Harvest Forecasting Improvement Using Federated Learning and Ensemble Model

Ohnmar Khin, Jin Gwang Koh, Sung Keun Lee

Abstract

Harvest forecasting is the great demand of multiple aspects like temperature, rain, environment, and their relations. The existing study investigates the climate conditions and aids the cultivators to know the harvest yields before planting in farms. The proposed study uses federated learning. In addition, the additional widespread techniques such as bagging classifier, extra tees classifier, linear discriminant analysis classifier, quadratic discriminant analysis classifier, stochastic gradient boosting classifier, blending models, random forest regressor, and AdaBoost are utilized together. These presented nine algorithms achieved exemplary satisfactory accuracies. The powerful contributions of proposed algorithms can create exact harvest forecasting. Ultimately, we intend to compare our study with the earlier research's results.

Keywords: Harvest forecasting | Federated learning | Ensemble algorithms | Multiple aspects

I. INTRODUCTION

Agriculture is the major source of the economy where about 65 percent of the people rely on agriculture. A huge number of crops are widely grown. According to survey results of the agriculture sector, many farmers face many problems in planting crops. Nowadays, agriculture has become much more important than in the past because plants are the main source of life to improve human living standards. Crops were mainly accustomed to feeding humans and animals. Harvest forecasting is crucial nowadays. Forecasting is required for successful agriculture. Agricultural precision is a trendy farming technique that utilizes the soil, weather conditions, and crop yield data. In the earliest days, monitoring was done manually by humans in that field. However, it requires research contribution and processing time. To predict crop production with the best quality and quantity, farmers are suggested to use the modernized digital image processing technique. By using this, harvest can be predicted and classified effectively. In the agriculture sector, paddy is widely grown, and people eat rice as a staple food. Most farmers grow various crops that lack knowledge of prediction in a very ineffective way.

The proposed research work provides a more accurate and robust crop prediction

* This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01489) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

* Sunchon National University

Manuscript: 2023.10.20
Confirmation of Publication: 2023.11.22

Corresponding Author: Sung Keun Lee, e-mail: sklee@scnu.ac.kr

system for agriculture simply. The objective of the system is:

(a) To investigate the harvest which is most grown.

(b) To classify the crops using federated learning and ensemble algorithms.

(c) To build a maximum yield and profit for farmers.

(d) To create a complete crop prediction system to get good model accuracy.

The previous vast number of papers are demonstrated in literature. In the last two years, crop yield prediction performance collation using supervised machine learning algorithms was researched by S. Khaki and L. Wang [1]. The authors compared the implementation of decision trees, artificial neural networks, and k-nearest neighbor algorithms to select the best one. The usefulness of the presented system provides a simple solution with effective outcomes. It helps India become a food-secure country. N BanuPriya, D Tejasvi, and P Vaishnavi proposed ensemble methods for crop yield prediction in [2]. Random Forest Regressor (RFR) was significant when compared to others. However, it is necessary to develop an application that farmers can utilize in their native language. The paper used deep neural networks [3]. The proposed model is extremely tested with other favored approaches such as Lasso, Shallow Neural Networks (SNN), and Regression Tree (RT). The outcomes also displayed that those environmental aspects had a more significant impact on crop yield than genotype. The palm oil yield prediction with machine learning has been presented [4]. This technology performed unique challenges in the study and powerful model outcomes for the forecasting of palm oil yields with the considerably smallest computational problems. The authors were extremely challenged with factors such as environmental facts, management practices, crop genotype, etc. [5]. This study consisted of recurrent neural networks (RNNs), convolutional neural networks (CNNs), deep fully connected neural networks (DFNN), random forest (RF), and LASSO. The crop yield prediction accuracy is different in soil conditions, weather predictions, and management practices that were explained. This presented a new predictive algorithm for crop yield prediction [6]. First, it achieved an RMSE of 8% or less. Second, it recognized a dozen surroundings by managing interchanges for corn and soybean yield. Third, it quantitatively examined crop yield into grants from the soil, weather, management, etc. The powerful grant of the new prediction algorithm is its capacity. It performed a Systematic Literature Review to extract and synthesize the classifiers and then features are utilized in crop yield prediction analyses [7][13]. Due to this study, the most features are temperature, rainfall, and soil type. The most used classifier is Artificial Neural Networks (ANN) among these algorithms. Due to this further study, Convolutional Neural Networks (CNN) are the most well-utilized algorithm in these studies, and the other spread-operated

algorithms are Deep Neural Networks (DNN) and Long-Short Term Memory (LSTM)[14]. Bomn and Zidek reported how spatial dependencies are comprised of statistical standards for the crop [8]. The prototype is focused on delivering efficient forecasts which stabilize the outcomes of noisy data. Prior publications are developed to adapt the spatial non-stationarity occurring from distinct between-region differences in farming policy and procedure. In addition, the scope of potential dimension-reduction strategies is tested for enhanced prediction. The study explained that precise predictions with well-timed are extremely essential [9]. Various types of crops are used. Climatic changes mainly impact crop yield predictions. Factors such as climatic situations, genotype, and ground types are used in predicting the yield. Based on the various previous paper studies, the prediction experiments are run by utilizing the different deep machine learning algorithms and ANN models that are executed to predict the yield, and the outcomes are investigated. Today's Indian economy relies on agribusiness [10]. In India, more additional 70% of individuals have taken it as a major career for a certain crop. Crops gain due to environmental situations such as climate, ground quality, heavy rain, deficiency, seed cracks, etc. The farmers do not get high products. Therefore, farmers forecast the crop yield using machine learning methods like the KNN model and Linear Regression using recorded farming facts. A comparative study is complete with decision tree operations to achieve heightened precision, and model implementation is computed. CSM offers very satisfactory results compared to the previous prediction models [11][12]. This research represents the abilities of different classifiers in predicting climate trends such as rainfall, temperature, humidity, soil, etc., and concludes the main procedures like preprocessing, feature extraction, classification models, and prediction are appropriate to predict weather sensations.

II. DATA AND FEATURES

The current data was downloaded from the Kaggle India dataset. It was used as a study region and by using this dataset, built a prediction model for crops on the different parameters. It is the numerous appropriate crops to cultivate on a specific farm using the various parameters. The columns in the dataset have Nitrogen, phosphorus, Potassium, pH values of the soil, humidity, temperature, and rainfall.

The dataset is composed of eight attributes and various types of crops (labels). The dataset consists of 2200 data. Agricultural products are extremely dependent on climate factors such as rain, soil, temperature, etc. This dataset contains rainfall, weather, and soil data. The eight attributes are described as N (Nitrogen), P (phosphorus), K (Potassium), temperature, humidity, pH, rainfall, and label (various types of crops). There are 22 types of crops. The histogram graph of the seven attribute crops is shown in Figure 1.

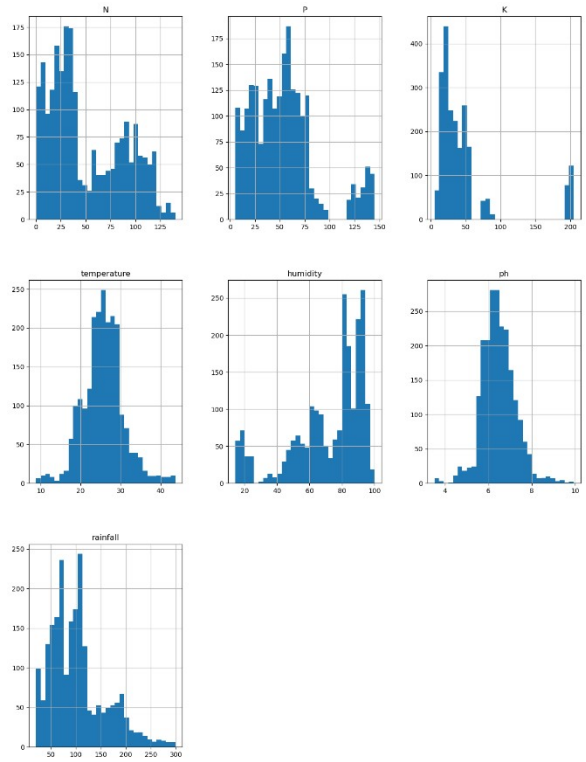

Fig. 1. Histogram graph for seven attributes

The following step is to conduct the feature extraction. In machine learning, feature extraction is the technical sense. It improves the machine learning model performance. Feature extraction chooses the appropriate information. It is nice to compute and develop features, but some points are needed to ban unrelated features. Since we don't know the ideal number of features, experiment with the approaches with potential features and select the numeral features with a more acceptable implementation. All extracted feature values are shown in Figure 2. This means that Nitrogen, phosphorus, Potassium, temperature, humidity, pH, and rainfall features are extracted.

	N	P	K	temperature	humidity	ph	rainfall
0	90	42	43	20.879744	82.002744	6.502985	202.935536
1	85	58	41	21.770462	80.319644	7.038096	226.655537
2	60	55	44	23.004459	82.320763	7.840207	263.964248
3	74	35	40	26.491096	80.158363	6.980401	242.864034
4	78	42	42	20.130175	81.604873	7.628473	262.717340
...	...	...	...	...	...	...	...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507
2196	99	15	27	27.417112	56.636362	6.086922	127.924610
2197	118	33	30	24.131797	67.225123	6.362608	173.322839
2198	117	32	34	26.272418	52.127394	6.758793	127.175293
2199	104	18	30	23.603016	60.396475	6.779833	140.937041

Fig. 2. Extracted features.

III. MATERIALS AND METHODS

In this part, the flow of the system architecture is depicted. The overall flow diagram is shown in Figure 3. The diagram inputs the crop yield dataset for input data and arranges the dataset, after that processes the preprocessing stage. This step removed the unwanted data or unnecessary noise and checked the missing value in the dataset. If the missing value has occurred, must replace it. Standard scalar means to remove the mean and scales of each feature that are variables to unique variants. The missing data handling increases the machine learning model performance. Feature scaling is required before creating training data and testing data.

The next step is to implement data splitting. The input data is split into training and testing data. Most data are stored in training and a small amount is stored in testing. Test data is used to evaluate the crop yield predicting model.
To build a model, the training set is applied. It consists of the data set which

is used for training the system. Training rules and applied algorithms provide the appropriate information to associate with input–output decisions. The program is conducted by utilizing deep learning and ensemble algorithms. Then, all the appropriate information is executed from the data. Later, the experimental outcomes are acquired. Normally, for training data, 80% of the data is done.

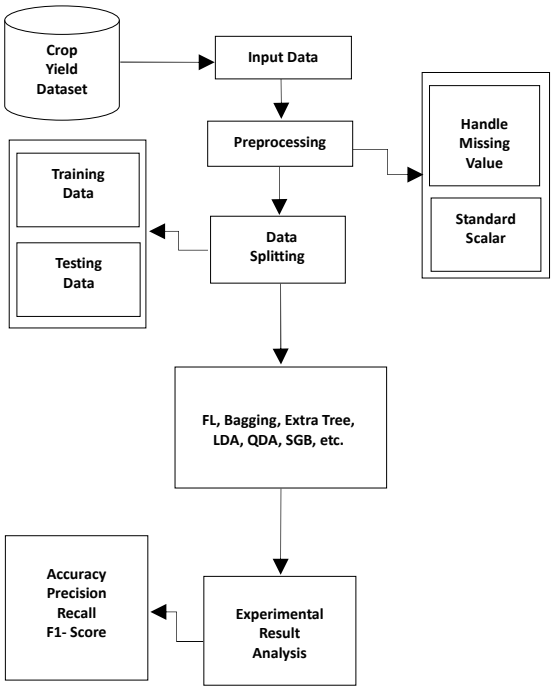

Fig. 3. Overall flow diagram

Testing data is used to verify the system that is producing an accurate outcome after being prepared. Typically, for testing, 20% of the dataset is utilized. Test data is applied to calculate the accuracy of the method. Subsequently, classification or model selection is implemented. Machine learning algorithms such as FL, bagging, extra trees, LDA, QDL, SGB, Blending, random forest regressor, and Ada boost are implemented. Finally, the experimental results are shown with accuracy, precision, and recall.

A. Yield Prediction Using Deep Learning

Deep learning (DL) is the machine learning technique that directly executes the features from the input dataset. When the data is boosted, machine learning (ML) techniques are inadequate in terms of performance. Deep learning provides a more satisfactory implementation such as accuracy. In this research, federated learning has been experimented with. The FL gains better accuracy. The experimental result is expressed below.

B. Federated Learning (FL)

Federated learning relates to AI. It trains the model itself, not a centralized one. The data is used for a particular device. Federated learning creates safe AI. Its early stage has challenges with design and deployment.

Users utilize the centralized copy to their needs. Federated learning trains the input data and tests. It smarts from time to time. Transfer the train results from the copy to the server. The results are saved on the server. The model result is finer than the earlier performance. In addition, the computation time is faster than twice of normal execution.

Modernize the version and model and create the user's data. Federated learning working processes are represented in Table 1. By executing the above processes, federated learning evaluates correctness as shown in Figure 4.

Table I. Federated learning working processes

1	First, load the data and convert the trained dataset into a federated dataset.
2	Execute the architecture of deep networks and fix the devices for the train.
3	Initialize the model. send () for the training function. Model updates and loss are examined utilizing the model. get () method.
4	The working out is done.
5	The computation time is less than twice the time used for normal execution. Communication time over the network is 0.07 seconds.

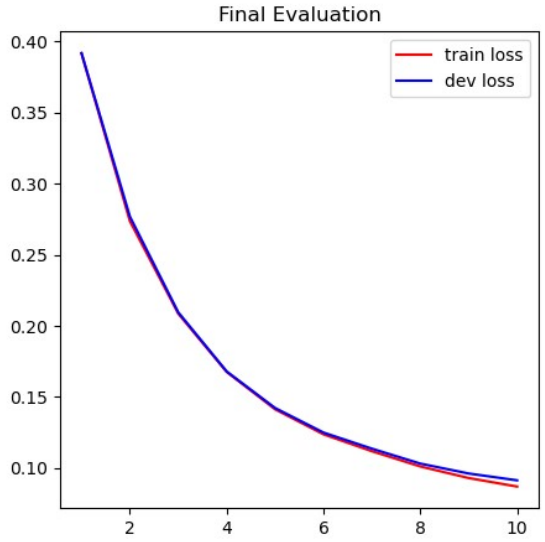

Fig. 4. Federated learning's evaluation

C. Yield Prediction Using Ensemble Algorithms

Ensemble Algorithms of machine learning are crucial determination aid tools for crop yield prediction. The machine learning process starts with the data to create more acceptable decisions. Split the input data set into two parts train and test sets. The sci-kit-learn library is used for all functions and machine-learning algorithms in this notebook. The most famous methods for merging the predictions from various models are the bagging classifier, extra tees classifier, LDA classifier, QDA classifier, SGB classifier, blending models, random forest regressor, and Adaboost. Boosting ensembles consist of Ada boost and stochastic gradient boosting. Bagging makes numerous models from separate subsamples of the training dataset. Boosting makes multiple models and individual learning to set the forecast mistakes of a previous model in the chain. These ensemble machine-learning algorithms improve the models' execution of the problems.

D. Bagging Model

It performs the best algorithm that has high variance. Bagging takes numerous samples from the training dataset and trains a model for an individual sample. The final prediction output is averaged from all the predictions of sub-models. The below model uses a bagging classifier with the classification. Construct a bagging model for classification working for the bagging classifier class. The bagging model accuracy is 98%.

E. Extra Trees

Extra trees are unexpected trees that are built from the training dataset samples. Create the extra tree classification using this extra tree classifier. In this extra tree analysis, effectiveness is associated with its technique categories of accuracy, precision, recall, and f1-score of the model through the optimization of the proposed extra tree model.

F. Making predictions with Linear Discriminant Analysis (LDA)

LDA is an easy, well-understood, and useful technique for classification. LDA creates predictions by calculating the possibility in which a renewed set of

inputs belongs to an individual class. The initial development is called the LDA. The cross-validation scores of LDA are 96%.

G. Quadratic Discriminant Analysis (QDA)

QDA is a popular extension. Each class utilizes its estimation of variance. Between them, QDA gets better accuracy than LDA. The cross-validation scores of QDA are 0.98181818, 0.98863636, 0.98863636, 0.98181818 and 0.98409091 respectively. The confusion matrix of QDA is constructed.

H. Stochastic Gradient Boosting (SGB)

Boosting the ensemble algorithm makes the series that tries to correct the model's errors before the series. The model creates predictions that are weighted by displayed accuracy and the outcomes are merged for the final output prediction. Two boosting algorithms are AdaBoost Gradient Boosting and Stochastic Gradient Boosting. Stochastic is a considerably complex ensemble strategy. It is perhaps the best approach for enhancing implementation via ensembles. Build a gradient boosting model for classification utilizing the stochastic gradient boosting classifier.

I. Random Forest Regressor

It is a meta-estimation that works on different samples of one dataset and uses average forecasting accuracies. Lastly, manages the over-fit. The random forest regressor's result is 87%. The sample size is handled with the max_samples if bootstrap=True.

J. AdaBoost Classifier

AdaBoost classifier fits the actual data and then works as an extra classifier on the same data. However, the importance of wrongly classified samples is modified such that the next classifier concentrates better on complex circumstances. In this experimentation, AdaBoost classifier accuracy is poor. It gains 20% accuracy.

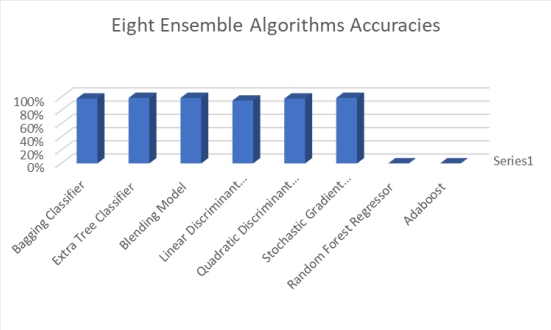

Fig. 5. Eight ensemble algorithms' accuracies

Figure 5 displays the experimental results for the bagging classifier, extra tees classifier, LDA classifier, QDA classifier, SGB classifier, blending models, random forest regressor, and Adaboost. The figure shows the above eight ensemble algorithms' accuracies in the pie chart. Among the eight ensemble algorithms, the bagging classifier, the extra trees classifier, the stochastic gradient boosting classifier, and the blending models get 99% accuracy. These three classifiers are the best. The quadratic discriminant analysis classifier gains 98% accuracy, the linear discriminant analysis classifier obtains 95% accuracy, the random forest regressor achieves 87%, and Adaboost is only 20%.

IV. RESULTS AND DISCUSSION

The Jupyter Notebook is a powerful development. It investigates the crop

recommendation dataset with a plot library. Afterward, experiments with the proposed distinct machine learning algorithms and acquire the most satisfactory accuracy. The proposed methods of federated learning accepted 97%, the bagging classifier got 99%, the extra tree classifier gained 99%, the blending model got 99% accuracy, the linear discriminant analysis (LDA) classifier obtained 95%, quadratic discriminant analysis (QDA) classifier acquired 98%, stochastic gradient boosting (SGB) classifier received 99% respectively, the random forest regressor achieves 87%, and Adaboost is only 20%. Cross value (cv) is important for high prediction accuracy. cv is adjusted for each model to achieve good results. We assigned the cv values to '3', '4', or '5'. Among these classifiers, the blending model is the best. The proposed algorithms' accuracies are indicated in the bar chart in Figure 6.

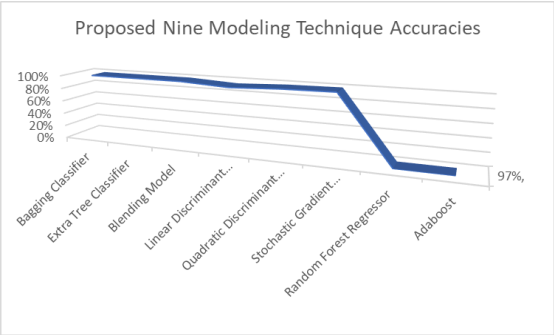

Fig. 6. Proposed algorithms' accuracies comparison

What optimum crops should be grown based on several parameters? It supports farmers with numerous optimal crops to make the right decision about farming for maximum yield and profit. By utilizing this, farmers make knowledgeable decisions before cultivation. The implementation results are described in Table 2. Based on the given dataset, the accuracies differ among these classifiers.

Table 2. Results of training and testing data

Proposed Models	Training Set	Testing Test
Federated Learning (FL)	1.0	0.98
Bagging Classifier	0.99	0.98
Extra Tree Classifier	1.00	0.99
Linear Discriminant Analysis (LDA) Classifier	0.96	0.95
Quadratic Discriminant Analysis (QDA) Classifier	0.99	0.98
Stochastic Gradient Boosting (SGB) Classifier	1.00	0.99
Blending Model	1.00	0.99
Random Forest Regressor	1.00	0.87
Adaboost Classifier	0.23	0.20

The current results are compared with K−NN, Decision Tree, Random Forest, GaussianNB, XGBoost, Logistic Regression, and SVC of the previous research [11]. The comparison is shown in Figure 7. The result indicates that the proposed algorithms offer very satisfactory results compared to the previous prediction models. Achieving the current substantial model accuracy percentages of this research is a challenging task. This research represents the abilities of different classifiers in predicting climate trends such as rainfall, temperature, and humidity.

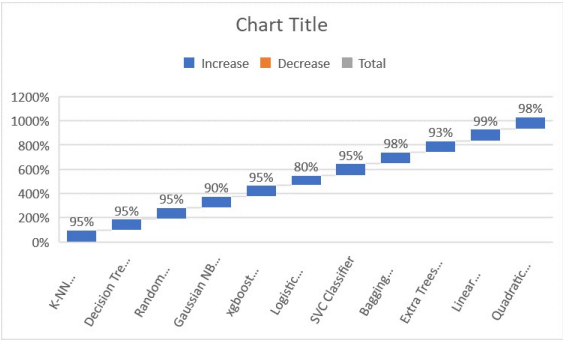

Fig. 7. Comparison results of the previous research

V. CONCLUSION

The current methods reduce crop failures and assist farmers in making informed decisions. The classifiers offer the best results to predict the harvests. Harvest forecasting benefits the farmers by predicting the expenses and resources. Current research supports agriculture. To sum up, it lessens the system's powerless issues and profits food production. Shortly, an Android application with the farmer's native language needs to be developed.

REFERENCES

[1] Saeed, K.; Lizhi, W.; "Crop yield prediction using deep neural networks," *Frontiers in Plant Science,* vol.10, Article 621, pp.1−10, May 2019

[2] N, B, Priya.; D, Tejasvi.; P, Vaishnavi.; "Crop yield prediction based on Indian agriculture using machine learning," *International Journal of Modern Agriculture,* vol.9, pp.1963−1973, Aug. 2020

[3] Doshi, Z., Nadkarni, S., Agrawal, R., Shah, N., "AgroConsultant: intelligent crop recommendation system using machine learning algorithms," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 1−6, Aug. 2018.

[4] Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., Khan, N., "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction," *IEEE Access,* vol. 9, 63406−63439, Apr. 2021.

[5] Saeed, K.; Lizhi, W.; Sotirios, V, A.; "A CNN−RNN framework for crop yield prediction," *Frontiers in Plant Science,* vol.10, Article 1750, pp.1−14, Jan. 2020

[6] Javad, A.; Lizhi, W.; Sotirios, V, A.; "An interaction regression model for crop yield prediction," *Scientific Reports*, Sep. 2021

[7] Luke, B.; James, V, Z.; "Efficient stabilization of crop yield prediction in the Canadian Prairies," *Elsevier*, pp. 223−232, Sep. 2011

[8] Bornn, L., Zidek, J. V., "Efficient stabilization of crop yield prediction in the Canadian Prairies," *Agricultural and Forest Meteorology,* vol. 152, 223−232, Jan. 2012.

[9] K. Pravallika; G. Karuna; K. Anuradha; V. Srilakshmi; "Deep Neural Network model for proficient crop yield prediction," *E3S Web of Conferences 309*, pp.1−10, 2021

[10] Ashwini, I, P.; Ramesh, A, M.; Vinod, D.; "Crop yield prediction using machine learning techniques," *International Journal of Scientific Research in Science, Engineering, and Technology*, vol.7, no. 3, Mar. 2019

[11] Zeel, D.; Subhash N.; Rashi A.; Neepa S.; "Agri Consultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms," *IEEE*, Aug. 2018

[12] HanByeol Oh; JongHyun Lim; SeungWeon Yang; YongYun Cho; ChangSun Shin; "A Study on the Prediction of Strawberry Production in Machine Learning Infrastructure," *Smart Media Journal,* Vol. 11, No. 5, pp. 9−16, Jun. 2022

[13] Jiuqing Dong; Alvaro Fuentes; Sook Yoon; Taehyun Kim; Dong Sun Park; "Towards Improved Performance on Plant Disease Recognition with Symptoms Specific Annotation," *Smart Media Journal,* Vol. 11, No. 4, pp. 38−45, May 2022

[14] Eunji Lee; Hyungwook Park; Eunju Kim; "A Study on LSTM−based water level prediction model and suitability evaluation," *Smart Media Journal*, Vol. 11, No. 5, pp. 56−62, Jun. 2022

Authors

Ohnmar Khin

She received a Bachelor of Computer Science in the Department of Computer Studies from Yadanabon University in 2008 from Yatanabon and a Master of Science degree in the Department of Computer Studies from Yangon University, Myanmar in 2011. Now, she is a Doctoral Candidate in the Department of Multimedia Engineering at Sunchon National University Graduate School in South Korea.
Her research interests are AI-based image processing, crop yield prediction algorithms, deep reinforcement learning, and smart agriculture.

Jin-Gwang Koh

He received the B. S., M.S., and Ph. D. degrees in computer science and engineering from Hongik University, Seoul, Korea, in 1982, 1984, and 1997, respectively. He joined the department of Computer Science and Engineering, Sunchon National University, Suncheon, Korea, in 1988, where he is currently a professor. He is the President of the KISM(Korea Institute of Smart Media) from 2018 to the present. His current research interests are database, wireless sensor networks

Sung-Keun Lee

He received his B.E., M.E., and Ph.D. degrees in Electronic Engineering from Korea University in 1985, 1987 and 1995. From 1996 to 1997 he worked at Samsung electronics network research team. He was a visiting professor at ECE, Georgia Tech, USA from 2017 to 2018. Since 1997, He has been a professor in the Department of Multimedia Engineering at Sunchon National University, Korea.
His research interests are reinforcement learning-based QoS guarantee technology, AI-based solar power prediction systems, and multimedia communication.