Proceedings of the 21st ICSMGE, Vienna, Austria, 14 – 19 June 2026. Pistrol, Adam & Schweiger (eds.)
Published by: ÖGG, Austrian Society for Geomechanics, Salzburg, Austria, ISBN 978-3-9503898-4-5
https://doi.org/10.53243/ICSMGE2026-981

An Overview of ML-Based Surrogate Models for MSW Landfill Performance Assessment and Optimization

Jagadeesh Kumar Janga, **Krishna R. Reddy**
Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago, USA, kreddy@uic.edu.

ABSTRACT: A coupled thermo-hydro-bio-mechanical (CTHBM) model was recently developed and validated to holistically assess municipal solid waste (MSW) landfill performance. The model integrates a two-stage anaerobic biodegradation model, an elasto-visco-bio-plastic mechanical model, a heat conduction model, and a two-phase flow model, capturing crucial coupled-process interactions to predict key landfill indicators such as methane generation, temperatures, and settlements, among others. Although the CTHBM model provides a robust framework for landfill performance assessment, these simulations can become computationally expensive, particularly when numerous high-resolution and large-scale simulations are required for uncertainty quantification and engineering optimization. To address this, a surrogate-modeling approach was adopted, where machine learning (ML)-based prediction models were trained on simulation data generated by the CTHBM model. The data used to train the AI models was obtained by varying key input variables, including placement conditions, waste properties, and leachate injection parameters. Several ML models were assessed, including random forests and extreme gradient boosting (XGBoost), support vector regression (SVR), and artificial neural networks (ANNs). These models were trained to predict different landfill performance indicators, including stabilization period, settlement, cumulative methane generation, and the evolution of maximum temperatures. The results demonstrated that AI-based surrogate models could accurately predict various landfill performance indicators when the hyperparameters are carefully tuned. Specifically, SVR and ANN were able to accurately predict stabilization periods ($R^2 \sim 0.95$), while ANN provided the best performance in settlement predictions ($R^2 \sim 0.99$). Further, exceptional accuracy was obtained using Bayesian ANNs for methane and maximum temperature evolution predictions. Finally, an optimization algorithm leveraging Bayesian techniques – using predictions from the AI-based surrogates – was developed to control leachate injection operations aimed at maximizing energy recovery, while regulating landfill temperatures within safe limits. Overall, this study highlights the potential of AI-based surrogate models for computationally efficient landfill performance assessment and optimization, offering a scalable solution for sustainable landfill management.

KEYWORDS: Municipal solid waste, Landfill, Coupled-process modeling, Machine learning, Bayesian optimization.

1 INTRODUCTION

A comprehensive coupled thermo-hydro-bio-mechanical (CTHBM) model was previously developed for holistic landfill performance assessment (Kumar and Reddy 2021a). CTHBM model uses a two-phase-flow model to simulate gas and liquid flow within the municipal solid waste (MSW) matrix, a heat conduction model based on Fourier's heat transfer equation to simulate the heat transport, and an elasto-visco-bio-plastic model to simulate mechanical stress-induced strains, biodegradation-induced strains, and long-term creep-induced strains. Furthermore, to simulate the biodegradation of organics in MSW, the CTHBM model employs a two-stage anaerobic biodegradation model. In addition to capturing individual processes, the CTHBM model accounts for their coupled interactions such as: how moisture availability influences biodegradation, and in turn, how biodegradation affects moisture content through water consumption; how heat generated from biodegradation elevates landfill temperatures and how landfill temperatures effect biodegradation rates; and how mass loss from biodegradation impacts mechanical strain development among others. Accounting for these detailed processes and their coupled interactions, the CTHBM model provides a valuable tool for assessing the performance of landfills in both short- and long-term thereby informing the design and operational strategies for landfill components such as leachate recirculation systems.

The CTHBM model has been successfully validated using various laboratory and pilot-scale studies (Kumar and Reddy 2020a, b). Furthermore, the applicability of the CTHBM model in assessing the spatial and temporal variability in landfill settlements, biodegradation rates, and temperature changes among other entities has been successfully demonstrated in Kumar and Reddy (2021a) using a full-scale landfill cell, while the usefulness of the same for informing leachate recirculation strategies through parametric studies has been demonstrated in Kumar and Reddy (2021b).

Although the CTHBM model enables robust and generalizable predictions of landfill behavior, its computational demands, extensive parametrization involved, and difficulty of implementation pose challenges in its practical application. Furthermore, landfills present several opportunities for resource recovery, including energy generation from landfill biogas, heat recovery from elevated internal temperatures, and beneficial site reuse (e.g., landfill mining and site reclamation) following stabilization. However, efficient recovery of these resources while ensuring safe landfill operations through controlled leachate recirculation necessitates evaluating numerous design and operational scenarios. Using the CTHBM model for such large-scale optimization will be computationally intractable and time-consuming.

In this regard, recent advancements in artificial intelligence (AI), machine learning (ML), and deep learning (DL) offer an excellent opportunity to develop computationally efficient prediction tools for complex systems such as landfills. In the context of landfills, prior studies have applied ML models to experimental data for predicting landfill gas generation and/or settlement (Janga and Reddy 2025a). However, these datasets are typically limited by site-specific waste characteristics and experimental conditions, restricting the generalizability of such tools. In this regard, surrogate models – which are ML or DL models trained on outputs of numerical models to emulate their behavior – provide an excellent option for computationally efficient prediction tools. The use of ML models as surrogates to complex numerical models simulating sub-surface domains such as groundwater flow modeling, and geotechnical engineering problems among others has seen an upward trend in the past decade (Luo et al. 2023, and Furtney et al. 2022), however, the use of the same for landfill performance-prediction and subsequent optimization is previously unexplored.

In this regard, we have recently evaluated several ML models as surrogate models to the CTHBM model for predicting crucial landfill performance indicators including

landfill stabilization period (Janga and Reddy 2025b), final surface settlements (Janga and Reddy 2025c), CH_4 generation, and maximum temperatures (Janga and Reddy 2025d) with an objective to develop a robust practical framework using ML surrogate models – informed by the CTHBM model – for computationally efficient prediction tools that can be useful for large-scale optimization enabling maximum resource recovery while ensuring landfill safety. This study provides an overview of different surrogate models developed by the authors to date in this regard and provides directions for future research.

2 MACHINE LEARNING BASED SURROGATE MODELS

2.1 *Data generation using the CTHBM model*

The data for training the ML models is generated using the CTHBM model. Comprehensive descriptions of the thermal, hydraulic, biodegradation, and mechanical models and how these processes are coupled within the CTHBM framework are detailed in Kumar and Reddy (2021a). These details are omitted here for brevity purposes.

The landfill cross-section shown in Figure 1 is utilized to simulate long-term landfill behavior and generate data with varying placement conditions, waste properties, and leachate recirculation parameters.

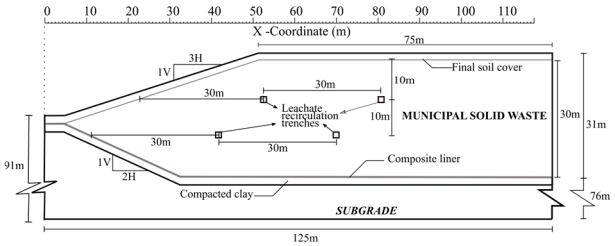

Figure 1. Full-scale landfill cross-section employed in this study

The input variables that were varied to generate the data include placement conditions (placement moisture content and dry density), initial solid degradable fraction, thermal properties (thermal conductivity and specific heat capacity), and waste degradability (maximum hydrolysis rate). The hydraulic conductivity is simulated as a function of void-to-inert phase ratio in the CTHBM model, and hence it varies with varying placement density and solid degradable fraction. Additionally, injection pressure and duration were varied to simulate various operational conditions for leachate recirculation. A total of 275 simulations were performed.

Outputs of CTHBM simulations include the spatial distribution of temperatures with time, the spatial distribution of the degree of degradation with time, the cumulative methane (CH_4) gas production from the landfill with time, and the spatial distribution of waste displacements with time, among others. For the purpose of developing the surrogate models, the following data has been retrieved from all the CTHBM simulations: landfill stabilization period (time taken to generate 95% of the total methane generated in a simulation), final surface settlements at various spatial locations (X = 50, 60, 70, 80, 90, 100, 110, and 120 m) at the end of simulation period, cumulative methane generation with time, and maximum temperature in the landfill at different times.

The number of data points used to develop the surrogate models varied depending on the output entity. To predict the landfill stabilization period, a total of 275 data points were used to train different ML models (Janga and Reddy 2025b). To predict surface settlement of landfills, 2184 data points were used (settlements at 8 different spatial locations for 273 simulations) (Janga and Reddy 2025c), and to predict the temporal evolution of cumulative CH_4 generation and maximum temperatures, 93,759 data points were used (300-360 time points per simulation from 274 simulations) (Janga and Reddy 2025d). This data has been split into training and testing sets for the ML-based surrogate model development and testing. Details regarding the input variables used to train different surrogate models to predict different indicators, and the corresponding ML models assessed, along with the hyperparameter tuning method used, are presented in Table 1.

Table 1. Inputs and outputs of various surrogate models and the corresponding ML models assessed as surrogates to the CTHBM model, along with the method used for hyperparameter tuning.

Landfill Performance Indicator (Y)	**Inputs (X)**	**ML models assessed**	**Hyper-parameter tuning**
Stabilization period	Main variables[#]	RF, XGBoost, SVR, ANN.	Bayesian optimization
Surface settlements	Main variables[#] + X-coordinate*	RF, XGBoost, SVR, ANN.	Bayesian optimization
Total CH_4 generated at a given time (t_i)	Main variables[#] + time (t_i)	BNN	Manual
Maximum temperature at a given time (t_i)	Select variables[1] + CH_4 generated at (t_i) + time (t_i)	BNN	Manual

[#]***Main variables*** *include: placement moisture content, placement density (dry), initial solid degradable fraction, maximum hydrolysis rate, thermal conductivity of waste, specific heat capacity of waste, leachate injection pressure, leachate injection duration*
X-coordinate *of the spatial location where surface settlement is being predicted*
[1]**Select variables:** Variables that mainly influence biodegradation (maximum hydrolysis rate, and injection parameters) were replaced with predicted CH_4 generation at time t_i.
Abbreviations*: RF – Random Forest, XGBoost – Extreme Gradient Boosting, SVR – Support Vector Regression, ANN – Artificial Neural Network, BNN – Bayesian Neural Network*

2.2 *Surrogate models' performance*

The input variables in the data retrieved from the CTHBM model simulations were pre-processed using a standard scaling technique (normalization based on mean and standard deviation values) to avoid the influence of differences in magnitudes of different variables. The data was standardized using only the training data to avoid unintended propagation of information regarding the test data during the model training. Further details on detailed data distribution in training and testing sets for different surrogate models, a brief background on various ML models assessed, and the hyperparameter tuning process can be found in the respective studies (Janga and Reddy 2025b, c, d).

2.2.1 *Landfill stabilization period*

Of the four ML-based surrogate models – trained using the data from the CTHBM model for their efficacy, SVR and ANN showed excellent accuracy in predicting the landfill stabilization period (years), while tree-based ensemble models, RF and XGBoost, seemed to overfit the data, as high testing errors were observed. The coefficient of determination (R^2) and root mean squared error (RMSE) (years) of predictions made by the ML models on previously unseen test data for the four models are as follows: RF – 0.75 and 1.95, XGBoost – 0.77 and 1.87, SVR – 0.96 and 0.80, and ANN – 0.94 and 0.93.

The predictions of the best-performing models were interpreted using partial dependence plots (PDPs) and SHAP values to understand the variability in surrogate model predictions with variations in different input variables. The results showed that the surrogate model prediction trends appropriately represented the underlying process relationships. For example, the effects of leachate injection parameters were

only significant in landfill with lower initial moisture contents; increasing moisture content values resulted in lower stabilization period predictions indicating faster stabilization; increasing organic fraction resulted in increase in stabilization period predictions; and increase in placement density (dry) values resulted in lower stabilization period. Similar interpretations were made for interactions between different variables, and the surrogate model predictions showed agreement with underlying process relationships. Further details on the performance of different models and their PDPs can be found in Janga and Reddy (2025b).

2.3 *Final surface settlement at different spatial locations*

Four machine learning (ML)-based surrogate models were trained using CTHBM simulation data to predict final surface settlement (m) at various spatial locations. Among them, the artificial neural network (ANN) with a single hidden layer (20 neurons, ReLU activation) demonstrated excellent generalization to testing data, achieving an R^2 of 0.99 and an RMSE of 0.09 m. In comparison, SVR and RF exhibited higher prediction errors, even while predicting at the same spatial locations used in training. While XGBoost yielded high accuracy when predicting at the same spatial locations as the training data, it failed to generalize to new locations, as demonstrated in a representative case study. This limitation stems from the limited number of spatial locations included in the training set and the tendency of tree-based models like XGBoost to treat such features as categorical rather than continuous. Additional details on the comparative performance of these ML models can be found in Janga and Reddy (2025c).

2.4 *Cumulative CH_4 generation with time*

As demonstrated in the previous sections, artificial neural networks (ANNs) effectively captured the behavior of the CTHBM model. However, like all data-driven surrogate models, their predictive performance is inherently constrained by the range and diversity of the training data. Predictions made beyond this range can be unreliable. Therefore, it is critical to recognize the limitations of the training data and interpret model predictions with appropriate caution.

One approach to address such limitations and improve the reliability of surrogate model predictions is by quantifying the epistemic uncertainty – the uncertainty arising from a lack of knowledge – associated with model predictions. Several techniques are available to estimate epistemic uncertainty, including Bayesian learning via variational inference (VI), Monte Carlo (MC) dropout, and ensemble methods, among others. We employed VI to estimate uncertainty (Janga and Reddy 2025d). ANNs with probabilistic model parameters trained using VI are commonly referred to as Bayesian Neural Networks (BNNs). In BNNs, the model parameters (weights and biases) are treated as probability distributions – typically Gaussian – which are learned during training. When a BNN is queried multiple times with the same input, it yields slightly different outputs, reflecting the uncertainty in the model's knowledge. The mean of these multiple predictions is treated as the final prediction, while the spread (e.g., standard deviation or confidence interval) represents the associated uncertainty.

Therefore, a BNN with two hidden layers (each containing 20 neurons, ReLU activation on layer 1 and Sigmoid on layer 2) was trained to predict cumulative CH_4 generation (as a percentage of the CH_4 potential) over time. The model achieved excellent predictive accuracy, with R^2 = 0.9988 and RMSE = 1.07%. Validation against 18 previously unseen test cases showed strong agreement with CTHBM model outputs. Notably, wider uncertainty bands were observed for test cases drawn from regions with sparse training data, highlighting the lower confidence. Further interpretation of surrogate model behavior through PDPs showed that the surrogate model has effectively captured the underlying process relationships between the input variables and the predicted outputs.

2.5 *Maximum temperature in the landfill with time*

We trained a BNN with 2 hidden layers (10 neurons each, Sigmoid activation on layer 1, and Tanh on layer 2) as a surrogate to predict the maximum temperature evolution. However, only selected inputs were used for training the BNN to predict maximum temperature, as listed in Table 1, to remove redundant variables and the associated noise. The resulting model performance showed that the model could accurately predict CTHBM data (R^2 = 0.9916 and RMSE = 1.61°C). Testing the model's performance on previously unseen cases showed excellent agreement between the surrogate model predictions and CTHBM simulation results, although higher uncertainty bands were observed in samples derived from regions with sparse training data. PDPs used to interpret the surrogate model behavior confirmed that the surrogate model was able to accurately capture the underlying process relationships between input and output variables.

Furthermore, using the surrogate model predictions obtained for CH_4 generation and for maximum temperature, an optimization case study was implemented to showcase the efficiency of a surrogate-based framework in identifying optimal operational scenarios that can result in maximum energy recovery. In this regard, a full-scale landfill cell studied by Kumar and Reddy (2021a), shown in Figure 1, was employed alongside the waste properties and placement conditions that are typical of a Midwestern landfill. The BNN-model-based optimization framework used was aimed at maximizing the biogas energy recovery efficiency and heat energy recovery potential, while ensuring the maximum temperatures in the landfill do not exceed a predefined threshold of 65°C to ensure landfill safety. The outcomes of the optimization indicated that the proposed framework successfully identified the optimal leachate injection parameters that can improve energy recovery potential from landfills. Refer to Janga and Reddy (2025d) for further details on the BNN model details, and optimization case studies.

2.6 *Variable importance in surrogate model predictions*

The importance of different variables has been calculated using permutation variable importance for all the developed surrogate models. These importance values, in absolute terms, measure the change in model outputs or model accuracy (R^2 or RMSE) by randomly shuffling one feature at a time. To assess the relative importance of each variable in predicting different landfill performance indicators in various surrogate models, we ranked the variables based on their absolute importance values as obtained in the respective studies. The resulting rankings are presented in Table 2.

These variable importance rankings in surrogate model predictions show that these models, although purely data-driven, capture the underlying relationships adequately. As seen, for variables including biodegradation (cumulative methane generation and stabilization period), moisture content was proven to be the most influential variable. Whereas for predicting settlements, initial solid degradable fraction and placement density (dry) were shown to be of highest importance, as these variables mainly affect the amount of organic solids present in the MSW matrix, which is the main influencing factor that decides the amount of biodegradation-induced settlement. These rankings show that surrogate models reasonably capture the nuances in underlying process relationships between input variables and outputs.

Table 2. Variable importance rankings for the best-performing surrogate models

X	Stabilization period	Final surface settlement[#]	CH_4 generation[1]	T_{max}*,[1]
MC	1	4	1	3
SDF_0	3	1	2	2
DD	5	2	4	5
b	4	7	7	7
TC	2	8	3	1
SHC	7	6	5	4
IP	6	3	8	8
ID	8	5	6	6
Ref.	*I*	*II*	*III*	*III*

- **Abbreviations:** MC - placement moisture content, SDF_0 - initial solid degradable fraction, DD - placement density (dry), b - maximum hydrolysis rate, TC - thermal conductivity, SHC - specific heat capacity, IP - injection pressure, ID - injection duration
- [#]X-coordinate of the spatial location was the third most important variable for the corresponding surrogate model; however, to maintain consistency for the analysis presented in this table, we excluded it from the rankings provided.
- [1]Time was the most important input variable as per the original analysis performed on the surrogate models; we excluded the same here to ensure consistency in the rankings provided.
- *T_{max} – Maximum temperature in the landfill
- *b, IP, and ID were not direct input variables to the surrogate model. These input variables were replaced with CH_4 generated at a given point in time (which was ranked second most important, but excluded here for consistency). Hence, the importance of these variables was considered the same as that in CH_4 prediction.
- *I, II, III – Janga and Reddy (2025b, c, d) respectively*

3 FUTURE RESEARCH

The performance of surrogate models that were developed as part of our initial research, as described in this paper, proves the efficacy of ML-based surrogate modeling in computationally efficient and accurate predictions of landfill indicators. However, the data generated for this research is specific to the landfill cross-section and leachate injection locations presented in Figure 1. Since these configurations vary with site conditions, the surrogate models developed here are not directly generalizable to landfills with different geometries. Future research should include additional variables that correspond to the landfill geometry (e.g., height and width of a landfill) and leachate injection configurations (e.g., number of injection trenches), to improve the generalizability of the surrogate models for broader applications.

The surrogate models developed in this study are data-driven and might not extrapolate well outside the training data range. Generating data that corresponds to the vast-diversity of landfill waste properties and geometries can be computationally prohibitive, given the computational costs associated with the CTHBM model. In this regard, physics-informed machine learning, where simplified differential equations describing the underlying process are included in the loss function in addition to the data loss during model training, can help in making the surrogate models generalize well outside the training data range, while also reducing the associated data requirements (Chen et al. 2023, Lan et al. 2023, Cuomo et al. 2022). Furthermore, the developed surrogate models, along with the CTHBM model, should be validated further using field monitoring data to improve the reliability of the predictions.

4 CONCLUSIONS

This study provided an overview of our initial research on developing machine learning (ML)-based surrogate models for the coupled thermo-hydro-bio-mechanical (CTHBM) model to provide computationally efficient prediction tools for landfills. ML models were trained on CTHBM data to predict landfill stabilization period, final surface settlements, CH_4 generation with time, and maximum temperature with time. To predict landfill stabilization period and settlements, four different ML models were assessed, of which artificial neural networks (ANNs) were commonly able to predict both these variables with reasonable accuracy. To improve the reliability of surrogate model predictions, uncertainty-aware Bayesian neural networks (BNNs) were trained to predict CH_4 generation and maximum temperature with time. Both BNNs showed exceptional accuracy in predicting CTHBM model data. Surrogate model behavior was further interpreted using partial dependence plots and permutation feature importance, confirming the efficacy of surrogate models in capturing the underlying process relationships. Overall, ML-based surrogate models have exceptional promise in serving as computationally efficient prediction tools. These can be further useful for broader engineering optimization efforts aimed at maximizing energy recovery through controlled leachate recirculation.

5 REFERENCES

Chen, Y., Xu, Y., Wang, L., and Li, T. 2023. Modeling water flow in unsaturated soils through physics-informed neural network with principled loss function. *Computers and Geotechnics*, 161, 105546.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F. 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3), 88.

Furtney, J. K., Thielsen, C., Fu, W., and Le Goc, R. 2022. Surrogate models in rock and soil mechanics: Integrating numerical modeling and machine learning. *Rock Mechanics and Rock Engineering*, 55, 2845–2859.

Janga, J. K., and Reddy, K. R. 2025a. Artificial intelligence prediction of landfill gas generation and settlement. In *Roshan Dash, R., Mohapatro, S., Behera, M. (eds) Pollution Control for Clean Environment — Volume 2,* pp. 231-241. Springer, Singapore.

Janga, J. K., and Reddy, K. R. 2025b. Data-driven prediction of landfill stabilization period using interpretable machine learning. *Computers and Geotechnics*, 183, 107202.

Janga, J. K., and Reddy, K. R. 2025c. Data-driven machine learning surrogate to CTHBM model for MSW landfill settlement prediction. In: Proceedings of Geo-Congress 2026. ASCE (Under review)

Janga, J. K., and Reddy, K. R. 2025d. Optimization of Leachate Recirculation for Enhanced Landfill Energy Recovery: A Bayesian Framework with Neural Network Surrogates. *ACS ES&T Engineering*. DOI: 10.1021/acsestengg.5c00141

Kumar, G., Reddy, K. R., and McDougall, J. 2020a. Numerical modeling of coupled biochemical and thermal behavior of municipal solid waste in landfills. *Computers and Geotechnics*, 128, 103836.

Kumar, G., Reddy, K. R., and Foster, C. 2020b. Modeling elasto-visco-bio-plastic mechanical behavior of municipal solid waste in landfills. *Acta Geotechnica*, 16, 1061-1081.

Kumar, G., and Reddy, K. R. 2021a. Comprehensive coupled thermo-hydro-bio-mechanical model for holistic performance assessment of municipal solid waste landfills. *Computers and Geotechnics*, 132, 103920.

Kumar, G., and Reddy, K. R. 2021b. Effects of leachate recirculation system variables on long-term bioreactor landfill performance using coupled thermo-hydro-bio-mechanical model. *International Journal of Geomechanics*, 21(5), 04021059.

Lan, P., Su, J., and Zhang, S. 2023. Surrogate modeling for unsaturated infiltration via the physics and equality-constrained artificial neural networks. *Journal of Rock Mechanics and Geotechnical Engineering*, 16(6), 2282-2295.

Luo, J., Ma, X., Ji, Y., Li, X., Song, Z., and Lu, W. 2023. Review of machine learning-based surrogate models of groundwater contaminant modeling. *Environmental Research*, 238, 117268.