# Predictive models for Lettuce quality from Internet of Things-based hydroponic farm.

Sethavidh Gertphol*1, Pariyanuj Chulaka#2, Tanabut Changmai*3
*Department of Computer Science, Faculty of Science
#Department of Horticulture, Faculty of Agriculture
Kasetsart University, Bangkok, Thailand
1sethavidh@gmail.com, 2agrpnc@ku.ac.th, 3tanabut.cha@ku.th

***Abstract*—As changes in the environment affect quality and quantity of crop, yield forecasting becomes very important for farmers. Thailand's economy depends on agriculture for a very long time, and lettuce is sold at high price. The authors had deployed smart hydroponic lettuce farms using the Internet of Things to collect environmental data and control the farm's operation in real time. The experiment generated large dataset which was used to create regression models using machine learning techniques in this study. Features used include environmental data such as the amount and intensity of light, humidity, temperature, together with weekly measurement of plant growth. Target variables are total fresh weight, nitrate content, number of leaf, and leaf area. RMSE was used as a measure for model selection. Models were created using several linear regression techniques, as well as newer techniques, such as, SVR, MLR, and ANN and the best ones were selected for each week prediction. Using only environmental data up to the 3rd week of planting, our predictive model performs 24.44% better than SVR when predicting total fresh weight, 13.93% better when predicting nitrate content, 0.47% better for number of leaves and 12.04% better for leaf area. When using additional plant growth data, our model is 13.09%, 5.52%, 19.47% and 5.06% better than SVR when predicting total fresh weight, nitrate content, number of leaf and leaf area, consecutively.**

***Keywords— Architectures; Lettuce; Predictive models; Machine Learning; Smart Farm; Hydroponics***

## I. INTRODUCTION

Thailand depends on agricultures for a very long time and most Thai people work mainly in the agricultures. The country's land and environment are suitable for agriculture.. However, farmers have difficulties in predicting the crop yield accurately since the yield is influenced by various factors, including environmental factors and External factors are affecting the plant's quality[2][5]. With the coming of the Internet of Things, which is a new type of computing system where small electronic devices equipped with sensors are used to detect the operating environment, smart agriculture or smart farming [12][13] is possible. The application of Internet of Things to grows crops has the potential of saving labor and resources, more fine-grained control in watering and fertilization, and more accurate gathering of information about planting environment.[1]

Machine learning is developed and extended from pattern recognition. Machine learning is related to a study and creation of algorithm that can learn and predict the data, using computer system with supervised learning or unsupervised learning technique, in order to acquire the model or pattern of the data in various forms, depending on the requirement and learning goal of the machine. Normally, this technique will perform a model validation process on every learning process. Learning algorithm operates by using the model acquired from sample dataset, in order to predict or make a decision later, instead of following procedural instruction from computer program. This technique allows the machine to learn from the sample dataset or the environment, with the goal to develop or improve the system's efficiency. [11] Machine learning play a role in forecasting crop yield. Many researchers improve the algorithm to decrease error rate, such as applying Machine learning techniques to predict yield by clustering, K-means, K-Nearest Neighbor, linear regression, artificial neural network, Support Vector Regression (SVR), Random Forests (RF) and Regression Trees (RT) [20][21]. These are a powerful and flexible framework of the key characteristics of the Machine learning techniques that make them widely used in many domains, and highly applicable to incorporate expert knowledge into the system. Machine learning enables better decision making and informed actions in real-world scenarios without (or with minimal) human intervention. [22]

In this research, the authors developed a predictive model for the crop yield, using machine learning techniques, where the system learned from the dataset which is a collection of environmental data, gathered from the hydroponic lettuce farms using the Internet of Things system[1]. The system collects environmental data every 5 minutes for the entire duration of the farming and we measured the plant's growth every week. The amount of crop yield and harvestable were also measured at the last week of farming. The collected dataset was used for teaching several predictive models. In this research, the author chose to create the following models, namely, Support Vector Regression (SVR), Multiple Linear Regression (MLR), Artificial Neural Networks (ANNs) and The Linear model module implements generalized linear models. All of the aforementioned models were accessed from the library of SCIKIT Learn, using Python Version 2.7; the tool the author used for testing and building the process crop yield predictive model for this research.

The paper is organized as follows. Section II contains Literature Review. Section III describes Dataset consisting of Input Variable, Target Variables and Prediction scenario. Section IV outlines out Methodology and explains about how to prepare and normalize data and apply prediction. Section V shows our results, while Section VI discusses other interesting aspects of gained from the research..

## II. LITERATURE REVIEW

Lettuce is highly popular among the people who care for their health, and sold at a very high price in the market. In Thailand, lettuce is widely produced in the hydroponics farm using Nutrient Film Technique (NFT). The lettuce produced from hydroponic farm has better taste, and requires less time to grow, in comparison to soil-grow lettuce. Hydroponic farming also facilitates the control of necessary nutrients for plant. [5]

One concern is the leftover nitrate content in the vegetable. This is caused by excessive use of nitrogen-based fertilizer, and as a result, the vegetable accumulates high amount of nitrogen. Once human or animal consume such vegetable, it can cause serious health issue. Therefore, there are efforts to reduce the amount of leftover nitrate content in the leafy vegetable [6]. In hydroponics farming that uses nitrogen-based fertilizer solution, the farmer must soak the plant in pure water for 5 – 7 days before harvest [7]. Also,

978-1-5386-8164-0/18/$31.00 ©2018 IEEE

the farmer must control lettuce's nitrate content at different level for individual farming season: less than 2,500 mg/kg during rainy season and summer, and less than 3,000 mg/kg during winter [8].

Internet of Things is a novel concept for a computer system that has small electronic equipments, equipped with sensor, that detect the system's environmental data during operation. The system will be able to record the data, and control its operation or create new feature for the system. Automatic operation is among the goals for adaptation to Internet of Things, which can be used whether in the household area or office area. Smart agriculture or smart farming [12][13] is an adaptation of Internet of Things for agricultural purpose, with the potential of labor saving and efficient resource usage, as well as the ability to gather more accurate data about the farming's environment, especially for the hydroponics farming[1].

Predicting agricultural crop yield is one of the aspects that have been studied for a very long time. There are several methods available for predicting agricultural crop yield, such as, Base regression [3], Biological system modeling [14], statistics [15], and machine learning. [12][21][22] Crop yield prediction allows the farmer to prepare storage area for the produce, and to create appropriate marketing plan for the harvested crop, as well as to anticipate the expected income in the future, and thus plan his business development accordingly.

Machine learning is used for making the model that presents accurate prediction [16]. It is used for many aspects of the crop yield prediction, whether for short-term prediction [17], long-term prediction [16], and time series prediction [18]. The widely used prediction models are, namely, Support Vector Regression (SVR), Multiple Linear Regression (MLR), Artificial neural networks (ANNs) and The Linear model. The model used for crop yield prediction must support variables that affect the amount of crop yield, such as, environmental variables [2][4], topographical variables, as well as growth-related variables, in order to provide accurate prediction result. The result can be validated, using root mean square error (RMSE) and mean absolute error (MAE) to identify the model's quality, and using mean absolute percentage error (MAPE) to identify any error [19].

## III. DATA SET

In order to measure the lettuce's quality according to method used in agricultural research, the vegetable's growth will be measured every week [9], namely, the number of leaves, plant's height and width, and stem's diameter. Another measurement will be performed during the harvesting week, namely, the amount of leaf, the plant's height and width, the stem's diameter, fresh weight, dry weight, nitrate content, vitamin c content. Doing so presents clear picture of the plant growth, during each week of farming leading to the harvesting week. The intensity and light duration of sunlight also affects the lettuce's growth and quality [10].

The dataset used for this research was gathered from the previous research, which collected the lettuce farming data in Nonthaburi Province. The dataset was gathered from February 2017 to August 2018 and consisting of 326 samples among 8 farming cycles. The important features of our dataset consist of environmental factors during farming, growth factors measured during the 1st – 5th week of farming, on a weekly basis and growth factors measured during the harvesting on the 6 week. This research presents the data analysis methods that provide prediction result with highest accuracy.

### A. INPUT VARIABLES

1) Environmental data was gathered from the server that monitors and collects data from the Internet of Things built into the lettuce farm. The data was recorded on a daily basis and considered the following variables: light intensity, humidity, air temperature, water temperature, EC, and pH value of nutrient solution. Each parameter was then summarized into weekly maximum, minimum, and average value. The data about light duration needed special handling since receiving specific level of sunlight causes the plant to growth at different rate. The light duration were divided into several intensity level: 540 (Lux), 1080(Lux), 2160(Lux), 3240(Lux), 4320(Lux), 5400(Lux), 8100(Lux), 10800(Lux) and 16200(Lux). We then calculated how long the plants received each intensity level per week. There are in total 28 features about the environment per week. Since the duration of each planting was 6 weeks, we have in total 168 environmental features.

2) Lettuce's growth data that was gathered on a weekly basis during the 1st – 5th week of farming: number of leaf (leaf), plant's height (cm), plant's width (cm), stem's diameter (mm.) and leaf's area (cm2). Consequently, there are 25 features for the 5 planting weeks (5 features per week).

### B. TARGET VARIABLES

1) Lettuce yield that was gathered on the harvesting were: number of leaf (leaf), plant's height (cm.), plant's width (cm.), stem's diameter (mm.), leaf's area (cm2), shoot, root and total fresh weight (g), shoot, root, and total dry weight (mg), and nitrate content (mg./kg), for a total of 12 target variables.

### C. PREDICTION SCENARIO

The author designed two scenarios for prediction according to the real-world ability of the farmer to collect data. The scenarios are as follow

1) Scenario 1: this dataset is used for the predictive model that considers the crop yield during farming. So this means that the farmer has to measure plant's growth every week which could be time-consuming. The features used in this scenario are the environmental data and growth data during the 1st – 5th week of farming, which can be customized into weekly environmental data for 168 features, in combination with growth data, for 25 features, for a total of 194 features.

2) Scenario 2: this dataset is used for the predictive model that only considers the environmental factors and the farmer does not have to measure plant's growth every week. The features used in this scenario are the 168 weekly environmental data only.

It is expected that the model from scenario 1 will be more accurate than the one from scenario 2 since it uses more features. However, it is interesting to see how much better the scenario 1's model is and whether it is worthwhile to collect the plant's growth data or not. Because plant measurement use a lot of time comparing to automatic environmental data collect from the Internet of Things devices, scenario 2 will save labor cost which could be significant if the farmer has large planting.

Data from the 6th week, which is the harvesting week, is used as the target for the model to learn in both scenarios.

## IV. METHODOLOGY

The author uses SCIKIT Learn, which was developed using Python version 2.7, as a tool for creation of the predictive models.

### A. Prepare data and Normalize

Environmental data was prepared into two groups, namely, daily environmental data of each week and accumulated environmental data of each week. Growth data was prepared into 2 groups, namely, growth data from the 1st – 5th week of the farming and growth data from the harvesting week.

The data was then normalized where the range of each feature is between 0 to 1. Then the data was split into training and testing sets with the testing set comprises 10 percent of the data points. We then used 10-fold cross validation to rotate the testing set for each validation.

### B. Processing

Combination of environmental data and growth data on a daily basis works best with predictive model that supports multiple variables. Since the first group of models uses linear equation with multiple variables, the author thus chooses the following linear models, namely, SGDRegressor, BayesianRidge, LassoLars, ARDRegression, PassiveAggressiveRegressor, TheilSenRegressor,

LinearRegression, as well as later developed models that also support multi variables, such as, SVR, MLR, and ANN.

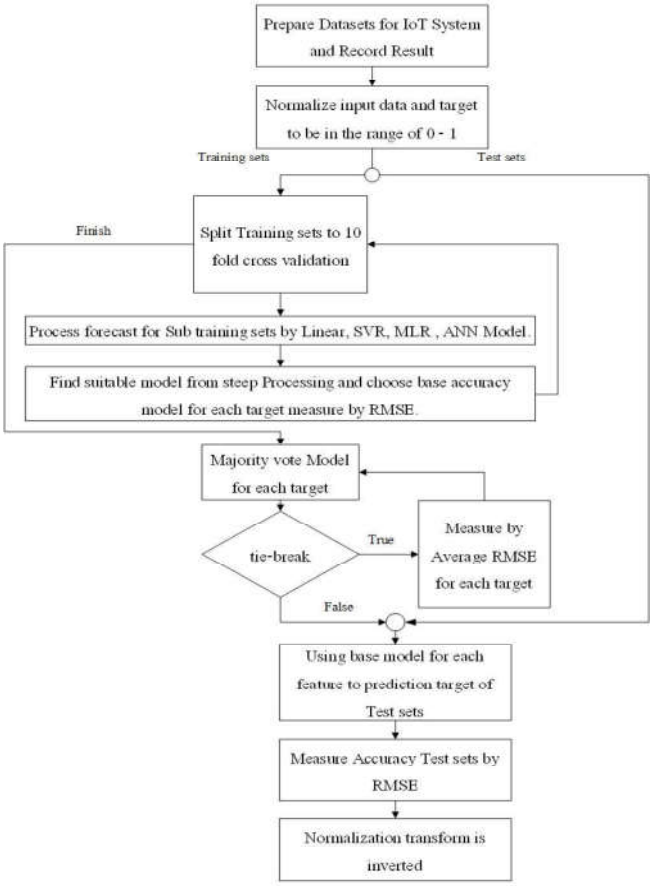


FIGURE 1. Process Flow Charts

For each target variable, we use each previously mentioned algorithm to create one prediction model for each week. This means that the week1 model uses only the data available for the first week to make prediction. The week2 model uses two-week worth of data and so on until the week5 model that uses all 5 weeks data for prediction. This reflects the availability of additional data to the farmer when planting season progresses.

During each validation fold, we used 10 algorithms to create 10 prediction models for each week. From those 10 models we calculated the RMSE of each model and picked the one with lowest RMSE as the winner of that fold. After the validation were done across 10 folds, we find the majority vote where the algorithm that wins the most often is declared winner. If there is a tie, we calculated the average RMSE of the models produced by the tied algorithms, and the one that is responsible for the lowest RMSE is declared winner.

The process then repeats for another week. When we find the winner for all 5 weeks, the process repeats again for another target variable. When all target variables are done, we repeat the process for the second scenario. In the end, we have one best model for predicting a target variable using data for each week according to our scenario.

## V. RESULTS

Due to limited space, we present only the result for the most important indicators of crop yield's quality, namely, total fresh weight, nitrate content, number of leaves, and leaf area. All of the aforementioned factors affect produce's selling price

Figure 2 to 5 shows the error rate of our method compared with SVR for both scenarios. The reason that we pick SVR for comparison is that it often produces a winner for each week. However, it is clear that our method outperforms SVR especially during later weeks. For example, our method has 0.1149 error predicting the total fresh weight at week5 compared with 0.1845 using SVR.

When comparing weekly predictions, the later weeks provides better accuracy. This is not a surprise because we have more data when the planting progresses. However, our method in Scenario 1 (that also uses plant's growth data) results in the most drop of error rate during later week especially when predicting total fresh weight and number of leaves. Error rate of SVR Scenario 1 in comparison almost flat-line after week 3 when predicting the number of leaves. This shows that SVR fails to make use of extra information available during later weeks to provide better predictions.

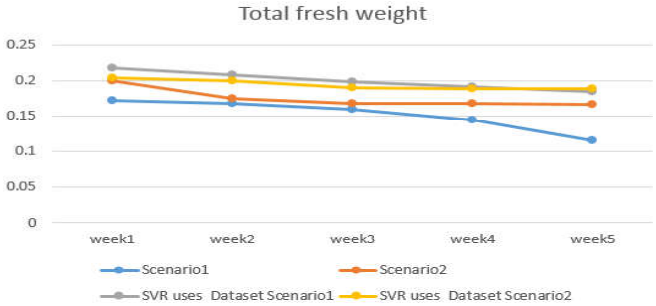


FIGURE 2 Error rate of total fresh weight each week.

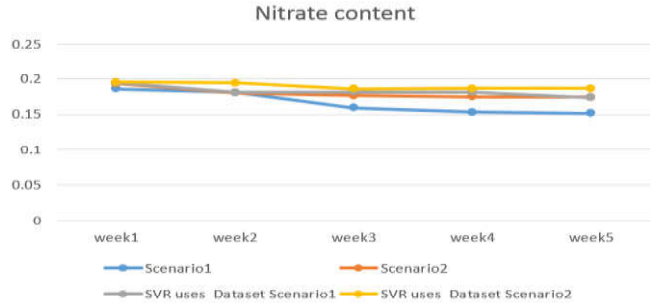


FIGURE 3 Error rate of Nitrate content each week.

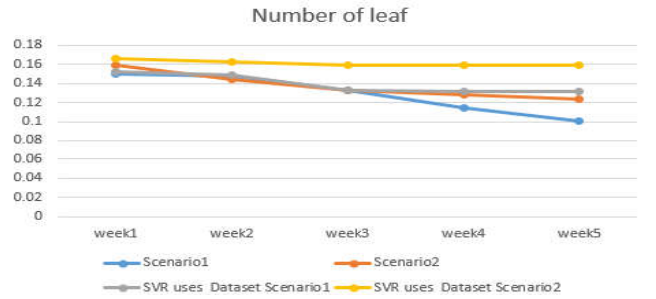


FIGURE 4 Error rate of Number of Leaf each week.

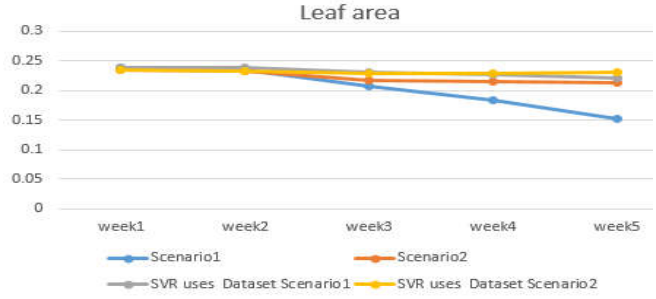


FIGURE 5 Error rate of Leaf area each week

Predictions using Scenario 1 is better than Scenario 2 as expected, but it is interesting to see that the error rates of the two Scenarios are mostly identical during the first 3 weeks when predicting many targets (except nitrate content). Figure 6 shows predictive power of our method during 3rd week for the 4 target variables. It shows that dataset gathered for a duration of 3 weeks is

sufficient to provide comparable result for both Scenarios. Even though dataset acquired from the 4th – 6th week gives an edge for Scenario 1, it requires additional period of 1-3 weeks to provide prediction with higher accuracy, therefore, the farmer will have less time to resolve issues if result from models show low quality. It also means that it may not be worthwhile to collect plant's growth data during the first 3 weeks because it does not improve accuracy much.

Our predictive model as week 3 performs 24.44% better than SVR when predicting total fresh weight, 13.93% better when predicting nitrate content, 0.47% better for number of leaves and 12.04% better for leaf area. When using additional plant growth data, our model is 13.09%, 5.52%, 19.47% and 5.06% better than SVR when predicting total fresh weight, nitrate content, number of leaf and leaf area, consecutively. These show detail in table 1.

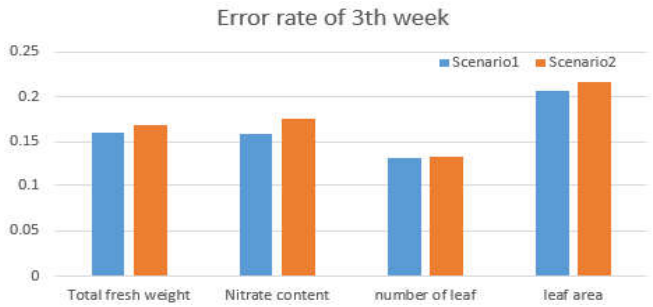


FIGURE 6 Error rate of week 3 models using our method.

Table 1 shows the winning algorithms and their corresponding error rate for each week and each target variable.

| | | | week1 | week2 | week3 | week4 | week5 |
|---|---|---|---|---|---|---|---|
| Fresh weight | Scenario1 | algorithm | ANN | TheilSenRegressor | ARDRegression | BayesianRidge | BayesianRidge |
| | | error | 0.1722 | 0.1676 | 0.1596 | 0.1441 | 0.1149 |
| | Scenario2 | algorithm | SGDRegressor | LinearRegression | LinearRegression | TheilSenRegressor | TheilSenRegressor |
| | | error | 0.1997 | 0.1757 | 0.1681 | 0.1688 | 0.1672 |
| Nitrate content | Scenario1 | algorithm | TheilSenRegressor | ARDRegression | ARDRegression | ARDRegression | ANN |
| | | error | 0.1853 | 0.1802 | 0.1591 | 0.1532 | 0.1512 |
| | Scenario2 | algorithm | SGDRegresso | LinearRegression | LinearRegression | ARDRegression | BayesianRidge |
| | | error | 0.1927 | 0.18029 | 0.176 | 0.1744 | 0.1741 |
| Number of leafs | Scenario1 | algorithm | ANN | BayesianRidge | BayesianRidge | ARDRegression | ANN |
| | | error | 0.1504 | 0.1473 | 0.1326 | 0.1147 | 0.1006 |
| | Scenario2 | algorithm | ANN | LinearRegression | SGDRegressor | SGDRegressor | ANN |
| | | error | 0.1592 | 0.1446 | 0.1332 | 0.1284 | 0.1241 |
| Leaf area | Scenario1 | algorithm | SGDRegressor | BayesianRidge | BayesianRidge | ARDRegression | ARDRegression |
| | | error | 0.2384 | 0.235 | 0.2069 | 0.1843 | 0.152 |
| | Scenario2 | algorithm | SGDRegressor | MLR | LinearRegression | TheilSenRegressor | PassiveAggressiveRegressor |
| | | error | 0.2342 | 0.2333 | 0.2173 | 0.2144 | 0.2129 |

TABLE 1 Winning algorithm and error rate.

## VI. DISCUSSION AND CONCLUSION

This research presented data collected from Internet of Things for further analysis on a weekly basis, namely, environmental data, growth data, and crop yield data. The gathered data was used for creation of the dataset that provide accurate prediction, in combination with machine learning model. The authors focus on convenient use of the farmer. The findings tell us that we cannot rely on single model to produce accurate crop yield prediction. Therefore, the author implemented Integrated Models for the crop yield prediction. Our research also shows that predicting at week 3 may be the most efficient because using environmental data is comparable to using both environmental and plant's growth data.

This research has not yet implemented Feature Selection in creation of the machine learning model. Therefore, the result was not quite good enough. Moreover, measuring plant's growth may involve some levels of error, such as, while the author measured the lettuce's height, someday the lettuce was standing straight, and the other day it might sway to the side a little from the blowing wind. Other error could be found from the stem's width, whereas the stem's width is subjected to the amount of sunlight and temperature that day. High amount of sunlight and high temperature cause the lettuce's leaf to wither and the plant to unfold. On the contrary, on a day with less sunlight and lower temperature, the lettuce's leaf will not unfold. Therefore, measuring the plant's width on the day with lower sunlight and temperature would result in a smaller width, in comparison to the measurement made on the day with higher sunlight and temperature. Ultimately, these errors caused the model to fail to achieve a good learning result.

## REFERENCES


[1] Tanabut Changmai, Sethavidh Gertphol and Pariyanuj Chulaka. "Smart Hydroponic Lettuce Farm using Internet of Things". International Conference on KST, 2018, Chiang Mai, Thailand.

[2] Haedong Lee and Aekyung Moon. "Development of yield prediction system based on real-time agricultural meteorological information". International Conference on Advanced Communication Technology, 2014, Pyeongchang, South Korea.

[3] I.Becker-Reshef, E.Vermote, M.Lindeman and C.Justice. "A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data". Remote Sensing of Environment, vol 114, Issue 6, 2010, pp. 1312-1323

[4] S. Skakun, B. Franch, J.-C. Roger, E. Vermote, I. Becker-Reshef, C. Justice and A. Santamaría-Artigas. "Incorporating yearly derived winter wheat maps into winter wheat yield forecasting model" IGARSS, 2016, Beijing, China.

[5] Arnat Tancho. 2005. "Soilless Culture with Hydroponics". TRIO ADVERTISING AND MEDIA CO.,LTD, Chiang Mai. Thailand, Vol 1, pp. 26-42.

[6] Anjana Shahid Umar, Muhammad Iqbal. "Nitrate accumulation in plants, factors affecting the process, and human health implications. A review". Agronomy for Sustainable Development, 2007, Vol 27, Issue 1, pp 45–57

[7] Arnat Tancho. 2005. "Soilless Culture with Hydroponics". TRIO ADVERTISING AND MEDIA CO.,LTD, Chiang Mai. Thailand, Vol 1, pp. 26-42.

[8] Opinion of the Panel on Contaminants in the Food chain, "Nitrate in vegetables Scientific". European Commission to perform a scientific risk assessment on nitrate in vegetables, The EFSA Journal (2008)Journal number, 689, 1-79.

[9] Narasak Boonmee, Use of Fish-culture Wastewater for Lettuce Production in Hydroponics. Kasetsart University, Bangkok,Thailand, 2013.

[10] Petchthai, P.and Thongket, T. "Effect of Light Intensity and Light-Exposure Duration on Growth and Quality of Lettuce" Songklanakarin Journal of Plant Science, Vol. 4, No. 3, 2017 pp. 54-59.

[11] Tom M. Mitchell. "Machine Learning" China Machine Press, Beijing China, 2003.

[12] S R Prathibha, H. Anupama, and M. P. Jyothi, IOT Based Monitoring System in Smart Agriculture. Electronics and Communication Technology (ICRAECT). 16-17 March 2017.

[13] S. Jaiganesh, K. Gunaseelan, and V.Ellappan "IOT Agriculture to improve Food and Farming Technology". Emerging Devices and Smart Systems (ICEDSS 2017). 3-4 March 2017.

[14] Building an operational system for crop monitoring and yield forecasting in Morocco. Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 2013. Page(s) 466 – 469

[15] H. Kerdiles, F. Rembold, O. Leo, H. Boogaard and S. Hoek. "CST, a freeware for predicting crop yield from remote sensing or crop model indicators: Illustration with RSA and Ethiopia". International Conference on Agro-Geoinformatics, 2016 , Page(s):1 – 6.

[16] Monique Pires Gravina de Oliveira, Felipe Ferreira Bocca, Luiz Henrique Antunes Rodrigues. "From spreadsheets to sugar content modeling: A data mining approach". Computers and Electronics , Vol 132 ,2017, Page(s) 14–20.

[17] P.J.C.Vogler–Finck, P.Bacher and H.Madsen. "Online short-term forecast of greenhouse heat load using a weather forecast service". Applied Energy, Vol 205, 2017, Page(s) 1298-1310

[18] LiHong-ying, HouYan-lin, ZhouYong-juan and ZhaoHui-ming. “Crop Yield Forecasted Model Based on Time Series Techniques”. Journal of Northeast Agricultural University (English edition), Vol 19, Issue 1, 2012, pp.73-77

[19] M.van der Velde and L.Nisini. “Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015” Agricultural Systems, Available online 20 July 2018.

[20] A.T.M Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman and Rashedur M Rahman. “Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh”. SNPD 2015, June 2015, Takamatsu, Japan

[21] Monique Pires Gravina de Oliveira, Felipe Ferreira Bocca and Luiz Henrique Antunes Rodrigues. “From spreadsheets to sugar content modeling: A data mining approach”. Computers and Electronics in Agriculture 132 (2017), pp 14–20.

[22] Anna Chlingaryan, Salah Sukkarieh and Brett Whelan. “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review”. Computers and Electronics in Agriculture, vol 151, August 2018, pp 61-69