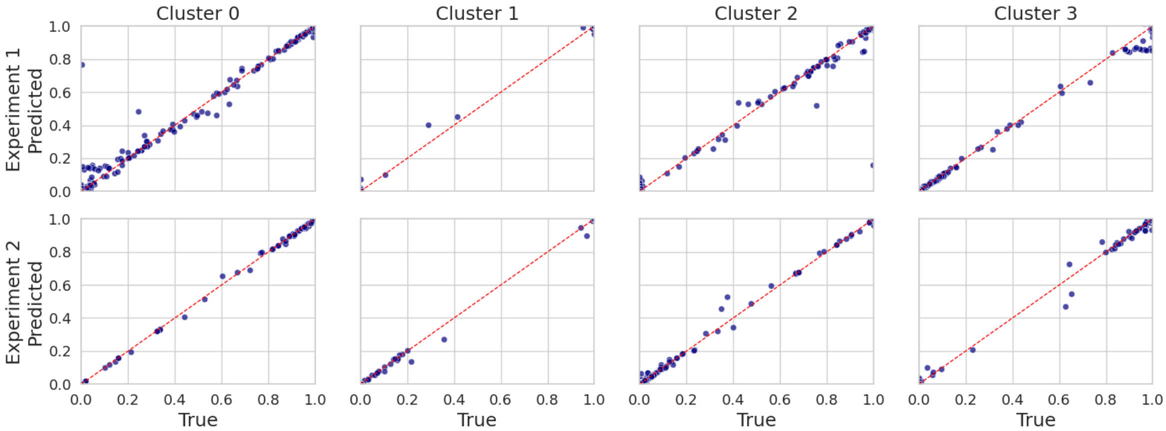

Figure 3: Predicted vs. true GMM-derived phase probabilities for each cluster and experiment. Each point represents a single data instance. All probabilities are normalized such that, for each instance, the sum of phase probabilities equals 1. Strong alignment with the 1:1 line confirms accurate real-time inference from sensor data

It should be noted that the models developed in this study are specific to the composting setup, sensor configuration, and high-altitude conditions of the experimental facility. As such, their predictive performance may not directly transfer to other environments or reactor types. While this study demonstrates the feasibility of combining soft clustering and supervised learning for compost phase prediction, generalization across composting systems remains an open challenge. Retraining with site-specific data may be required, and future work should explore multi-site datasets to develop more robust, transferable models.
In actual field implementations, sensor reliability may be impacted by issues such as fouling, calibration drift, or power/connectivity interruptions. Such realities could also introduce noise or gaps in input data, potentially degrading predictive accuracy. Future deployments should consider sensor accuracy drift and calibration protocols to ensure robust operation in real-world composting environments.
For practical implementation, the proposed framework can be embedded into process control or monitoring systems using programmable logic controllers (PLCs), IoT dashboards, or SCADA interfaces. Model components must be retrained or tuned using historical data specific to the target composting system to reflect local substrate characteristics, system dynamics, and control parameters. This added flexibility contrasts with conventional threshold-based monitoring schemes such as those described by Phu et al. (2023), which rely on static criteria rather than adaptive inference from real-time signals.
That said, these results support the viability of using snapshot-based supervised learning to infer latent composting phase probabilities in real time. The approach requires only instantaneous process measurements and avoids the need for time-series modelling, making it computationally efficient and well-suited for integration into resource-constrained or embedded composting systems. By reconstructing the soft phase structure learned through unsupervised clustering, this method provides a lightweight but powerful tool for intelligent compost

monitoring, with potential applications in anomaly detection, process tracking, and the development of phase-aware control strategies.

4. Conclusions

This study proposed a machine learning framework integrating GMMs for phase identification with supervised learning for real-time composting inference. GMMs identified four latent phases—pre-composting, mesophilic, thermophilic, and maturation—capturing smooth transitions and overlaps often missed by rule-based methods. Supervised models trained using H2O AutoML achieved high predictive accuracy across both experiments, with $R^2 > 0.99$ and RMSE < 0.032 for all phases. Tree ensembles consistently outperformed other algorithms, effectively capturing nonlinear temperature-humidity composting dynamics. The framework enables label-free, low-latency monitoring from real-time sensor data, scalable to resource-limited systems. Although the framework demonstrated excellent performance on the present dataset, its generalizability to other composting systems and environments remains to be validated. Future studies incorporating diverse composting configurations and climatic conditions are necessary to extend the applicability of this approach beyond the current experimental context.

Acknowledgments

This research is funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS), project number “FRGS/1/2024/WAS04/XMU/02/1”. It is also supported through the PublicPrivate Research Network (PPRN), project number “2024/XMUM/772”, and the Xiamen University Malaysia Research Fund (XMUMRF) under projects “XMUMRF/2024-C14/IENG/0075” and “XMUMRF/2023-C12/IENG/0060”.

References

Alarefee H.A., Ishak C.F., Othman R., Karam D.S., 2023, Effectiveness of mixing poultry litter compost with rice husk biochar in mitigating ammonia volatilization and carbon dioxide emission, Journal of Environmental Management, 329, 117051.

Ding S., Huang W., Xu W., Wu Y., Zhao Y., Fang P., Hu B., Lou L., 2022, Improving kitchen waste composting maturity by optimizing the processing parameters based on machine learning model, Bioresource Technology, 360, 127606.

Duarah P., Haldar D., Patel A.K., Dong C.-D., Singhania R.R., Purkait M.K., 2022, A review on global perspectives of sustainable development in bioenergy generation, Bioresource Technology, 348, 126791.

Finore I., Feola A., Russo L., Cattaneo A., Di Donato P., Nicolaus B., Poli A., Romano I., 2023, Thermophilic bacteria and their thermozymes in composting processes: a review, Chemical and Biological Technologies in Agriculture, 10(1), 7.

Food and Agriculture Organization of the United Nations, 2025, Crops and livestock products, FAO <fao.org/faostat/en/#data/QCL> accessed 11.06.2025.

Giusti E., Marsili-Libelli S., 2010, Fuzzy modelling of the composting process, Environmental Modelling & Software, 25(5), 641–647.

Huang L., Hou J., Liu H., 2024, Machine-learning intervention progress in the field of organic waste composting: Simulation, prediction, optimization, and challenges, Waste Management, 178, 155–167.

Mulya K.S., Tan J.P., Yeat S.P., Yeat C.N.C., Farooque A.A., Zhou S., Woon K.S., 2025, Toward carbon mitigation resiliency in the agriculture sector: An integrated LCA-GHG protocol-IPCC guidelines framework for biofertilizer application in paddy field, Journal of Environmental Management, 389, 126005.

Patiño-Forero Á.A., 2024, Dataset Supporting: Dynamic Modeling of Poultry Litter Composting in High Mountain Climates using System Identification Techniques, Mendeley Data, V2.

Patiño-Forero Á.A., Méndez F., Zuluaga D., 2023, Automation of a composting plant, I + T + C Investigación, Tecnología y Ciencia, 1(17).

Phu S.T.P., Asari M., Nguyen D.B., Dinh C.L., 2023, Developing a thermal-composting system for recycling household biodegradable solid waste, Chemical Engineering Transactions, 106, 607–612.

Walling E., Vaneeckhaute C., 2021, Novel simple approaches to modeling composting kinetics, Journal of Environmental Chemical Engineering, 9(3), 105243.

Yeow T.A., Agustin Y.E., Lee X.Y., Ismail A.A.S., Hui Y.W., Rahman N.A., Tan S.T., Ngan C.L., Tan J.P., 2024, Upcycling lignocellulosic palm biomass via Black Soldier Fly Larvae (BSFL) composting incorporated with ex-situ fermentation by Bacillus subtilis, Industrial Crops and Products, 222, 119764.