Muş Alparslan Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi
3(1),12-21,2022

Araştırma Makalesi/Research Article

Comparison of Different Machine Learning Methods for Estimating Agricultural Products

Halit ÇETİNER
Vocational School of Technical Sciences, Isparta University of Applied Sciences, Isparta, Turkey, ORCID: 0000-0001-7794-2555

Corresponding Author:halitcetiner@isparta.edu.tr

Geliş tarihi:29.04.2022 Kabul tarihi:27.05.2022

ABSTRACT

Regression analysis was carried out with two different machine learning methods in order to realize the yield estimations of crops that have different effects on the world economy. The two methods used in regression analysis are SVM (Support Vector Machine) and HTGA (Histogram Based Gradient Augmentation) regression analysis methods. Both regression analysis methods try to find the yield estimation underlying many different agricultural problems with the least error. In this sense, in order to carry out experimental studies, yield estimations were made using precipitation, temperature, pesticide input values found in the World Data Bank and FAO World Agricultural Organization databases in the 35-year interval from 1961 to 2016. As a result of efficiency estimations, 94% R^2 score was reached with HTGA, while 91% R^2 score was reached with SVM Poly core. In the SVM method, regression analysis was performed using 3 different kernels. Poly core value results reached 91% R^2 scores, while RBF and Linear core values reached 81% and 69% R^2 scores, respectively.

Keywords: SVM, HTGA, Yield estimation, Regression analysis

1. INTRODUCTION

In order for the food supply to continue without decreasing, the problems in agriculture must be eliminated. Agriculture is an important area in both the global economy and employment. In this important field, it deals with different problems such as crop productivity forecasting, food supply and security, climate change. Within the scope of this article, methods to find a solution to the crop yield estimation problem have been researched in order to support agricultural producers among the mentioned problems. At the same time, research has been carried out in order to find out whether the yield estimation parameters used in this study are sufficient and to find solutions to this problem. If the yield estimation problem is solved, agricultural producers or farmers have the opportunity to make future plans for the post-harvest period. They have conducted a study to analyse the effects on wheat crop due to the climate effect in the southeast region of Turkey [1]. As a result of the study, it is estimated that the wheat yield will decrease depending on the temperature as a result of the effects of climate change. In support of this study, it is stated that the increase in temperature differences and uncertainties in precipitation patterns also affect crops and food security [2], [3]. In another study, it is seen that yield estimation is carried out by using soil characteristics and spatial information as inputs in addition to climate characteristics for the purpose of crop yield estimation. It is stated that yield estimation is very effective in agricultural disaster scenarios, food security, supply and trade policies due to its predictability [4].

According to the world agriculture organization FAO, there has been a large increase in food insecurity in the last five years [5]. Researchers doing research in the field of yield estimation have also taken into account the relationship with the population while examining the studies to be done to ensure food security and to continue the food supply. It is calculated that a population of two billion will be added to the current population in the next thirty years [4], [6]. Here, although there is no population that provides food insecurity, it is important to ensure the sustainability of production in a way to meet food consumption in reducing food-related obstacles and problems. Experimental studies have been carried out with machine learning methods in order to find solutions to an important food problem due to the critically important points mentioned. In order to conduct experimental studies with two different machine learning techniques called HTGA and SVM, a data set with a wide range of data and kernel parameters affecting efficiency in the literature was selected [7]. Crops in this dataset consist of paddy field, sorghum, plantains and others, potatoes, cassava, corn, yams, wheat, soybeans, sweet potatoes. Briefly, when we examine these products, cassava is a tuber that can be produced in arid and infertile soils that are grown intensively in the sub-Saharan region of Africa [8]. Corn crop, on the other hand, is a staple food that is expected to increase exponentially, as with other crops, the problems in food sustainability [9]–[11]. Plantains and others, on the other hand, are a high vitamin C, carbohydrate and potassium storage crop from the Musaceae family [12]. Potato crop is an important food containing magnesium, iron, vitamin C [13]. Rice is a crop on which yield estimation and planting adjustments are made, since it is nutritious and supply continuity is important [14]. Sorghum (Sorghum bicolor L. Moench) is used extensively to wean babies as a staple grain food, especially since it can grow in hot weather [15]. Crops are gaining more and more importance in terms of food security, supply and sustainability of crop supply [16]. Different varieties of sorghum are used not only for human nutrition but also for animal husbandry [17]. Soybean, which is a crop with difficulties in yield estimations, is a food crop that requires a lot of water and is consumed by converting different types of production such as soy flour and meal [18]. As in other crops, it is seen that the need for soybean increases in direct proportion to the world population [19]. Sweet potatoes are a high nutritional value crop with lobed leaves, starch-storage roots, can be produced in different climatic conditions [20]. Yams are a tropical root tuber crop used in the food and pharmaceutical industries [21]. Wheat, the final crop, is a high-nutrient, protein-providing and high-calorie crop like corn [22].

When the researches on crop yield estimation are examined comprehensively, it is seen that statistical-based models have been developed [22]–[24]. Parameters with seasonal effects such as temperature and precipitation were used as inputs to these models. In the study, in addition to these parameters, the usage rates of pesticides in tons were added. Two different models have been developed using SVM and HTGA, which are important methods of machine learning algorithms. Three different models were tested within the SVM model. In this way, four different models were used within the scope of the article in order to make a yield estimation. Apart from these, although linear and logistic regression methods were also tested, R^2 was not added because it could not reach the desired success rate. The important contributions of the models developed in this context to the literature are summarized below.

- The data has been normalized to a certain range in order to make an effective estimation of the yield from the data consisting of precipitation, temperature and pesticide inputs from 1961 to 2016 in a 35-year interval from different countries.
- Regression analyses were performed for efficiency estimation with poly-core SVM and HTGA methods.
- As a result of the analysis, 94% R^2 score success rate was achieved with HTGA, while 91% R^2 score success rate was reached with SVM.
- In the SVM method, three different models were created with Poly, RBF and Linear kernel values. Among the models created, it was determined that the best one in terms of yield estimation was the Poly kernel.
- To make an accurate comparison by measuring the performances of the mentioned models, the models created were compared in terms of R^2 score, MSE, RMSE, MAE and MAPE measurement metrics.

The remainder of the article consists of three parts. In the next section, the data set used in the training and testing of the models is detailed. In the second part, performance metrics obtained with HTGA and SVM are presented. In the last part, the study is concluded in a way that will guide future studies.

2. MATERIALS and METHODS

2.1. Materials

Within the scope of this article, HTGA and SVM methods are used in regression analyses to perform yield estimation. Experimental studies of Linear regression and Logistic regression methods, especially HTGA, SVM methods, were carried out on the dataset of ten different products from 1961 to 2016 [7].

Table 1. Details of dataset used in crop yield estimation

Raw input year	Inputs					Outputs
Year	Country	Average temperature (℃)	Crop	Average Rain (mm)	Pesticide (ton)	Yield
1990	Canada	8.19	Potatoes	537	29.568	250.994
1991	Brazil	21.42	Wheat	1761	58.349	14.232
….		….	…	…	….	…
1997	India	24.92	Corn	1083	52.279	17.111
….		….	…	…	….	…
2000	Germany	10.96	Soybeans	700	35.273	20.000
….		….	…	…	….	…
2010	Japan	16.12	Plantains and others	1668	55.576	65.135
….		….	…	…	….	…
2013	Turkey	20.21	Sorghum	593	39.440	26.544
….		….	…	…	….	…
2016	Zimbabwe	19.76	Cassava	657	2550.07	46.000

The data set used is presented in Table 1 in detail. Table 1 also includes precipitation and temperature data affecting yield. At the same time, pesticides, which are expressed as abiotic components, are included in the data set as an environmental factor, apart from climate data. When the parameters used in crop yield estimation are examined in the literature, it is seen that the researchers commonly use pesticides, air temperature and precipitation in crop yield estimation [25]–[27]. In this article, cassava, corn, plantains and others, potatoes, paddy field, sorghum, soybeans, sweet potatoes, wheat and yams, which have the highest food consumption in the world, were preferred in the formation of the data set [28]–[31].

Table 2. Total yield amounts of the data set used on the basis of countries

Country	Yield
India	327420324
Brazil	167550306
Mexican	130788528
Japan	124470912
Australia	109111062
Pakistan	73897434
Indonesia	69193506
America	55419990
Turkey	52263950
Spain	46773540

In Table 2, it is seen that the yield totals in all crops are listed from the highest to the least. Apart from these ten countries, yield detail information of one hundred and one more country is available in the dataset. According to the ranking in the table, it is seen that the yield of India is much higher than other countries.

2.2. Methods

Environmental and abiotic factors such as soil properties, temperature and climate variables are needed to determine the parameters that affect the yield [27]. Yield estimates for ten different crops from 101 countries need to be made. Yield estimation of each crop on a yearly basis from the input and output values of different countries has been focused on in two study areas. In the first, it is focused on how the data obtained from FAO and the World Bank are formed. In the second study area, how meaningful information can be obtained from the raw data, the scope and content of which are specified in Table 1. This article focuses more on the second area of study. Using the specified information, HTGA and SVM methods were applied to the dataset of ten different crops. Regression analyses were performed on the yield estimates of the crops using the specified machine learning methods. SVM is a popular machine learning method that is heavily used in tasks such as classification, regression analysis by increasing the distance between different datasets [32]. SVM performs classification or regression analysis by creating planes between class clusters. It is calculated according to the equation given in Equation 1. Defined in Equation 1, $\emptyset$ represents the feature map, while x represents the data points.

$$K(x,z) = \emptyset(x)^T \emptyset(z) \quad (1)$$

HTGA has native support for null values in the dataset. During the training process of HTGA algorithms, the missing values in the data set decide at each division point. In this decision-making process, it is decided whether to go to the child branch or to partition to left and right branches. In general, samples with missing values in the data set are assigned to left or right child branches. The HTGA algorithm is a method inspired by the LightGBM [33] algorithm. LightGBM is much faster than conventional Gradient Boosting Decision Tree based regression algorithms [33]. LightGBM introduces a new method for calculating all split points with gain value.

3. RESULTS AND DISCUSSION

In the experimental results section, the data obtained as a result of the regression analysis obtained by SVM, Histogram Based Gradient Augmentation (HTGA) machine learning methods are presented.

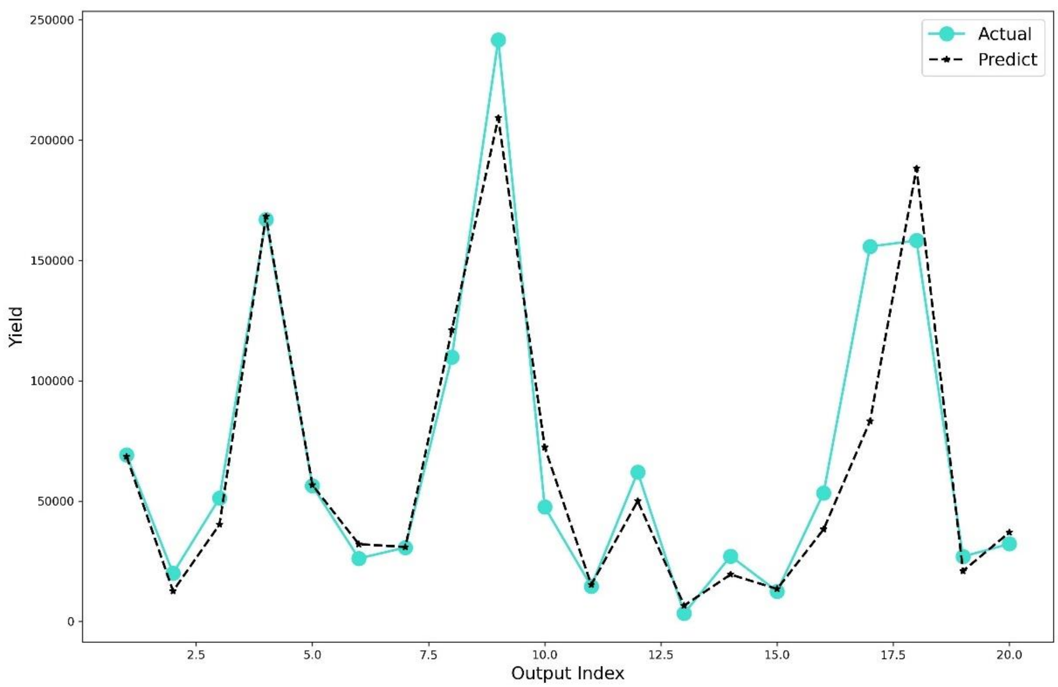

Figure 1. Crop yield estimation with 20 index HTGA

In the regression analysis with the SVM method, not only the Poly kernel value but also the RBF and Linear kernel values were used. However, while 91% success rate was achieved in Poly kernel value, 0.81 and 0.69 R^2 score values were obtained in RBF and Linear kernel values, respectively. The main reason for choosing SVM and HTGA machine learning methods in regression analysis is that R^2 score values are higher than other machine learning methods. While a 0.75 R^2 score value was obtained in the Linear Regression [34] method, a 0.68 R^2 score value was obtained in the Logistic Regression [35] method. Since there is too much difference in scores with SVM and HTGA methods, this study focuses on these two machine learning methods. It is expected that the comparison of the specified methods will form an infrastructure on how to design algorithms and evaluate parameters in the development of a new yield estimation method. In Figure 1, the results obtained at ten different crop yields are shown in a narrow field with an index of 20. In this area, deviations at the lower and upper points are striking. It is seen that there are deviations at the peak between the indices of 7.5 and 10, while deviations occur at the lower points between the indices of 15 and 20. The main reason for this is that some of the ten different crops have lower than average R^2 scores.

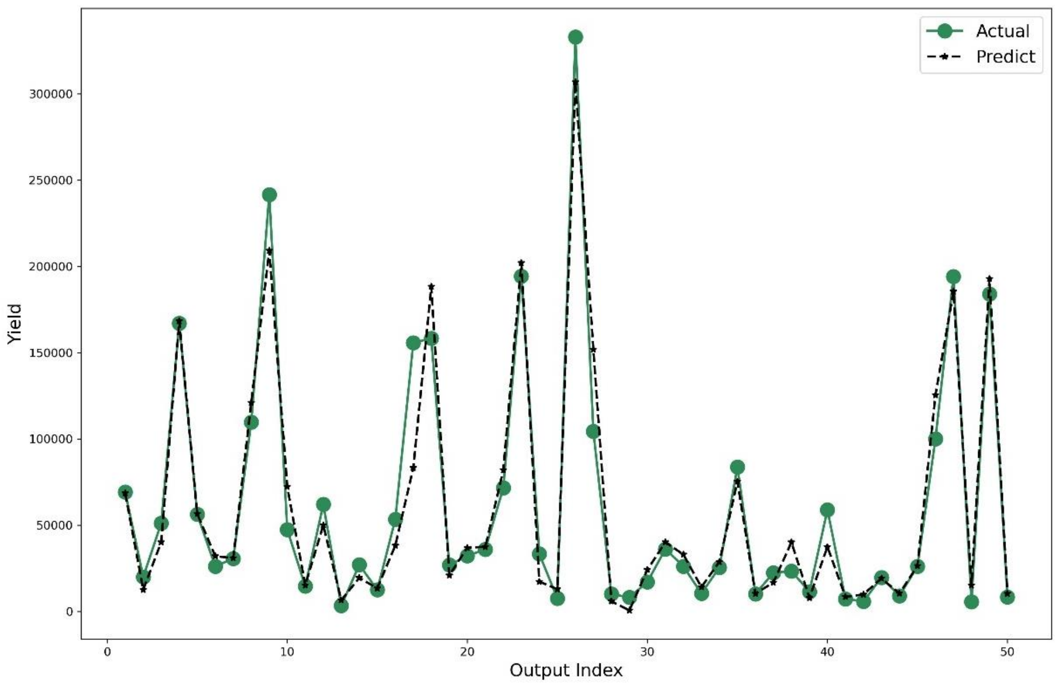

Figure 2. Crop yield estimation with HTGA at 50 scale

In Table 3, the results of the regression analysis obtained by the HTGA method are given. In the results obtained, it is seen that the productivity estimation of soybean is quite low. Yield results for cassava crops were highest, with 70% success rates in maize, plantains, paddy, sorghum and wheat.

Table 3. Analysis results of ten different crops with HTGA regression method

Algorithm	Crop	R^2 Score
HTGA	Cassava	0.9155
HTGA	Corn	0.7621
HTGA	Plantains and others	0.7056
HTGA	Potatoes	0.8837
HTGA	Paddy field	0.7186
HTGA	Sorghum	0.7308
HTGA	Soybeans	0.4458
HTGA	Sweet potatoes	0.8262
HTGA	Wheat	0.7620
HTGA	Yams	0.8680

The estimation results obtained as a result of the use of pesticides, air temperature and precipitation parameters proven in [25]–[27] studies in the literature are shared in a large index in Figure 3. Unlike Figure 1 and Figure 2, the examined length of the data was determined as 20, 50 and 100. The deviations at the peaks and lower points in 20 indices, 50 indices and 100 indices were examined in detail for the HTGA method.

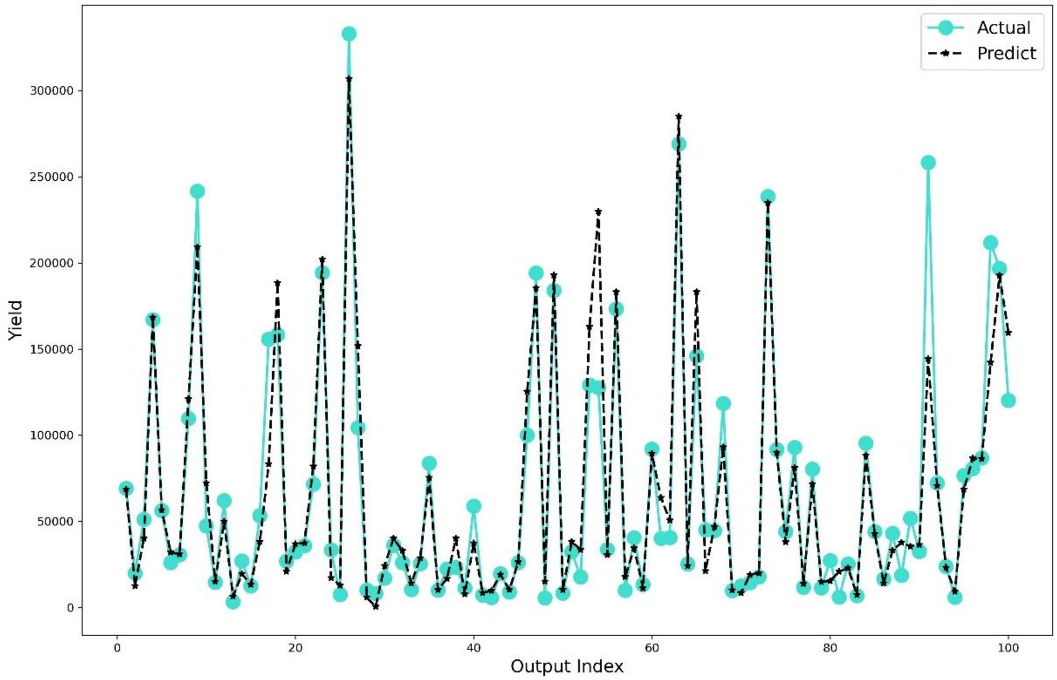

Figure 3. Crop yield estimation with HTGA in 100 scale

In the SVM method, regression analysis was performed using 3 different kernels. Since the Poly core value results were higher than the RBF and Linear core values, only the Poly core results were shared in the article findings. In general, 91% success rate was achieved in the use of Poly core, while 0.81 and 0.69 R^2 rates were obtained in the use of RBF and Linear cores, respectively. For this reason, Poly core regression analysis results are presented in Figure 4-5-6.

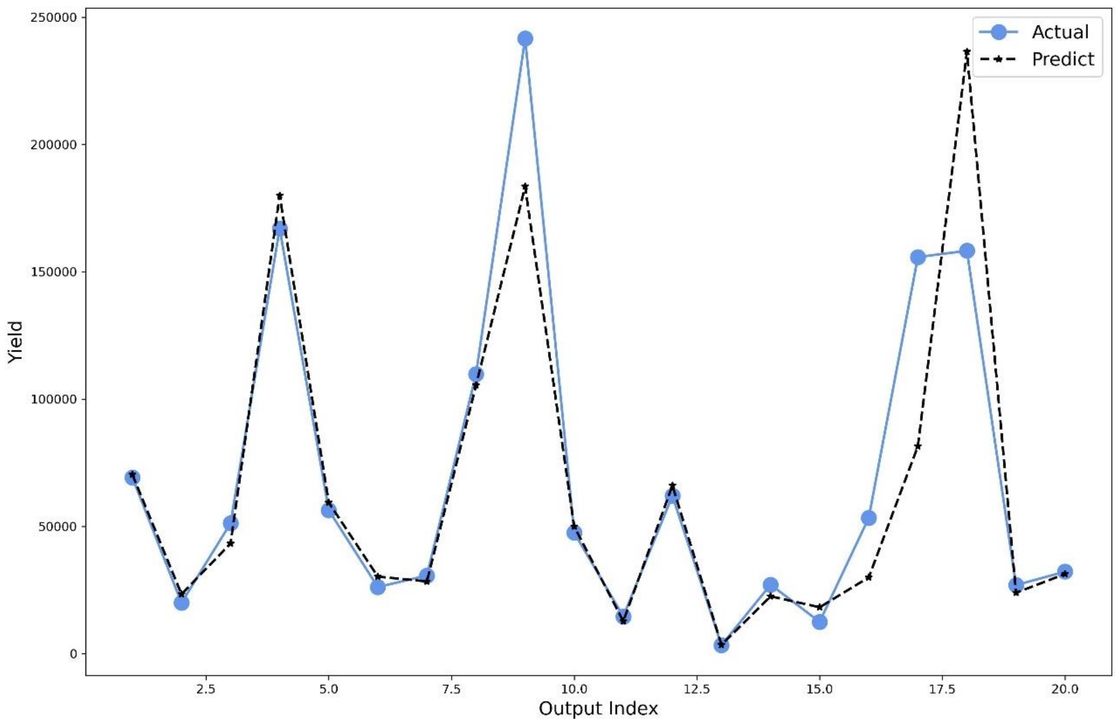

Figure 4. Crop yield estimation with SVM in 20 scale

In Figure 4, there are deviations between 7.5 and 10, and between 15 and 20. The deviations in Figure 4 can also be explained by examining Table 4. In Figure 5, there are deviations in the ranges of 7 and 9, 14 and 17, 28 and 29.

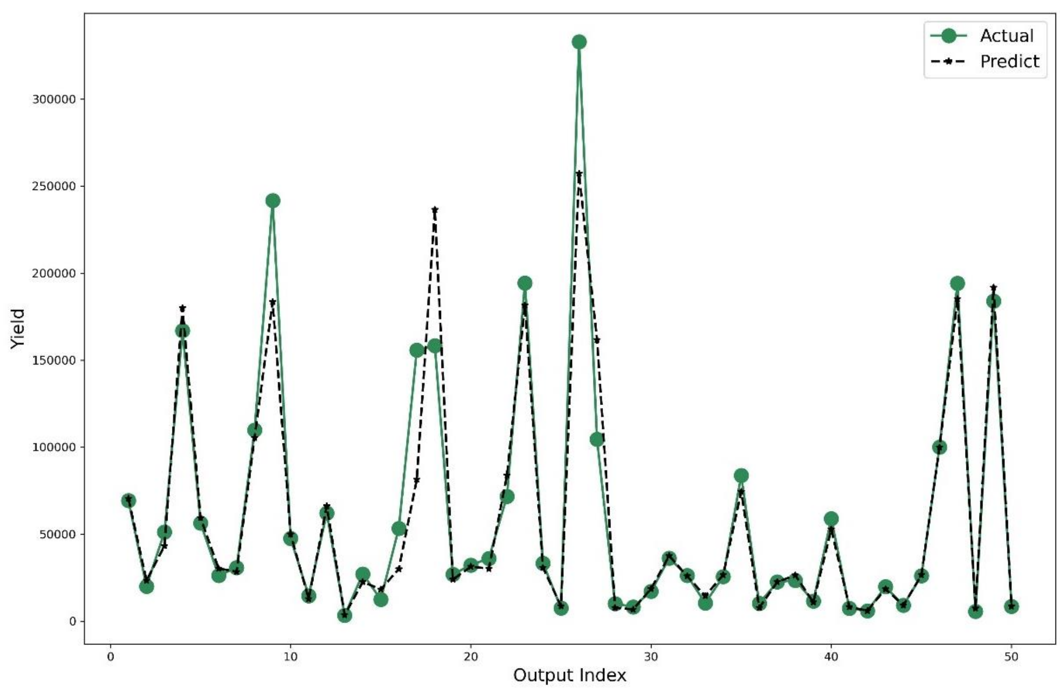

Figure 5. Crop yield estimation with SVM in 50 scale

The main reason for these deviations is the low success rate crops given in Table 4. When Table 4 is examined in detail, the lowest R^2 ratio was obtained in the Plantains crop. The main reason for the deviations in Figures 4, 5 and 6 indicated is the result of low success in the plantains crop.

Table 4. Analysis results of ten different crops with SVM regression method

Algorithm	Crop	R^2 Puan
SVM	Cassava	0.8359
SVM	Corn	0.8322
SVM	Plantains and others	0.5384
SVM	Potatoes	0.7416
SVM	Paddy field	0.8707
SVM	Sorghum	0.8416
SVM	Soybeans	0.7706
SVM	Sweet potatoes	0.8262
SVM	Wheat	0.9183
SVM	Yams	0.8499

In the evaluation of SVM regression analysis, 20, 50 and 100 interval index values were determined. The deviations from the peaks and bottoms at 20 indices, 50 indices, and 100 indices are clarified in Table 4 for the SVM method.

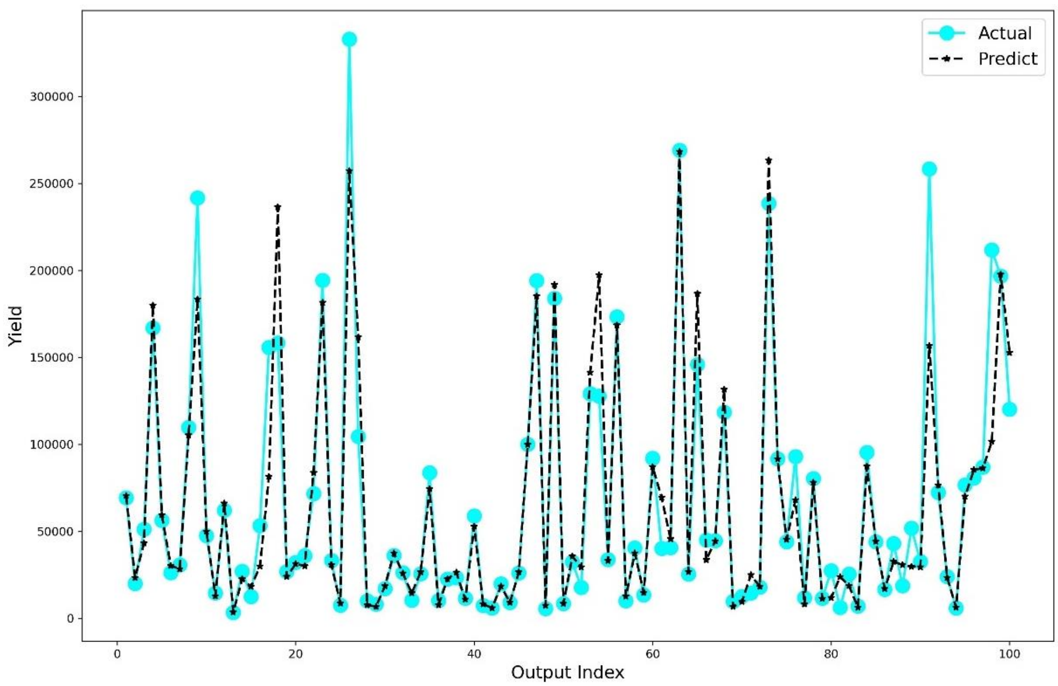

Figure 6. Crop yield estimation with SVM in 100 scale

According to the HTGA algorithm, the best yield estimation was obtained in the order of cassava, potatoes, yams, sweet potatoes, corn, wheat, sorghum, paddy field, plantains and other, soybeans. The best R2 success rates range from 0.91 to 0.44. The product with the best R2 success rate with HTGA is cassava, the lowest success rate was obtained in soybeans product. In SVM, wheat, paddy field, yams, sorghum, cassava, corn, sweet potatoes, soybeans, potatoes, plantains and others were obtained in product order. The R2 success rates obtained with SVM ranged from 0.91 to 0.53. Wheat was the product with the best R2 success rate with SVM, while the lowest success rate was obtained with plantains and others.

4. CONCLUSION

A large data set is used to perform crop yield estimation to assist with the problems of agricultural producers. The used dataset provides information on a wide range of production, both by country and by crop. Input parameters in Table 1 of different crops from 1961 to 2016 in the data set were used in the regression analysis. As a result of the analysis, 94% R^2 success rate was reached with HTGA, while 91% R^2 success rate was reached with SVM. In the SVM method, regression analysis was performed using 3 different kernels. Since the Poly core value results were higher than the RBF and Linear core values, only the Poly core results were shared in the article findings. In addition, within the scope of the article, a 0.75 R^2 score value was obtained in the Linear Regression [34] method, while a 0.68 R^2 score value was obtained in the Logistic Regression [35] method. Since these values are considerably less than the HTGA and SVM Poly core results, they are not detailed. Although the data set used is relatively successful, there are deficiencies in spatial data. It is highly recommended to use not only environmental and abiotic values, but also spatial values in yield estimation [4]. For these reasons, it is aimed to propose a new model from different machine and deep learning methods in further studies.

REFERENCES

[1] Ö. Vanli, B. B. Ustundag, I. Ahmad, I. M. Hernandez-Ochoa, and G. Hoogenboom, "Using crop modeling to evaluate the impacts of climate change on wheat in southeastern turkey," *Environ. Sci. Pollut. Res.*, vol. 26, no. 28, pp. 29397–29408, doi: 10.1007/s11356-019-06061-6 , 2019.

[2] W. Nasim *et al.*, "Future risk assessment by estimating historical heat wave trends with projected heat accumulation using SimCLIM climate model in Pakistan," *Atmos. Res.*, vol. 205, pp. 118–133, doi: https://doi.org/10.1016/j.atmosres.2018.01.009, 2018.

[3] I. Ahmad *et al.*, "Yield Forecasting of Spring Maize Using Remote Sensing and Crop Modeling in Faisalabad-Punjab Pakistan," *J. Indian Soc. Remote Sens.*, vol. 46, no. 10, pp. 1701–1711, doi: 10.1007/s12524-018-0825-8, 2018.

[4] J. Cao *et al.*, "Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine," *Eur. J. Agron.*, vol. 123, p. 126204, doi: https://doi.org/10.1016/j.eja.2020.126204, 2021.

[5] I. FAO, "WFP (2015). The state of food insecurity in the World. Meeting the 2015 international hunger targets: taking stock of uneven progress. Rome, FAO." 2017.

[6] F. Dodds and J. Bartram, *The water, food, energy and climate Nexus: Challenges and an agenda for action*. Routledge, 2016.

[7] R. Patel, "Dataset," , https://www.kaggle.com/patelris/crop-yield-prediction-dataset, 2022.

[8] O. O. Abayomi-Alli, R. Damaševičius, S. Misra, and R. Maskeliūnas, "Cassava disease recognition from low-quality images using enhanced data augmentation model and deep learning," *Expert Syst.*, vol. 38, no. 7, p. e12746, doi: https://doi.org/10.1111/exsy.12746, 2021.

[9] M. B. Cole, M. A. Augustin, M. J. Robertson, and J. M. Manners, "The science of food security," *npj Sci. Food*, vol. 2, no. 1, pp. 1–8, 2018.

[10] T. Stevens and K. Madani, "Future climate impacts on maize farming and food security in Malawi," *Sci. Rep.*, vol. 6, no. 1, pp. 1–14, 2016.

[11] D. Tilman, C. Balzer, J. Hill, and B. L. Befort, "Global food demand and the sustainable intensification of agriculture," *Proc. Natl. Acad. Sci.*, vol. 108, no. 50, pp. 20260–20264, 2011.

[12] A. F. Al-daour and M. O. Al-shawwa, "Classification of Banana Fruits Using Deep Learning," 2020.

[13] K. A. Beals, "Potatoes, nutrition and health," *Am. J. Potato Res.*, vol. 96, no. 2, pp. 102–110, 2019.

[14] Y. Zhu and M. Ghosh, "Temperature control, emission abatement and costs: key EMF 27 results from Environment Canada's Integrated Assessment Model," *Clim. Change*, vol. 123, no. 3, pp. 571–582, 2014.

[15] C. Gonzalo-Martín, A. García-Pedrero, and M. Lillo-Saavedra, "Improving deep learning sorghum head detection through test time augmentation," *Comput. Electron. Agric.*, vol. 186, p. 106179, doi: https://doi.org/10.1016/j.compag.2021.106179, 2021.

[16] C. W. Mundia, S. Secchi, K. Akamani, and G. Wang, "A Regional Comparison of Factors Affecting Global Sorghum Production: The Case of North America, Asia and Africa's Sahel," *Sustainability* , vol. 11, no. 7, doi: 10.3390/su11072135, 2019.

[17] Z. Lin and W. Guo, "Sorghum Panicle Detection and Counting Using Unmanned Aerial System Images and Deep Learning," *Front. Plant Sci.*, vol. 11, doi: 10.3389/fpls.2020.534853, 2020.

[18] R. Battisti, P. C. Sentelhas, and K. J. Boote, "Inter-comparison of performance of soybean crop simulation models and their ensemble in southern Brazil," *F. Crop. Res.*, vol. 200, pp. 28–37, doi: https://doi.org/10.1016/j.fcr.2016.10.004, 2017.

[19] P. C. Sentelhas, R. Battisti, G. M. S. Câmara, J. R. B. Farias, A. C. Hampf, and C. Nendel, "The soybean yield gap in Brazil–magnitude, causes and possible solutions for sustainable production," *J. Agric. Sci.*, vol. 153, no. 8, pp. 1394–1411, 2015.

[20] D. R. Rankine *et al.*, "Parameterizing the FAO AquaCrop Model for Rainfed and Irrigated Field-Grown Sweet Potato," *Agron. J.*, vol. 107, no. 1, pp. 375–387, doi: https://doi.org/10.2134/agronj14.0287, 2015.

[21] N. Verter and V. Bečvářová, "An analysis of yam production in Nigeria," *Acta Univ. Agric. Silvic. Mendelianae Brun.*, vol. 63, no. 2, pp. 659–665, 2015.

[22] Y. Cai *et al.*, "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agric. For. Meteorol.*, vol. 274, pp. 144–159, doi: https://doi.org/10.1016/j.agrformet.2019.03.010, 2019.

[23] R. Alvarez, "Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach," *Eur. J. Agron.*, vol. 30, no. 2, pp. 70–77, doi: https://doi.org/10.1016/j.eja.2008.07.005, 2009.

[24] A. B. Potgieter, D. B. Lobell, G. L. Hammer, D. R. Jordan, P. Davis, and J. Brider, "Yield trends under varying environmental conditions for sorghum and wheat across Australia," *Agric. For. Meteorol.*, vol. 228–229, pp. 276–285, doi: https://doi.org/10.1016/j.agrformet.2016.07.004, 2016.

[25] Y. Li, "Prediction of energy consumption: Variable regression or time series? A case in China," *Energy Sci. Eng.*, vol. 7, no. 6, pp. 2510–2518, doi: 10.1002/ese3.439, 2019.

[26] J. A. Mathieu and F. Aires, "Assessment of the agro-climatic indices to improve crop yield forecasting," *Agric. For. Meteorol.*, vol. 253–254, pp. 15–30, doi: https://doi.org/10.1016/j.agrformet.2018.01.031, 2018.

[27] J. Zhang, Y. Chen, and Z. Zhang, "A remote sensing-based scheme to improve regional crop model calibration at sub-model component level," *Agric. Syst.*, vol. 181, p. 102814, doi: https://doi.org/10.1016/j.agsy.2020.102814, 2020.

[28] G. Azzari, M. Jain, and D. B. Lobell, "Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries," *Remote Sens. Environ.*, vol. 202, pp. 129–141, 2017.

[29] Z. Jin, G. Azzari, M. Burke, S. Aston, and D. B. Lobell, "Mapping smallholder yield heterogeneity at multiple scales in Eastern Africa," *Remote Sens.*, vol. 9, no. 9, p. 931, 2017.

[30] A. Kern *et al.*, "Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices," *Agric. For. Meteorol.*, vol. 260, pp. 300–320, 2018.

[31] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and remotely sensed data," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 858–861, 2015.

[32] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, "Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease," *Int. J. Mach. Learn. Comput.*, vol. 5, pp. 414–419, doi: 10.7763/IJMLC.2015.V5.544, 2015.

[33] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[34] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[35] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.