International Journal of Smartcare Home
Vol.1, No.1 (2021), pp.35-44
https://dx.doi.org/10.21742/26531941.1.1.04

Machine Learning Models using Paprika Leaf Growth Forecast Based on Environmental and Energy Data

Saravanakumar Venkatesan[1], Jonghyun Lim[2], Changsun Shin[3], and Yongyun Cho[4*]

[1,2,3,4]Department of Information and Communication Engineering, Sunchon National University, South Korea*
[1]skumarvsk1288@gmail.com, [2]sshb56@s.scnu.ac.kr, [3]csshin@scnu.ac.kr
[4]yycho@scnu.ac.kr*

Abstract

Paprika (Capsicum annuum L) is an extremely popular and widespread plant species in South Korea. The purpose of this paper is to develop the prediction method for Paprika growth and compare the leaf count of two areas of paprika (R1 and R2 mean Row planting) through investigation of the Greenhouse with the different environmental factors influencing its growth. The objectives of this paper are the following: (1) to use Machine Learning (ML) approach for crop growth prediction in greenhouse agriculture; (2) to research on the correlativity among different weather factors like input temperature, output temperature, wind speed, dew point, CO2, and humidity are connected to Paprika growth at the field level; and (3) to test growth data using the predictive machine-learning models Support Vector Machine (SVM), generic Random Forest (RF), Gradient Boosting Machines (GBM) and eXtreme Gradient Boosting (XGB). Compared to the principal component regression the machine learning models show the best skills in predicting Paprika growths. The Support Vector Machine method is used to provide the best performance in predicting Paprika growth. While measurements of one production period can predict crop development with sensible requirements, we need more attempts to allow this approach in several fields in the region.

Keywords: *Paprika leaf, Environmental, Correlation, Linear regression and machine learning*

1. Introduction

Paprika (Capsicum annuum L) production observation is necessary for perfecting dispensation and increasing the growth of greenhouse paprika. Leaf growth and leaf width are important facts for remembering growth. To supervise the growth of paprika by accurately getting performance-related attributes (leaf count of growth and leaf width) is a great applicative important thing to improve productivity and quality of paprika [1]. The linear regression model for measure production-related attributes, which are comparatively truthful, can achieve a comparatively accurate solution. Nevertheless, the processes require annihilating selection, thus creating it time-overwhelming [2].

The hypothesis of this method, the sensor-derived growing data can measure the production-related attributes, thus succeeding in non-destructive growth monitoring. The environmental factors using binary classification machine learning models (SVM, RF, GBM, XGB) and leaf

Article history:
Received (March 14, 2021), Review Result (April 10, 2021), Accepted (June 10, 2021)

eISSN: 2653-1941 IJSH

growth variables are adjustive to estimate the paprika growth. This research field has two types 1st the models are using training data, they get the sensor data under field conditions, randomness caused by non-linear data and cluttered environments is inevitable, which aims to sectionalize sensor data to extract attributes, thus potentially reducing the quality [3]. 2nd, the models rely on manual training and testing data planned characteristics, which is large procedure complications. The generality quality of the solution's non-linear data-set quality is poor. Researches should analyze a more workable and stronger overture. Linear regression (LG), which is a futuristic Machine learning (ML) Attitude can straight get an environmental dataset as input mechanically to learn to build feature representations. With a comfortable number of the dataset, machine learning can reach better exactness than orthodox methods [4]. This research has also been used to determine which environmental factors are most important in the growth of leaves. The study's major focus is to assess and compare the performance of the two beds using the linear regression method and machine learning, in which the correlation of paprika growth is identified by using leaf growth and environmental parameters. An ML is used to models the relationship between greenhouse paprika and environmental corresponding growth-related traits (SVM, RF, GBM, and XGB). Then pursuing the planned model, including paprika leaf data, environmental, and machine learning to create from the row planting, this paper will analyze the expected of using machine learning with sensors to accuracy the growth of similar attributes of paprika.

2. Related works

Recently, using data-analysis skills, many scientists have examined the agriculture environmental prediction problem. For modeling, the paprika growth patterns, several statistical and machine learning (ML) methods or developed. Paprika is conquering vegetable marketplaces, not only because of their aureate visuals aspect but and they are one of the high-grade begins of ascorbic acid and pigments, and phenolic bloated, substantive in the human diet. Hence, basic researches related to the production and betterment of the available imported varieties of low-level greenhouse conditions on the Bogota plateau are required to analyze this production of three-color hybrids of paprika using greenhouse conditions [5]. The nutrition water for crop growth in agricultural farming facilities. The supply of nutrition water is not configured with a precise plan but executed conventionally. They propose a prediction method for nutrition water demands based on big data analysis for optimal strawberry industry, analyzed this prediction of nutrition water for the strawberry industry using linear regression [6]. This study suggests a system that proposed to solve the requirements of a charging station infrastructure by providing waiting time information to electric vehicle (EV) drivers for optimal charging options. This system determines the number of vehicle movements, such as entering or leaving the charging station in real-time analyzed this electric vehicle waiting time prediction with open CV and support vector regression [7]. They used machine learning for predicting agricultural soil greenhouse gas emissions. I compared the performance of nine alternative machine learning models. LSTM revealed the best models in foreseeing both N2O and CO2 fluxes. Classical RF was found to be a fast and impressive data-driven model for forecasting CO2 fluxes [8]. The material created in present-day agricultural undertakes is given by a broad range of sensors that allow a high-performance comprehension of the operative condition environment and the activity itself, prompting increasingly precise and quicker basic leadership [9]. The cost of energy needed for smart farming agriculture has a significant impact on agricultural productivity. As a result, prior awareness of energy usage is critical for agricultural energy planning and policy growth. This analysis aims to see how well multiple linear

regression (MLR) and machine learning techniques like support vector regression (SVR) and Gaussian process regression (GPR) can forecast agricultural energy consumption [10]. This research is to comprehend the relation between greenhouse agriculture productions with varying Predictors. As well, we explore the efficiency of various machine learning models (SVM, GBM, XGB, and the RF) in paprika leaf growth and environmental energy prediction for the paprika.

3. Materials and method

In this research, we have used the Greenhouse paprika data in the year October and December 2019 and January to July 2020 (9 months). It bases these paprika data on leaf count, leaf-width and input temperature, output temperature, wind speed, dew point, CO2, and humidity are connected to Paprika growth, etc. we collected data from a local paprika greenhouse production in Korea. I have given a full expansion of the site and paprika production in this research. It provides growth and environmental properties in the additional material and variable [Figure 1].

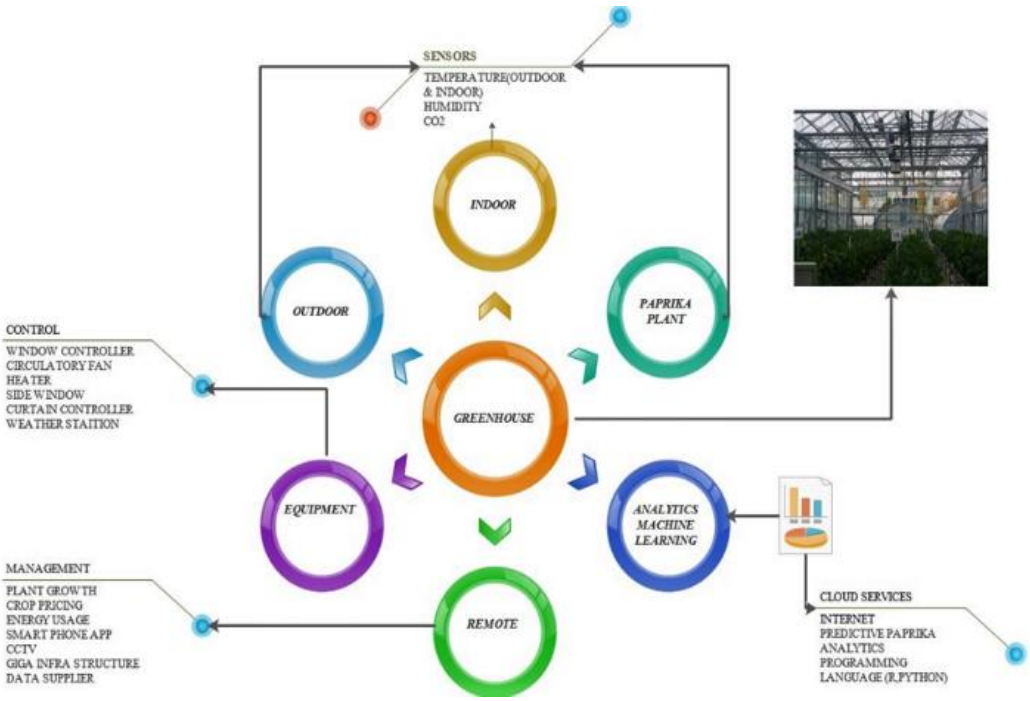

Figure 1. Paprika data collection

Two types of growing plants are using (R1 and R2 mean row planting) locally experienced paprika growths, where all the samples were healthy or unhealthy collected each plant and environmental data. It took the samples to the analytical study for a correlation in paprika ability after collecting the leaf production and environmental data. The leaf cyclically gives our end-to-end of the year, commonly environmental occurred as distinct peaks in January, especially after analyzed on different sensor applications. This paper was the per-processing data obtainable (270 training dataset and 180 validation-dataset).

3.1. Random forest

In the random forests, we can include more data. It can perform well on large databases. The random forest gives the highly accurate output from the collection of decision trees. Each decision tree draws the sample random data and it predicts the accurate result at the end. It makes efficient use of all predictive features and maintains accuracy even if the data is missing.

$$Gini = 1 - \sum_{i=1}^{c}(P_i)^2 \quad (1)$$

3.2. Support vector machine

The SVM algorithm develops a model that increases the separation between data points in each collection, with a hyperplane. The SVM function in R package e1071 can build structure a model given a testing data set and training data set to predict the classification of supplemental data points. SVM is useful because it is quick and there is no danger of over-add-on the data

$$\frac{1}{n}\sum_{i=1}^{n} l\big(f(w, x_i + b, y_i)\big) + \|\omega\|^2 \quad (2)$$

3.3. Gradient boosting machine

The gradient boosting machine function requires you to specify certain statements. You will begin by qualifying the formula. This will include your response and forecaster variables. Next, you will qualify the system of your response variable. We specify if nothing, then the gradient boosting machine will try to guess. At last, we will specify the data and the n. trees statement by default, the gradient boosting machine model will assume 500 trees, which can provide is a good estimate of our gradient boosting machine performance.

$$F(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \beta_m h(x; a_m) \quad (3)$$

3.4. eXtreme gradient boosting

The eXtreme Gradient Boosting is used for pedestrian detection classifiers in this paper. The flow of the boosting algorithm is usually based on the classification and regression tree. In this paper, a tree architecture using the eXtreme Gradient Boosting algorithm from the sample is adopted to get the first estimation result Y1 and the second tree with Y based on the variety between the real label and the predictive label in the previous step. By this, the algorithm error can be effectively reduced.

$$obj^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad (4)$$

4. Result and discussion

In this paper, machine learning models (SVM, RF, GBM, XGB) are used for comparison experiments for the prediction efficiency of paprika leaf and width growth production amount in the greenhouse paprika row (R1 and R2 means Row planting). To find out which input and output constant quantity has the maximum applied mathematical importance, the suggested research methodology conducted the forecast analysis experiments using leaf growth data, which are aggregated in the real testbed during 2 years (from 2019 to 2020), as the training data. The input variables include leaf width and leaf growth, outside temperature, inside temperature, humidity, wind speed, dew point, Carbon dioxide (CO2). Through the test, we find the

greenhouse environment created the maximum paprika correlation coefficients, the minimum error values of significance coefficients, and P-values (correlation coefficients = 0.49 and significance coefficients = 2.29 - 9 for R1 row planting and correlation coefficients = 0.62 and significance coefficients = 3.27-8 for R2 row planting. The result of correlations in R2 row planting shows the greatest correlation coefficient value of 0.62. It means that the linear regression procedure is an effective method to measure the result of all forecasters on paprika. The other result shows that factors, internal temperature, and Carbon dioxide CO2 are the most powerful factors on the released paprika leaf. The best R squared, and RMSE values were acquired by seeing all forecaster's input value (R2 = 0.96 and RMSE = 16.6, R2 = 0.99 and RMSE = 0.04 for passed off R1 and R2 row planting from temperature) [6]. The results from the two characteristics (R1, R2) LR method. From the experiments, the paprika growth uptake was calculable to immediately affect R2 row planting more than R1 row planting. The results from the linear regression model exposed almost related findings that internal and outside temperature was the most important items on R1 and R2 row planting individually, except for the CO2 which was found to be valuable on the R1 row planting, a fact that was also according in other research papers in the literary study. In visual percept of these collections, all input factors (i.e., internal and outside temperature, Carbon dioxide (CO2), humidity, dew point, wind speed) were chosen for the linear regression prognostic logical thinking of paprika growth.

[Figure 2] and [Figure 3] show the mathematical relation metrics for the R1 growth prediction times period gained from the four ML models. The results show on SVM best performs among all the ML models prediction times period (correlation coefficients = 0.91) R1 and R2. The orbit of the variability of SVM models can store actual data through their interior greenhouse paprika data, which act as maximum and minimum datasets. Equivalence to the precision outcomes got by the linear regression model for the same paper site and time intervals for data collection, the best value of correlation coefficients was 0.91. SVM could give better results with correlation coefficients = 0.91. The collection of machine learning models affected well in the training phase but did not execute as well in the prediction phase (correlation coefficients = 0.84 and 0.60 respectively) and this is because of their low effectivity of the basic cognitive process in data series forecast tasks.

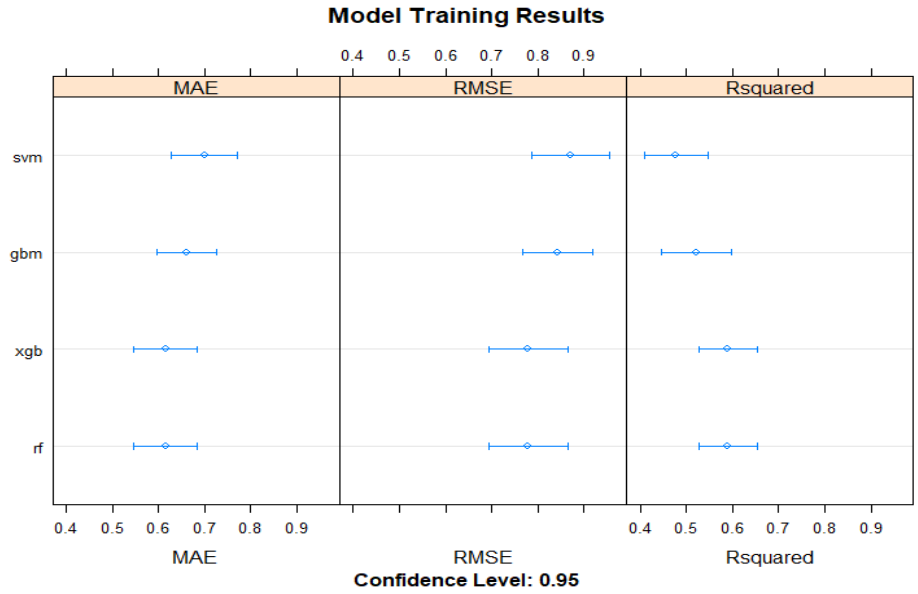

Figure 2. Squared error performance R1

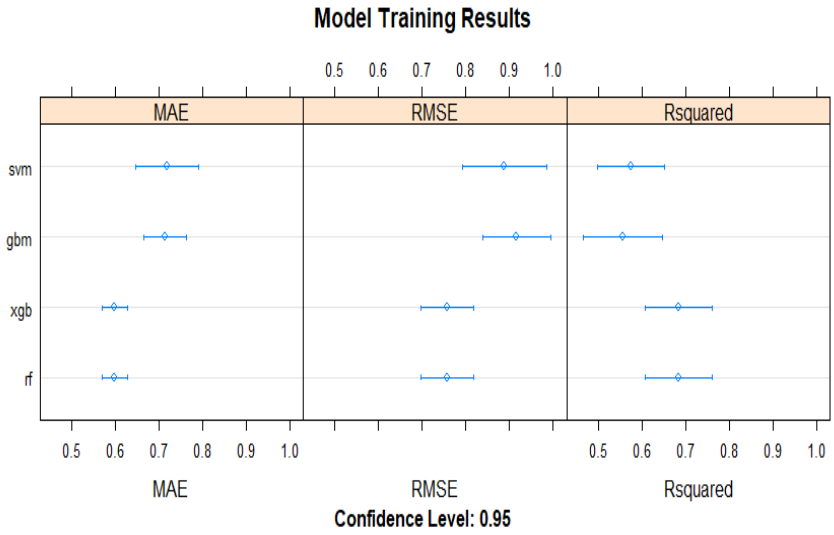

Figure 3. Squared error performance R2

Training error values in XGB (correlation coefficients = 0.77 and 0.60, respectively) and in RF (correlation coefficients = 0.76 and 0.59, respectively) are mainly because of insufficient amount of measured R1 row planting. It means that classical regression models (RF and XGB) are better than the GBM models, particularly in the forecast state. The SVM model is defined by heavy activity and shows the quality to fit a data set. SVM model overcomes all ML models, attempted to produce the best prediction of R2 row planting.

R1 row planting, the comparison with R2 predictions places within the ability of machine learning models to capture the best performance R2 row planting. SVM was the only ML model that complete this challenge and captured the pattern, outstripping all ML models (R = 0.87 and 0.60, respectively) and (the best R was 0.91 with R2 row planting) with less variability equivalence to the previous R1 predictions. The SVM can detect fewer peak patterns falling out in a short period. GBM and XGB performed slender well relative quantity in the prediction form (R = 0.80 and 0.64, respectively) than the training dataset. GBM is another best performance in ML model, the testing dataset forecast level and the experience that peak in the data using GBM input layer can inform this, making it easier for GBM to detect. Training data makes XGB low error accurate for predictions. ML followed outstripped other shallow ML models in the prediction stage (R = 0.42). The SVM model is characterized by compact support and the ability to fit a data set. All ML models, except SVM, failed to produce the best prediction of R2 row planting.

Table 1. The model's performance for R1

Model's	Training data			Validation data		
Performance Measures	RMSE	MAE	R-squared	RMSE	MAE	R-squared
SVM	0.87	0.69	0.65	1.15	0.93	0.72
GBM	0.80	0.63	0.58	0.82	0.64	0.67
XGB	0.74	0.59	0.51	0.72	0.64	0.57
RF	0.72	0.57	0.52	0.69	0.59	0.42

Table 2. The model's performance for R2.

Model's	Training data			Validation data		
Performance Measures	RMSE	MAE	R-squared	RMSE	MAE	R-squared
SVM	0.91	0.82	0.79	1.20	1.02	0.91
GBM	0.82	0.71	0.55	0.86	0.68	0.62
XGB	0.75	0.59	0.68	0.83	0.66	0.72
RF	0.70	0.56	0.62	0.70	0.64	0.68

Finally, this machine learning model is the best activity of the SVM models for leaf growth using environmental in greenhouse paprika. We selected the most befitting model from modes, listed in [Table 2] and [Table 3], which have the lowest predicting error when examination predicted data using a befitting test set. SVM is an effective algorithm for analyzing a large amount of data that are impossible to interpret manually, and its performance improves as more data is added. The SVM did not demonstrate a significant increase in estimation accuracy since this analysis only used 902 points leaf count, leaf width, and leaf area. However, in addition to higher precision, the use of SVM to estimate leaf area has benefits over regression analysis. The model's precision was improved simply by adding the node number, with no extra equipment or labor to measure leaf areas. However, several variables have been suggested for improving the precision of leaf area measurement, including leaf count and width, as well as environmental factors. Also carried out the analysis of the autocorrelation and the RMSE value. The above analyzed linear model that is the most appropriate model for the reasoning data sets depends on the environmental zones from which they originate in the greenhouse. The results got from ML models to a leaf growth in paprika provide valuable insights into the data structures studied and their components, which is a good basis for a suitable prediction. Table 2 and Table 3 include examples of the predictions generating the smallest RMSE and passing the training data and validation data.

5. Conclusions

This study aims to evaluate the paprika development's environmental life cycle in a Korean paprika area. The study's boundary scheme runs from the time inputs are introduced into the greenhouse to the time paprika is harvested. The most critical factors in paprika production are leaf development and the climate, with carbon dioxide (CO2) coming in first. Carbon dioxide (CO2) emissions were estimated to have a 0.94 environmental efficiency. If this index is greater than 0.85, showing that no energy has been removed from the system, it can increase environmental consumption efficiency by combining machine learning strategies to reduce inputs while still increasing paprika yield. Finally, in the development of paprika for environmental consumption, the RF, SVM, XGB, and GBM models predicted environmental efficiency and environmental indicators. The improved SVM model will make management decisions about paprika production at various times and in different situations in the future because of its high precision.

Acknowledgments

This research was supported by the Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University, 255, Jungang-ro, Suncheon-si, Jeollanam-do 57922, Republic of Korea. This research was supported by the MSIT (Ministry of Science and ICT),

Korea, under the Grand Information Technology Research Center support program (IITP-2020-0-01489) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation). This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20202020900060). This study has also been promoted by the GJ-RIP fund.

References

[1] M. H. Aminifard, et al., "Growth and yield characteristics of paprika pepper (Capsicum annum L.) in response to plant density," Asian Journal of Plant Sciences, vol.9, no.5, pp.276, **(2010)**

[2] H. Kirnak, et al., "Paprika pepper yield and quality as affected by different irrigation levels," Tarim Bilimleri Dergisi – Journal of Agricultural Sciences, vol.22, pp.77-88, **(2016)**

[3] S. Venkatesan, S. Sivamani, M. B. Lee, J. W. Park, C. Shin, and Y. Cho, "A comparative study of forecasting strawberry production based on random forest and support vector machine," Information Technology Convergence Engineering Journal in Korea, vol.9, pp45-52, (2019)

[4] M. L. Maskey, T. B. Pathak, and S. K. Dara, "Weather-based strawberry yield forecasts at field scale using statistical and machine learning models," Atmosphere, vol.10, pp.378, **(2019)**

[5] F. B. Peña and I. Z. de Polanía, 'Growth of three-color hybrids of sweet paprika under greenhouse conditions," Agronomía Colombiana, vol.33, no.2, pp.139-146, **(2015)**

[6] S. Venkatesan, "A prediction of nutrition water for strawberry production using linear regression," International Journal of Advanced Smart Convergence, vol.9, no.1, pp.132-140, **(2020)**

[7] R. Muhammad Fikri and M. Hwang, "Electric vehicle waiting time prediction with openCV and support Vector regression," Asia-pacific Journal of Convergent Research Interchange, vol.6, no.10, pp.137-146, **(2020)**

[8] A. Hamrani, A. Akbarzadeh, and C. A. Madramootoo, "Machine learning for predicting greenhouse gas emissions from agricultural soils," Science of The Total Environment, vol.741, pp.140338, **(2020)**

[9] P. B. Lohiya and G. R. Bamnote, "A machine learning model for the growth of agriculture industry," IOSR Journal of Engineering, vol.10, pp.2278-8719, **(2020)**

[10] Z. Ceylan, "Assessment of agricultural energy consumption of Turkey by MLR and Bayesian optimized SVR and GPR models," Journal of Forecasting, Sep., vol.39, no.6, pp944-56, **(2020)**

Authors

Saravanakumar Venkatesan

He is currently pursuing a Ph.D. in the Department of Information and Communication Engineering, Sunchon National University. He received his Bachelor's degree in Mathematics from Madras University and a Master of Science in Information and Communication Engineering from Sunchon National University in South Korea. His current research interests include Big Data Analytics, Data Mining, Mathematics

Jonghyun Lim
Completed Master's degree in Information and Communication Engineering from Korea. He currently studying for a Ph.D. degree in Information and Communication Engineering at Sunchon University. His area of interest includes System Software, Ubiquitous.

Changsun Shin
Received the Ph.D. degree in Computer Engineering at Wonk Wang University. Currently, he is a professor of the Department of Information and Communication Engineering at Sunchon National University. His researching interests include Distributed Computing, IoT, Machine Learning, and Agriculture/ICT Convergence.

Yongyun Cho
Received the Ph.D. degree in computer engineering at Soongsil University. Currently, he is an assistant professor of the Department of Information and communication engineering at Sunchon National University. His main research interests include System Software, Embedded Software, and Ubiquitous Computing.

This page is empty by intention.