# A Study on TVAE-based Data Augmentation and Verification to Predict Physiologically Active Ingredients of Medicine Plants According to Climate Change

Hyunjo Lee
*Department of General Education*
*Korea National University of Agriculture and Fisheries*
Jeonju, South Korea
o2near@gmail.com

Cheol-Joo Chae
*Department of General Education*
*Korea National University of Agriculture and Fisheries*
Jeonju, South Korea
chae.cheoljoo@gmail.com

***Abstract*—Climate change, including rising temperatures, droughts, and floods, has recently become a global concern. In the agricultural sector, it is anticipated that climate change will significantly affect the characteristics and productivity of crops. In particular, medicinal plants are used as raw materials in various industrial areas such as health functional foods, natural medicines, and living materials. However, their productivity is decreasing due to climate change. In this paper, we propose a model that can predict the physiological active ingredient index of *Cnidium Officinale Makino*, a representative medicinal crop vulnerable to climate change, based on different climate change scenarios. First, to address the issue of imbalanced data collection, we augmented the collected data using the TVAE algorithm, a structured data generation model. The quality of the augmented data was assessed using column shape and column pair trends, resulting in an average overall quality of 89%. Secondly, in order to predict the total contents of phenol and flavonoid, which are the main physiological active ingredients of *Cnidium Officinale*, the accuracy of the predictions was evaluated using five models: RF, SVR, XGBoost, AdaBoost, and LightBGM. After evaluating model performance, the XGBoost model demonstrated the highest accuracy in predicting the physiological active ingredients of *C. officinale*.**

***Keywords— Machine Learning; Data Augmentation, TVAE, Prediction Model; XGBOOST; Climate Change; Cnidium Officinale Makino***

## I. Introduction

Smart farm means creating new value in various agricultural fields, encompassing not only agricultural production but also distribution and consumption, through the convergence of agriculture and ICT [1, 2]. The primary goal of a smart farm is to enhance productivity by analyzing data collected from the smart farm and delivering the results. Various environmental factors, such as temperature, humidity, light intensity, and moisture, are known to be important in determining the growth conditions of crops. Climate change is significantly altering these environmental factors, leading to missed crop cultivation periods and creating uncertainty in cultivation areas, which has a significant impact on crop production [3]. To accurately understand the effects of climate change and reduce damage caused by it, various studies based on machine learning techniques are being conducted to determine key factors affecting crop production and to explore ways to improve productivity.

However, research on high-value medicinal plants has not been actively pursued, except for a few plants like ginseng or mushrooms. It is challenging to gather a substantial amount of data for machine learning due to limitations on the number of production farms and cultivation methods. To address the issue, we collects environmental and physiological response data for *Cnidium Officinale Makino*, a representative medicinal plant vulnerable to climate change. Additionally, a data augmentation technique is introduced to alleviate the imbalance in the collected data. Finally, using the augmented data, we make predictions for the physiological active ingredients index of *Cnidium Officinale* based on climate change scenarios.

The content of the paper is as follows. In Session 2, we explain the factors contributing to climate change and the experimental environment. In Session 3, we describe data augmentation technique for processing the collected data and using prediction models. We assess the performance of the prediction model for the physiological active ingredients of *C. Officinale* in Session 4 and conclude the paper in Session 5.

## II. Climate change scenarios and cultivation environment for *Cnidium Officinale*

Climate change scenarios include SRES(Special Report on Emission Scenarios), RCP(Representative Concentration Pathways), and SSP(Shared Socioeconomic Pathways). In this paper, we utilize the SSP scenario, specifically SSP1-2.6, SSP3-7.0, and SSP5-8.5 [4].

To manage the environmental impact of climate change, the environmental conditions, such as CO2 concentration, temperature, and humidity, inside the chamber are precisely controlled. The SPDS (Soil Plant Daylit System) chambers, made of plexiglass to allow natural light, were used, as shown in Fig. 1.

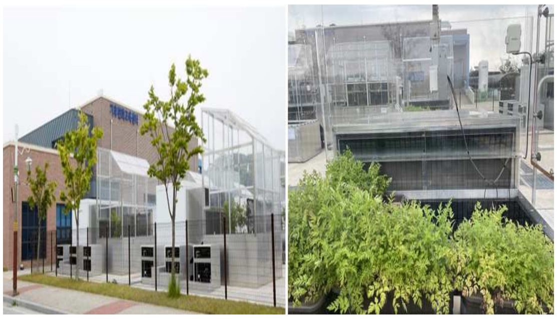

Fig. 1. SPDS for *Cnidium Officinale*

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(NRF-2021R1F1A1063042).

## III. Data Collection and Augmentation

### A. Data Collection

We collected environmental and physiological response data for *Cnidium Officinale* under different climate change scenarios (SSP1-2.6, SSP3-7.0, and SSP5-8.5) using the SPDS chamber at the Climate Change Education Center of Korea National University of Agriculture and Fisheries from May to September 2023. The data collection environment is depicted in Table 1.

Data collected for climate change scenarios includes environmental information, physiological response index, and physiological active ingredients index. First, environmental data (atmospheric CO2 concentration, temperature, relative humidity, atmospheric vapor pressure difference (VPD), and light intensity (PPDF)) was collected every hour. Second, the physiological response index of *C. Officinale* includes photosynthetic pigments such as chlorophyll a, chlorophyll b, total chlorophyll, and carotenoid content, as well as ratios like chlorophyll a/b and chlorophyll/carotenoid. It also encompasses energy transfer flow like photosynthetic reaction center deactivation index, heat release index, energy capture index, and energy delivery index. Additionally, it involves photosynthetic vitality measures such as energy conservation index, photosynthetic driving force, photosynthetic functional structure, and quantum yield. We collect the physiological response index once a month and measure it five times repeatedly. The physiological active ingredients index of *C. Officinale* finally includes the total extraction yield of phenol and flavonoid from the leaf and root (%), the total phenol content from each part of the leaf and root (mg GAE/g weight of C. officinale), and the total flavonoid content from each part of the leaf and root (mg QE/g weight of C. officinale). We collect the index of physiologically active ingredients once a month and measure it three times repeatedly.

The total number of data collected for each scenario includes 3,237 pieces of environmental information, 75 pieces of physiological response index data, and 45 pieces of physiological active ingredients index data.

TABLE I. Data collection environment

| | CO2 concentration | Temperature | Humidity |
|---|---|---|---|
| **SSP1-2.6** | 445 ppm | +1.8℃ | 60% |
| **SSP3-7.0** | 872 ppm | +3.6℃ | 60% |
| **SSP5-8.5** | 1,142 ppm | +4.4℃ | 60% |

### B. Data Augmentation

TVAE (Tabular Variational Autoencoder) is a model that applies VAE (Variational Autoencoder) to structured data. TVAE applies mode-specific normalization to address non-Gaussian and multimodal distribution problems and utilizes data generation through a conditional generator to enhance the quality of augmented data [5-7]. In this paper, the collected data is enhanced using a TVAE-based oversampling technique, which can address data imbalance between environmental information, physiological response index, and physiological active ingredient index.

We assessed the quality of augmented data based on column shape, column pair trends, and overall quality. Column shape represents the similarity in distribution between the original data column and the corresponding augmented data column, and was measured using the Kolmogorov-Smirnov (K-S) test. To calculate the K-S statistic, the numerical distribution was transformed into a cumulative distribution function (CDF), and the maximum difference between the two CDFs was measured [8-9]. Column pair trends represents the correlation coefficient between real data (R) and augmented data (S) for a pair of columns A and B, and was calculated using (1).

$$\text{score} = 1 - \frac{|S_{A,B} - R_{A,B}|}{2} \times 100\% \qquad (1)$$

Overall quality is calculated by averaging the values of the column shape and column pair trends. Table 2 presents the quality measurement results of the TVAE-based augmented data.

TABLE II. Quallity of the augmented data

| | Column shape | Column pair trends | Overall quality |
|---|---|---|---|
| **SSP1-2.6** | 85.22% | 94.37% | **89.79%** |
| **SSP3-7.0** | 85.55% | 93.34% | **89.45%** |
| **SSP5-8.5** | 86.87% | 93.66% | **90.27%** |

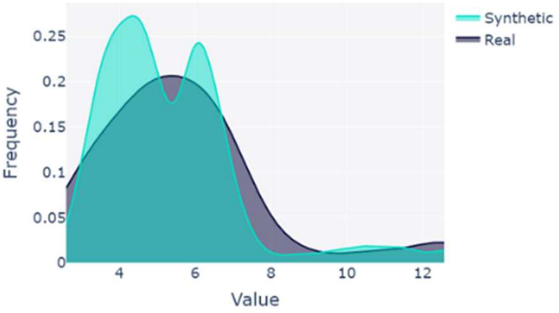

Fig. 2. Column shape of total chlorophyll (TChl) at SSP3-8.5

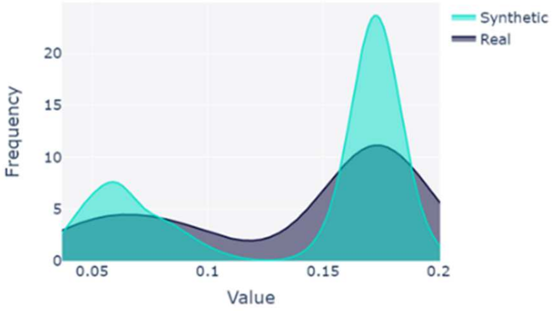

Fig. 3. Column shape of SFIabs at SSP3-8.5

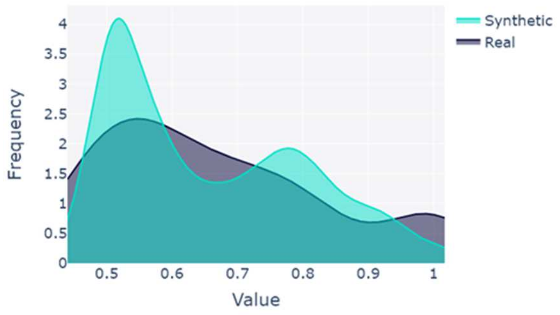

Fig. 4. Column shape of Eto/RC at SSP3-8.5

## IV. Performance Evaluation of Prediction Models for Verifying Augmented Data

To validate the augmented data, we make experiments to predict the contents of physiological active ingredients of *Cnidium Officinale* according to climate change. First, we augment both physiological response index data and physiologically active ingredients data, and then map them with environmental information data. 80% of the preprocessed data was used for training, 20% for testing, and k-fold cross-validation (k=5) was performed. To predict the content of physiological active ingredients in *C. officinale*, we used five methods: RF (Random Forest) [10], SVR (Support Vector Regression) [11], XGBoost [12], AdaBoost [13], and LightGBM [14]. For the five models, we measure Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Root Mean Squared Log Error (RMSLE), which are calculated by (2), (3), and (4), respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \qquad (2)$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y_i})^2}{n}} \qquad (3)$$

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y_i} + 1))^2} \qquad (4)$$

For each prediction model, we determine the optimal hyperparameters using the grid-based method [15-18]. The Random Forest (RF) parameters were set as follows: n_estimators=2000, criterion='squared_error', max_features=12, max_depth=10, min_samples_split=8, and min_samples_leaf=8. The SVR parameters were set to kernel='rbf', C=64, and gamma=8. The XGBoost parameters were configured as follows: learning_rate=0.1, n_estimators=890, max_depth=13, min_child_weight=5, gamma=0, subsample=0.9, colsample_bytree=0.8, objective='reg:squarederror', reg_alpha=10, and reg_lambda=0.1. The AdaBoost parameters were set to n_estimators=50, learning_rate=0.1, and loss='linear'. The parameters of LightBGM were set as follows: n_estimators=1000, learning_rate=0.05, max_depth=-1, num_leaves=90, and colsample_bytree=0.5.

To assess all the models, we predict the total phenol contents in leaf and root, and the total flavonoid contents in leaf and root, based on the climate change scenarios SSP1-2.6, SSP3-7.0, and SSP5-8.5. Next, we calculate the average prediction error for each model. Fig. 5. shows the result of the average prediction error.

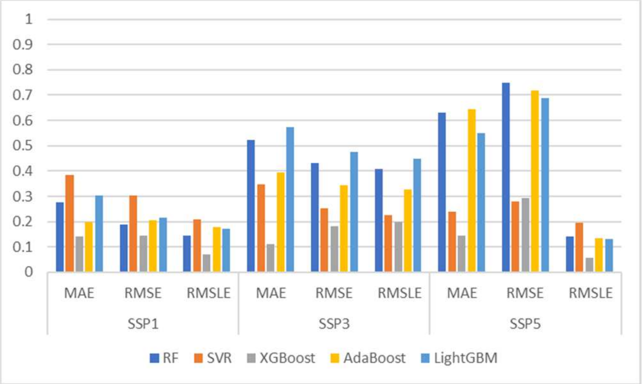

Fig. 5. The average prediction error for five prediction models

XGBoost exhibits excellent performance in predicting the physiological active ingredients of *C. Officinale*. XGBoost is a boosting-based ensemble machine learning model that applies regularization to prevent overfitting. Due to regularization, XGBoost can achieve high prediction accuracy.

## V. Conclusion

In this paper, we proposed a model that can predict the physiological active ingredient index of *Cnidium Officinale*, a representative medicinal crop vulnerable to climate change, based on different climate change scenarios. To predict the phenol and flavonoid components, which are physiological active components of C. officinale, environmental information was collected using an SPDS chamber. Physiological response data was collected monthly and measured five times, while the physiological active ingredients were collected monthly and measured three times. To address the issue of data imbalance in collecting environmental information, physiological reactions, and physiologically active ingredient data, we augmented the data using the TVAE algorithm. The overall quality of the augmented data is 89%, achieved by averaging column shape and column pair trends. With the augmented data and five prediction models (RF, SVR, XGBoost, AdaBoost, and LightGBM), we predicted the total phenol and total flavonoid contents in leaf and root. After evaluating prediction errors, XGBoost demonstrated the best performance in predicting the physiological components of C. officinale.

In the future, we plan to design and implement a prediction model that integrates various existing models. So we can derive key factors and correlations between features by expanding the collected data.

## References

[1] Y. Akkem, S. K. Biswas, A. Varanasi, "Smart farming using artificial intelligence: A review," Engineering Applications of Artificial Intelligence, 120, 105899, April 2023.

[2] I. Attri, L. K. Awasthi, T. P. Sharma, P. Rathee, "A review of deep learning techniques used in agriculture," Ecological Informatics, 102217, 2023.

[3] J. Kim, J. Han, "Agricultural management innovation through the adoption of internet of things: Case of smart farm," Journal of Digital Convergence, vol. 15, no 3, pp. 65-75, March, 2017.

[4] Climate Change Scenario, http://www.climate.go.kr/home/Eng/htmls/intro/sub3.html, (2023.10.30.)

[5] A. Kiran, S. S. Kumar, "A Comparative Analysis of GAN and VAE based Synthetic Data Generators for High Dimensional, Imbalanced

Tabular data," In 2023 2nd International Conference for Innovation in Technology (INOCON) , pp. 1-6, IEEE, March, 2023.

[6] O. Habibi, M. Chemmakha, M. Lazaar, "Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection," Engineering Applications of Artificial Intelligence, 118, 105669, February, 2023.

[7] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, "Modeling tabular data using conditional gan," Advances in neural information processing systems, 32, 2019.

[8] Kolmogorov–Smirnov test(2023), https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test (2023.10.30.)

[9] Cumulative distribution function(2023), https://en.wikipedia.org/wiki/Cumulative_distribution_function (2023.10.30.)

[10] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," Ore Geology Reviews vol. 71, pp. 804-818, 2015.

[11] F. Zhang, L. J. O'Donnell, "Support vector regression," Machine learning, Academic Press, pp. 123-140, 2020.

[12] T. Chen, G. Carlos, "Xgboost: A scalable tree boosting system," Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785-794, 2016.

[13] R. E. Schapire, "Explaining adaboost," Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37-52, 2013.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, 2017.

[15] W. Wang, Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," IOP conference series: materials science and engineering, IOP Publishing, vol. 324. 2018.

[16] D. Chicco, M. J. Warrens, G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, 2021.

[17] S. Basheer, R. M. Mathew, M. S. Devi, "Ensembling coalesce of logistic regression classifier for heart disease prediction using machine learning," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 12, pp. 127-133, 2019

[18] A. A. Mir, K. J. Kearfott, F. V. Çelebi, M. Rafique, "Imputation by feature importance (IBFI): A methodology to envelop machine learning method for imputing missing patterns in time series data," PloS one, vol. 17, no. 1, e0262131, January, 2022.

[19] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[20] S. Chatterjee, "A new coefficient of correlation," Journal of the American Statistical Association 116.536, pp. 2009-2022, 2021.