Materials Research Proceedings 59 (2026) 162-169 https://doi.org/10.21741/9781644903957-21

Prediction of Solvent Component and Composition for Absorption-Based Acid Gas Removal Unit using Optuna-LightGBM and K-means

Rafi Jusar WISHNUWARDANA[1,a*], Madiah OMAR[2,b], Haslinda ZABIRI[1,c], Kishore BINGI[3,d], Rosdiazli IBRAHIM[3,e]

[1]Chemical Engineering, Universiti Teknologi PETRONAS, Perak, Malaysia

[2]Integrated Engineering Department, Universiti Teknologi PETRONAS, Perak, Malaysia

[3]Electrical and Electronic Engineering Department, Universiti Teknologi PETRONAS, Perak, Malaysia

[a]rafi_24002236@utp.edu.my, [b]madiah.omar@utp.edu.my, [c]haslindazabiri@utp.edu.my, [d]bingi.kishore@utp.edu.my, [e]rosdiazli@utp.edu.my

Keywords: Acid Gas Removal Unit, Solvent Component, Solvent Composition, LightGBM, Optuna, K-Means

Abstract. Acid gas removal unit (AGRU) is a pivotal component of a natural gas processing plant. The primary purpose of acid gas removal is to reach the industrial pipeline standard of H_2S below four ppm and CO_2 below 2% per volume for pipeline quality. The most widely used technique is an absorption-based AGRU using amine as a solvent. MDEA is the most utilized solvent but has the drawback of absorbing CO_2. The mixture of other amine and physical solvents is necessary to assist the absorption of CO_2. However, the main problem of mixing solvents is parameter complexity. The machine learning method is utilized to find the most optimal solvent based on its operational parameters. LightGBM tuned with Optuna are used to classify the solvent component, followed by K-means to identify solvent composition. The algorithm is applied to six different solvent blends and two feed gas compositions, resulting in 37,786 data points. The LightGBM model tuned with Optuna performed excellently with accuracy of 0.98 and training time under 0.2 seconds. K-means showed the silhouette score averaging 0.5, showing that the data is not well clustered. This model demonstrates reliable capability in analyzing and distinguishing the solvent component and its respective composition.

Introduction

Natural gas was presented as the most environmentally friendly fossil fuel, with the CO_2 emission 23.2% and 57.4% lower than oil and coal, respectively [1]. The purification of natural gas from dangerous gas such as H_2S and CO_2 is critical in industrial chemistry to guarantee the quality of product gas. Hence, it is essential to fulfil the industrial pipeline standard of H_2S below 4 ppm and CO_2 below 2% per volume for pipeline quality [2,3]. To achieve this feat, the removal of H_2S and CO_2 from natural gas is essential.

In recent developments, several techniques have been implemented in the AGRU process, including physical and chemical solvent absorption, adsorption, membrane, etc. However, among all the techniques, absorption-based AGRU using amine as a chemical solvent is widely applied in the chemical industry [4-6]. Methyldiethanolamine (MDEA) is the most used amine considering its selectivity in absorbing H_2S [7]. Despite its selectivity over H_2S, it has a drawback in MDEA reacting with the CO_2. To tackle that drawback, MDEA needs to be blended with an amine that can assist the absorption of CO_2. Besides using a chemical solvent, the blending of MDEA with sulfolane (SFL) shows promising capability in enhancing the absorption of H_2S and energy saving. Also, MDEA, SFL, and (piperazine) PZ have been discovered to have the same ability as the SFL blend but have advantages at the CO_2 absorption [8]. Based on all previously mentioned research,

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 license. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under license by Materials Research Forum LLC.

the major problem of solvent blending requires careful optimization of operational conditions to achieve the best performance. Hence, the data-driven approach is growing in interest.

This paper will develop a more general model for determining solvent components as well as composition on six different solvents blending such as MDEA + SFL + PZ, MDEA +PZ, MDEA+ SFL, MDEA + (diglycoamine) DGA, MDEA + (methylethanolamine) MEA, and MDEA + Diisopropylamine (DIPA). At the same time, two different feed gas compositions are applied to the model, providing a large amount of the dataset for the model. The model employed in this paper is LightGBM for classifying the solvent component, and K means to determine the solvent composition ranges. LightGBM will be tuned with the Optuna Hyperparameter optimization framework to ensure high accuracy and rapid training on the model.

Methodology

In this paper, solvent components and composition of absorption-based AGRU were determined by employing Optuna-LightGBM model integrated with K-means models. The main pipeline (see Fig. 1).

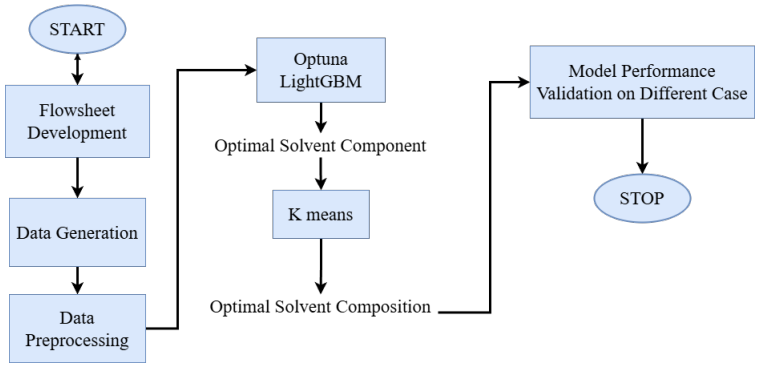


Figure 1 *Research flowchart.*

Flowsheet development

As mentioned beforehand, absorption-based AGRU using amine is extensively utilized in industry. To recreate the actual operation of the AGRU process, flowsheets were developed in Aspen HYSYS software as portrayed in Fig. 2, which has been validated in previous work [9] by using a case study from real plant data, resulting in an acceptable average error of 2%. Therefore, the developed flowsheet is suitable for generating data.

Data generation

Data can be retrieved from the validated flowsheet by simulating different values of a variable using the Aspen Simulation workbook. Several variables that influence the AGRU process are selected based on the previous literature. Large datasets can improve the model accuracy [10], and better data generalization [11]. Table 1 shows the variables that are going to be generated in this research based on the previous literature [9]. The parameters are defined using a specific framework with the Case study tool embedded in Aspen HYSYS software, with each case study generating 201 samples. To develop a generalized model, different datasets will be generated. It is critical to test the generalization ability of the model on different datasets by observing its performance error [15]. The dataset will be generated based on the feed gas composition that is summarized in Table 2. Case 1 will be utilized as the base dataset for model development.

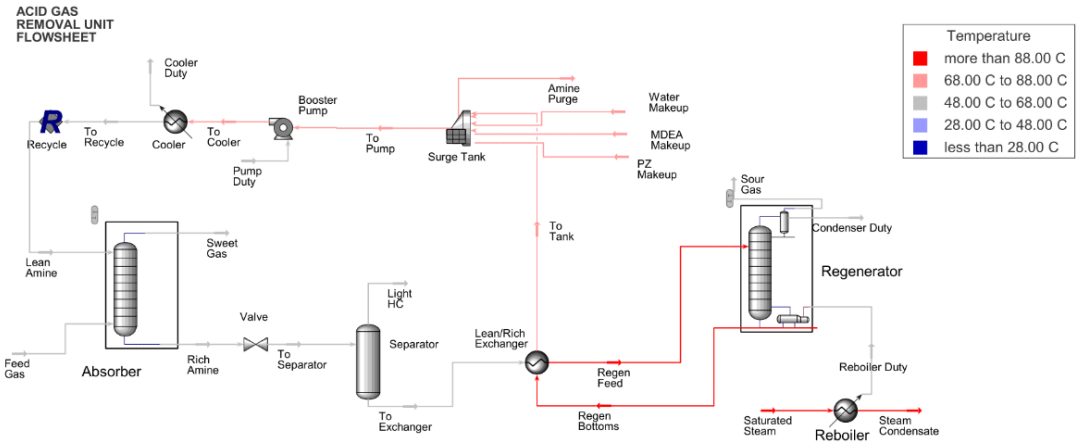


Figure 2 *Developed Aspen HYSYS flowsheet.*

Table 1 *Generated parameters in simulation.*

Variables		Minimum	Base	Maximum	Reference
MDEA + PZ	MDEA (wt%)	35	50	45	[9]
	PZ (wt%)	0	5	10	
MDEA + SFL	MDEA (wt%)	30.8	44	57.2	[8]
	SFL (wt%)	1.05	1.5	1.95	
MDEA + DGA	MDEA (wt%)	30.8	44	57.2	
	DGA (wt%)	1.05	1.5	1.95	
MDEA + SFL + PZ	MDEA (wt%)	30.8	44	57.2	
	SFL (wt%)	0.525	0.75	0.975	
	PZ (wt%)	0.525	0.75	0.975	
MDEA + MEA	MDEA (wt%)	8.4	12	15.6	[12]
	MEA (wt%)	19.6	28	36.4	
MDEA + DIPA	MDEA (wt%)	0	15	30	[13]
	DIPA (wt%)	30	15	0	
Temperature (°C)		38.50	55.00	71.50	[14]
Absorber Pressure (bar)	Feed	38.12	54.02	69.93	
	Lean	38.33	54.32	70.32	
	Top	37.3	52.85	68.41	
	Bottom	37.64	53.33	69.03	

Data pre-processing

In this research, data pre-processing begins by defining the model's input and output. Lean amine temperature, absorber pressure, H_2S and CO_2 composition on sweet gas are the inputs, while the solvent component and its composition are the targeted outputs. The next stage is changing the category for the solvent component into a numerical form. Furthermore, the data is moved to standardization to prevent bias because of inevitable feature domination [17]. The subsequent step involves data splitting into training and testing datasets. This research will apply 80% training data and 20% test data since training data should significantly outweigh testing data, ideally by three to four times [18]. The last step is balancing data using Synthetic minority oversampling technique (SMOTE) [19]. The purpose of data balancing is to prevent model to lean toward the majority class and performing poorly in the minority class

Table 2 *Sour gas composition.*

Component	Mole fraction	
	Case 1 [14]	Case 2 [16]
H2O	0.00234	0
i-Pentane	0.0052	0
n-Pentane	0.0056	0.0024
i-Butane	0.01243	0
n-Butane	0.0215	0.0018
CO_2	0.03001	0.0535
H_2S	0.00024	0.0005
Methane	0.72583	0.8086
n-Ethane	0.11717	0.0163
i-Ethane	0.07377	0
n-Propane	0.00234	0.0035
N2	0	0.1134

Model development

This research employed LightGBM to classify solvent components on AGRU. The key strength of LightGBM is its capability to achieve high accuracy while operating on less memory, making it compatible with regression and classification tasks [20]. Moreover, a leaf-wise strategy is utilized in this algorithm resulting in fewer nodes and enhanced computational performance, diverging from conventional methods [21]. Furthermore, Optuna hyperparameter tuning framework will be applied on LightGBM by following the steps below,

1. Determining the optimization goal and selecting hyperparameter boundaries. Two target function are defined, which are accuracy of more than 0.98 and a training time of less than 0.7 seconds. The chosen hyperparameters for LightGBM are summarized in Table 3.

Table 3 *LightGBM Hyperparameter ranges.*

Hyperparameter	Range
learning_rate	0.001 - 0.4
num_boost_round	3 to 100
max_depth	1 - 300
num_leaves	2 - 300
lambda_l1 and lambda_l2	$1e^{-12}$ - 0.03

2. Perform the optimization. Optuna was configure for having 2,000 optimization trial.
3. Implementing the final model and analyzing its performance. The best-performing hyperparameters from the trial data were extracted, and a LightGBM model was latter trained using these optimal values.

Lastly, K-means is applied to determine the solvent composition. The output will be the optimal concentration range of amine solvent. K-means is one of the unsupervised learning methods that has been proven by its capability and extensive application in current research and industry [22]. K-means works by initialising k cluster centres through random selection, defining the difference between each object and the initial clustering centre, and assigning it to the nearest clustering centre [23].

Model performance metrics evaluation

The performance of the LightGBM model was evaluated using metrics such as accuracy and training duration. Accuracy is calculated as follows,

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative prediction of the model. To evaluate the performance of the K-means model, evaluation metrics such as the Silhouette Score and Davies Bouldin Score were used to assess the performance of the unsupervised algorithm. These performance metrics are calculated as follows,

$$Silhoutte\ Score = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (2)$$

where b(i) is the smallest average distance between i and some other clusters and a(i) is the average distance between i and all other points of its cluster,

$$Davies - Bouldin\ Score = \frac{1}{k}\sum_{i=1}^{k}\left(\max_{i\neq j}\frac{s_i+s_j}{d_{ij}}\right) \quad (3)$$

where k Is the number of the clusters, si is the average distance between each point of the cluster i and its centroid (same for cluster j and sj), and dij Is the distance between the i'th and j'th cluster centroid.

Results and Discussion

Analysis of generated data

The correlation heatmap of solvent composition with the operating parameters such as temperature of the amine solvent, absorber pressure, and the composition of H_2S and CO_2 in the product gas\sweet gas are illustrated in Fig. 3. Generally, there is a significant correlation between the solvent and the temperature and pressure. The correlation heatmap shows the distinctive trait of each solvent blend on its capability to absorb H_2S and CO_2. For instance, there is a negative relationship between MDEA and all solvent blending with the H_2S, indicating that MDEA with a higher mass fraction tends to absorb H_2S than CO_2. This is why secondary amines such as PZ, SFL, DGA, MEA, and DIPA are needed to assist in the absorption of CO_2 from sweet gas. It can be observed that most of the secondary amines have a negative relationship with CO_2, such as PZ, SFL, DGA, and MEA, except DIPA. Given the personal characteristics of each blend, the classification of solvent components is necessary, followed by the determination of solvent composition.

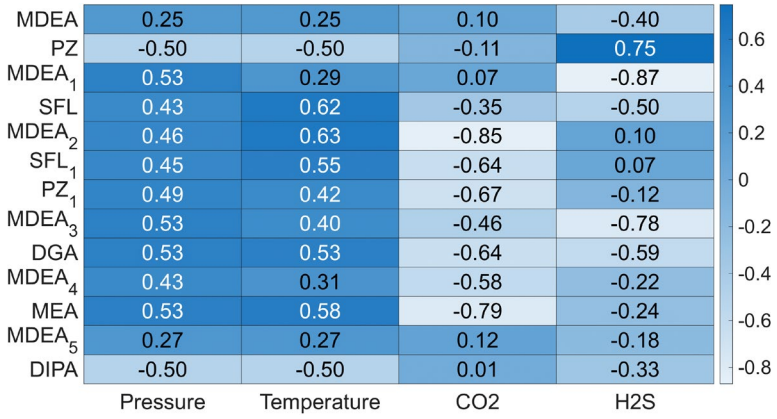


Figure 3 *Correlation matrix of solvent composition and operating parameters.*

Model tuning and validation using different dataset

Before discussing the model validation first, LightGBM model hyperparameter tuning with the Optuna was performed, resulting in a high accuracy of 0.98 and a training time of 0.17 seconds after performing the iteration of 2000 combinations of hyperparameters. Hyperparameter tuning was done on Scenario 1 as the base case before being validated in another case for generalization. Hyperparameters are required to achieve high accuracy as long as a fast model. Table 4 shows the

model's performance metrics applied to different scenarios. The machine learning models used in this research are LightGBM and K-means with Accuracy and training time as the performance metrics of LightGBM and Silhouette, and Davies Bouldin as the performance metrics of the means as the unsupervised learning methods.

Table 4 *Model validation on different case study.*

Model	Metric	1	2
LightGBM	Accuracy	0.98	0.989
	Training time (s)	0.17	0.193
K-means	Silhouette	0.481616	0.598307
	Davies-Bouldin	0.825007	0.599234

Overall, the generalization process was successfully conducted, as indicated by the minimum difference between the performance metrics for each scenario. LightGBM accuracy shows a very high score above 0.98 across all the scenarios, with case 2 being the highest one. The training time is considerably rapid, with the maximum training time of 0.193 seconds, occurring in scenario 2. The K-means performance metrics, show how the model can recognize the cluster of data. Basically, the output of the K-means model will determine the range composition of the solvent. It shows that the silhouette score doesn't have any difference between scenarios, with the difference not until the 0.5 score. The Davies Bouldin score has the same pattern as the Silhoutte score. From this discussion, it can be concluded that the generalization process for this model was successfully conducted.

Conclusion

This research presents the technique to define the solvent component and composition for absorption-based AGRU using six different solvent blends. In this paper, a general model of component and composition AGRU has been developed. The proposed model has a consistent accuracy of 0.98 across all scenarios, while the clustering performance shows adequate performance with the range of 0.48-0.59. This model successfully determines the solvent component and composition across six different solvent blends on two different feed gas compositions with a total of 37,786 data points.

Acknowledgement

The authors gratefully acknowledge the support provided by Universiti Teknologi Petronas for this research, funded through grant Yayasan Universiti Teknologi PETRONAS (YUTP), with grant number YUTP-PRG 015PBC-037 and YUTP-FRG 015LCO-362.

References

[1] R. Faiz, M. Al-Marzouqi, Insights on natural gas purification: Simultaneous absorption of CO_2 and H_2S using membrane contactors, Sep. Purif. Technol. 76 (2011) 351–361. https://doi.org/10.1016/j.seppur.2010.11.005

[2] S. Sheikh, Enhancements in modeling gas sweetening, in: Proc. Abu Dhabi Int. Petroleum Exhibition and Conference (ADIPEC), Society of Petroleum Engineers, 2015. https://doi.org/10.2118/177656-MS

[4] A.M. Tamidi, N.H. Yasin, CO_2 and H_2S co-removal from natural gas using membrane contactor technology, in: Proc. SPE Asia Pacific Oil and Gas Conference and Exhibition (APOG), Society of Petroleum Engineers, 2020.

[5] L.A. Pellegrini, S. Moioli, F.M. Munari, P. Vergani, B. Picutti, A. Uccelletti, The acid gas removal unit at Gasco's Habshan 5: Simulation and comparison with field data, in: Proc. Offshore Mediterranean Conference and Exhibition (OMC), 2015.

[6] M. Faqih, M.B. Omar, R.J. Wishnuwardana, N.I. Ismail, M.H.M. Zaid, K. Bingi, Prediction of solvent composition for absorption-based acid gas removal unit on gas sweetening process, Molecules 29 (19) (2024) 4591. https://doi.org/10.3390/molecules29194591

[7] W. Zhu, H. Ye, X. Zou, Y. Yang, H. Dong, Analysis and optimization for chemical absorption of H_2S/CO_2 system: Applied in a multiple gas feeds sweetening process, Sep. Purif. Technol. 276 (2021) 119301. https://doi.org/10.1016/j.seppur.2021.119301

[8] A. Esmaeili, T. Yoon, T.A. Atsbha, C.J. Lee, Rate-based modeling and energy optimization of acid gas removal from natural gas stream using various amine solutions, Process Saf. Environ. Prot. 177 (2023) 643–663. https://doi.org/10.1016/j.psep.2023.07.030

[9] M. Hakimi, M.B. Omar, R. Ibrahim, Application of a neural network for predicting H_2S from an acid gas removal unit (AGRU) with different compositions of solvents, Sensors 23 (2023) 1020. https://doi.org/10.3390/s23021020

[10] I. Izonin, R. Muzyka, R. Tkachenko, I. Dronyuk, K. Yemets, S.-A. Mitoulis, A method for reducing training time of ML-based cascade scheme for large-volume data analysis, Sensors 24 (15) (2024) 4762. https://doi.org/10.3390/s24154762

[11] P. Anandan, A. Manju, M.R. Reddy, Classification of massive data sets using a revolutionary grey wolf optimization algorithm and a deep learning model in a cloud-based setting, in: Proc. 2023 Int. Conf. Data Sci., Agents & Artif. Intell. (ICDSAAI), 2023, pp. 1–6.

[12] L.C. Law, N.Y. Azudin, S.R. Syamsul, Optimization and economic analysis of amine-based acid gas capture unit using monoethanolamine/methyl diethanolamine, Clean Technol. Environ. Policy 20 (2018) 451–461. https://doi.org/10.1007/s10098-017-1430-1

[13] B.K. Choi, S.M. Kim, K.M. Kim, U. Lee, J.H. Choi, J.S. Lee, I.H. Baek, S.C. Nam, J.H. Moon, Amine blending optimization for maximizing CO_2 absorption capacity in a diisopropanolamine–methyldiethanolamine–H_2O system using the electrolyte UNIQUAC model, Chem. Eng. J. 419 (2021) 129517. https://doi.org/10.1016/j.cej.2021.129517

[14] A. Sharifi, E.O. Amiri, Effect of the tower type on the gas sweetening process, Oil Gas Sci. Technol. 72 (2017). https://doi.org/10.2516/ogst/2017018

[15] S. Zhang, The study of model generalization ability for spam classification based on machine learning models, in: Proc. 2024 9th Int. Symp. Comput. Inf. Process. Technol. (ISCIPT), 2024, pp. 43–46.

[16] H.A. Behrooz, Robust synthesis of amine-based natural gas sweetening plants, Gas Sci. Eng. 113 (2023) 204970.

[17] Y. Li, Y. Cao, J. Yang, M. Wu, A. Yang, J. Li, Optuna-DFNN: An Optuna framework driven deep fuzzy neural network for predicting sintering performance in big data, Alexandria Eng. J. 97 (2024) 100–113. https://doi.org/10.1016/j.aej.2024.04.026

[18] L. Zhou, D. Garg, Y. Qiu, S.M. Kim, I. Mudawar, C.R. Kharangate, Machine learning algorithms to predict flow condensation heat transfer coefficient in mini/micro-channel utilizing universal data, Int. J. Heat Mass Transf. 162 (2020) 120351. https://doi.org/10.1016/j.ijheatmasstransfer.2020.120351

[19] S. Banerjee, K. Bhavna, T. Raychoudhury, Prediction of transport behavior of nanoparticles using machine learning algorithm: Physical significance of important features, J. Contam. Hydrol. 258 (2023) 104237. https://doi.org/10.1016/j.jconhyd.2023.104237

[20] O. Alshboul, G. Almasabha, A. Shehadeh, K. Al-Shboul, A comparative study of LightGBM, XGBoost, and GEP models in shear strength management of SFRC-SBWS, Structures 61 (2024) 106009. https://doi.org/10.1016/j.istruc.2024.106009

[21] R. Bakır, C. Orak, A. Yüksel, Optimizing hydrogen evolution prediction: A unified approach using random forests, LightGBM, and Bagging Regressor ensemble model, Int. J. Hydrogen Energy 67 (2024) 101–110. https://doi.org/10.1016/j.ijhydene.2024.04.173

[22] K. Yuan, G. Chi, Y. Zhou, H. Yin, A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description, Res. Int. Bus. Finance 59 (2022) 101536. https://doi.org/10.1016/j.ribaf.2021.101536

[23] G. Niu, Y. Ji, Z. Zhang, W. Wang, J. Chen, P. Yu, Clustering analysis of typical scenarios of island power supply system by using cohesive hierarchical clustering based K-means clustering method, Energy Rep. 7 (2021) 250–256. https://doi.org/10.1016/j.egyr.2021.08.049