Доклади на Българската академия на науките
Comptes rendus de l'Académie bulgare des Sciences
Tome 76, No 1, 2023

AGRICULTURAL SCIENCES
Plant breeding

ESTIMATION CHICKPEA SPECIES AND PRODUCTIVITY PER DECARE WITH SYNTHETIC DATA GENERATION METHODS

Kerim Karadağ#, Fırat Keskinbıçak

Received on August 15, 2022
Presented by H. Najdenski, Corresponding Member of BAS, on October 25, 2022

Abstract

Production increase in agriculture depends on some parameters such as improving arable land, activating spraying and irrigation activities. In addition to these, it is known that spraying and seed types have an effect on productivity. Therefore, proper selection of seed types is important. With the developing technology, big data consisting of scientific studies can be recorded digitally and used in the estimation or decision-making process. In this study, chickpea species diversity was made with classification process using machine learning methods by taking advantage of the characteristics of chickpea plant. In addition, productivity per decare was estimated by regression process. Accuracy was preferred as a success criterion for classification, and rmse success criterion was preferred for regression. The dataset was first used raw, and then experiments were made using synthetic data. To generate synthetic data, the synthetic minority oversampling technique method and also the n-shifting mean method proposed in this study were used. When the success rates of the results obtained were compared, the highest success rate was 90.6% in the classification made using only raw data. Likewise, the classification success rate of the dataset using the synthetic data created with the raw data was the highest 100%. For regression, the highest score was 0.17 for raw data and 0.16 for synthetic data. The high performance of the results showed that machine learning algorithms can be used in this field.

Key words: chickpea, synthetic data, machine learning, classification, regression

#Corresponding author.
DOI:10.7546/CRABS.2023.01.16

Introduction. As long as the need for nutrition in living things continues, agriculture will maintain its importance. The pulses group, which constitutes an important part of agriculture, has always taken place in the nutrition of people. Legumes are preferred because they contain high levels of crude protein and are rich in vitamins and minerals. In addition, since its straw contains higher quality protein than grains, it has also been used in animal husbandry [1]. Chickpeas constitute a large part of legumes as a field of use in agriculture. Since chickpea is the most drought resistant plant among legumes, it plays an important role in narrowing the fallow areas by entering crop rotation in arid and semi-arid areas [2]. The effort spent and the high yield are one of the important parameters for agriculture. Although chickpeas are resistant to drought, low rainfall also affects chickpea productivity. In order to increase productivity, researches such as the determination of the right product types, methods of obtaining more products from less area are carried out [3]. Productivity and species identification in agricultural production have a very important place for sustainability. Applications of machine learning (ML) technologies in agriculture, the use of which has increased rapidly in all areas in recent years, has an important place in determining the future of agriculture. ML is a set of values that draws conclusions from the available data, using mathematical and statistical methods, and makes predictions about the unknown with these results.

In their study, Belay et al. [4] developed a chickpea disease detection model using deep learning techniques. Khatri et al. [5] in their study performed wheat seed classification using machine learning methods. Ayele and Tamiru [6] used ML methods to classify chickpea species. Bayar [7] used ML methods and Moore-neighbour tracking algorithms to calculate the sphericity of chickpeas to increase the measurement accuracy of a chickpea pile volume and weight estimator. In their study, Kosmowski and Worku [8] investigated the feasibility of using visible/near infrared hyperspectral data collected with a miniaturized NIR spectrometer to identify barley, chickpea and sorghum cultivars in Ethiopia, and utilized ML algorithms to distinguish cultivars. Kirtis et al. [9] established a plant regeneration protocol for chickpea and then created a model prediction using different ML algorithms. Taheri-Garavand et al. [10] presented a convolutional neural network framework for automatic identification of chickpea cultivars using seed images in the visible spectrum (400–700 nm). Sinha et al. [11] performed a neural network-based estimation of the incidence of DRR, taking into account DRR incidence data from previous reports and weather factors. Azimi et al. [12] compared the performance of traditional ML methods and deep learning methods to classify water stress using chickpea images. Kini et al. [13] investigated the performance of ML algorithms for chickpea classification using some features. Saha and Manickavasagan [14] used machine vision and deep transfer learning to successfully classify different chickpea varieties in their study. Pourdarbani and Sabzi [15] developed a computer vision system to identify three similar chickpea varieties,

adel, arman and azad, using different artificial neural network hybrids. PODOLNY et al. [16] performed a ML analysis of flowering gene expression in the CDC frontier chickpea cultivar. POURDARBANI et al. [17] aimed to present a computer vision system for automatic classification of chickpea varieties. The aim of CANDAN et al. [18] was to study the applicability of infrared spectroscopy combined with ML techniques to evaluate the uptake and distribution of gold nanoparticles and single-walled carbon nanotubes in chickpea.

In this study, the data set obtained from the chickpea species planted using the experimental fields of the Faculty of Agriculture of Harran University was used. First of all, chickpea seed types were estimated by performing the classification process. Then, the regression process was performed and the productivity per decare estimation was made. For classification, Decision Trees (DT), Support Vector Machines (SVM) and k-Nearest neighbour (KNN) methods were used, and accuracy was used as a performance criterion. For regression, DT and SVM methods were used and rmse was used as a performance measure. The dataset was first used raw, and then experiments were made using synthetic data. Synthetic minority oversampling technique (Smote) and n-Shifting Mean (nSM) methods suggested in this study were used as synthetic data. In terms of time and simplicity of application, operations were carried out by choosing not all of the features, but the ones that were effective, and it was observed that the performances increased.

Data and methods. ***Dataset acquisition.*** In this study, the data set obtained from the chickpea species planted using the experimental fields of Harran University Faculty of Agriculture was used. This obtained dataset consisted of 64 samples and features of chickpea seed species. Forty-one of these samples belong to the *Cicer echinospermum* (CEch) chickpea species, and 23 of them belong to the *Cicer reticulatum* (CReti) chickpea species. As a feature, it belongs to chickpea species; seed weight, seed volume (calculated from the height and width of the seed), seed density (calculated from the weight and volume of the seed), seed colour, first flowering days, first pod setting days, maturation days, seed hull form, black spots, plant crown width, chickpea seed type and decare productivity were used. The properties of the dataset are given in Table 1.

Paradigm. In the data set obtained, 64 samples were composed of 41 CEch and 23 CReti chickpea species. First of all, chickpea seed species estimation was made by classification process using 64 samples and then productivity per decare estimation was made by regression process. Since the sample numbers of the chickpea seed types belonging to both groups were not equal, synthetic data were created for the lesser CReti seed type to avoid possible errors in the evaluation, and samples were created and evaluated in numbers close to the CEch sample numbers. The Smote and nSM methods were used to generate synthetic data. The sample numbers obtained using Smote consisted of a total of 87 samples, with 41 CEch and 46 CReti. The sample numbers obtained using NSM consisted of a total of 82 samples, with 41 CEch and 41 CReti. In the last stage, normal

T a b l e 1

The properties of datasets

Features	Feature descriptions	Feature value range	Normalization value range
100 pieces seeds weight (g)	Calculated by weighing	8.62–17.24	0.1–0.9
Seed volume (cm^3)		0.10–0.30	0.1–0.9
Seed density (g/cm^3)		0.39–1.24	0.1–0.9
Number of days to first flowering	Date of planting and the day the first flower appeared	121.00–143.00	0.1–0.9
Number of days of first pod tying	Date of planting and day of first pod tying	132.67–149.50	0.1–0.9
Number of days to maturity	Number of days from the day of planting to the period when all green parts of the plants turn yellow	174.00–223.67	0.1–0.9
Black spots	Yes – No		0.1–0.9
Plant crown width (cm^2)	Diameter was calculated by adding the vertical and horizontal lengths and dividing them by 2.	1014.01–2950.28	0.1–0.9
Parcel productivity (g/parcel)	Calculated by weighing the stems and grains of the harvested plants together	15.61–108.05	0.1–0.9
Plant growth form	Tilted – Semi-tilted		0.1–0.9
Seed colour	Light brown, light brown beige, grey, greyish brown, coffee beige, reddish brown, dark brown, salmon brown, greenish brown		0.1–0.9
Seed husk shape	Prickly, rough, erased		0.1–0.9
Decare productivity (kg/da)	Calculated proportionally from the parcel	52.03–360.17	0.1–0.9
Chickpea seed type	CEch and CReti	0–1	0.1–0.9

and synthetic datasets were normalized and processed. The working flow diagram is given in Fig. 1.

Classification and success criteria. In this study chickpea seed type and yield per decare were tried to be estimated by using ML algorithms. The main purpose of these algorithms is to obtain information from data using computational methods. There are many different methods used, but choosing the right algorithms is important. In this study DT, SVM and KNN methods were used for classification and DT and SVM methods were used for regression. DT creates a graph or tree that uses the branching technique to show each possible outcome of a decision. In a decision tree representation, each internal node tests a feature, each branch corresponds to the result of the parent node, and assigns the class label at the end of each leaf. To classify a specimen, a top-down approach is

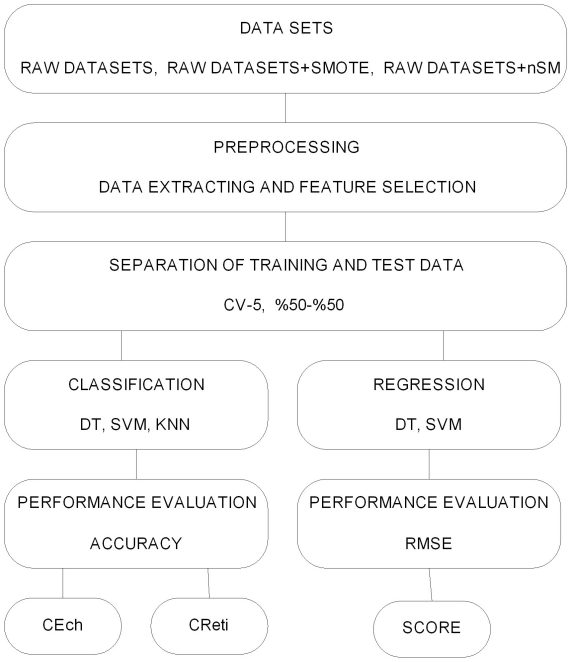

Fig. 1. Flow chart of the study

applied, starting at the root of the tree. For a particular feature or node, the branch that matches the value of that attribute's data point is considered until a leaf is reached or a label is decided [19]. SVM is a method that tries to separate the datasets belonging to different classes in such a way that they are the farthest from each other, and the datasets belonging to the same class are the closest. KNN chooses the nearest neighbour number k for the classification process and classifies the data according to the group membership of k [20]. Data sets were subjected to normalization process before being classified. Eq. 1 was used for normalization.

$$N = 0.8 * \frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})} + 0.1 \tag{1}$$

In order to balance the data sets, the Smote method, which is frequently compared in the literature, and the nSM method proposed in this study were used. In the NSM method, the data set belonging to the small class (clasS) is equalized to the data set belonging to the big class (clasB). First, the ratio between classes was calculated with Eq. 2 and the coefficient "n" was determined. If n is not an integer, it is rounded up.

$$n = \frac{clasB}{clasS} \tag{2}$$

The calculation of this ratio has been chosen because it is compatible with the examples in the space distribution of the samples belonging to the classes. Then, starting from first, the next value was formed by taking the average of the number of ratios. This process continued until the value of the largest class was reached by shifting. In the data set used in the study, as 41 of 64 samples belonging to two classes belonged to clasB and 23 of them belonged to clasS, the n value was 1.78, and since the n value was not an integer, the obtained value was chosen as $n = 2$ by rounding up.

Results and discussion. The dataset used in the study was processed as two groups for classification, chickpea seed types CEch and CReti. For the regression, the amount of product obtained per decare was estimated. Separation of training and test data was first done with cross validation-5 (CV-5), and then the training and test data (%50-training-%50 testing) were evaluated by taking them equally. For the classification process, the dataset was evaluated as raw, then it was used in the datasets created with synthetic data and the accuracy values of the results obtained using machine learning methods were estimated. The three most effective features in distinguishing in classification and regression processes were selected by the relief method. Seed type, black spot and seed husk shape were the most important distinguishing features for all three cases. When the five most important features were examined, in addition to the first three features for raw data, grain weight and plant crown width were determined as important features. In the dataset obtained with Smote, plant growth form and seed colour were other distinguishing features. Other important characteristics for the NSM dataset were grain weight and seed colour. When the raw dataset and the dataset created with Smote are compared, it is observed that 3 features are the same and 2 features are different. When the raw dataset and the dataset created with nSM were compared, it was observed that 4 features were the same and 1 feature was different.

The accuracy achievements obtained in the classification process using all features and only the five features that are effective in distinguishing are given in Table 2.

When the results obtained using all features are compared, the highest success rate was obtained with the KNN method by using the accuracy performance criterion when only the dataset with raw data was processed. It was observed that the success rates increased in the classification process by including synthetic data. The reason for this is due to the increase in the evaluation sensitivity between features as the number of samples increases. When the performances of synthetic data were evaluated among themselves, it was observed that the methods and datasets were superior to each other in the classification process using Smote and nSM methods. For example, it has been observed that the highest success rate for Smote is in the classification made using the DT method, where the training and test data are divided equally. For nSM, on the other hand, it was observed

Table 2

Classification accuracy values of the results obtained

Classification accuracy values of the results obtained using all features						
Classification methods	Raw data		Smote + Raw data		nSM + Raw data	
	%50-%50	CV-5	%50-%50	CV-5	%50-%50	CV-5
DT	75	79.7	100	92	87.8	97.6
SVM	84.4	92.2	97.7	97.7	90.2	93.9
KNN	90.6	93.8	97.7	98.9	95.1	95.1
Classification accuracy values of the results obtained using 5 effective features						
Classification methods	Raw data		Smote + Raw data		nSM + Raw data	
	%50-%50	CV-5	%50-%50	CV-5	%50-%50	CV-5
DT	62.5	100	100	100	100	100
SVM	93.8	96.9	100	100	100	100
KNN	87.5	93.8	95.3	100	95.1	95.1

that the training and test data were separated by CV-5 and classified using the DT method.

When the accuracy values obtained in the classification process using only 5 effective features are compared, the highest success rate was obtained with the DT method by using the accuracy performance criterion when only the dataset with raw data was processed. The success rates obtained in the classification process using synthetic data were 100% for Smote and nSM.

For the regression process, datasets created with synthetic data and raw datasets used the rmse values of the results obtained by machine learning methods were estimated. The results obtained are given in Table 3.

Table 3

Regression rmse values of the results obtained

Regresion rmse values of the results obtained using all features						
Classification methods	Raw data		Smote + Raw data		nSM + Raw data	
	%50-%50	CV-5	%50-%50	CV-5	%50-%50	CV-5
DT	0.17	0.17	0.19	0.16	0.16	0.17
SVM	0.18	0.18	0.18	0.16	0.17	0.18
Regresion rmse values of the results obtained using 5 effective features						
Classification methods	Raw data		Smote + Raw data		nSM + Raw data	
	%50-%50	CV-5	%50-%50	CV-5	%50-%50	CV-5
DT	0.16	0.18	0.17	0.17	0.16	0.17
SVM	0.16	0.19	0.17	0.18	0.16	0.17

When the results obtained using all features are compared, the highest success rate was obtained as 0.17 by using the rmse performance criterion when only the

dataset with raw data was processed. As in the classification process, it was observed that the success rates increased in the regression process by including synthetic data. When the performances of synthetic data were evaluated among themselves, the highest rmse value was 0.16 in the regression process using Smote and nSM methods.

When the rmse values obtained in the regression process using only 5 effective features are compared, the highest success rate was obtained with both methods by using the rmse performance criterion when only the dataset with raw data was processed. The success rates obtained in the regression process using synthetic data were 0.17 for Smote and 0.16 for nSM.

As a result, some experiments were made with the dataset in order to determine the seed type of chickpea and the productivity per decare. The dataset was first taken raw, then synthetic data was created and applied with classification and regression methods. The obtained results showed that the synthetic data was higher than the raw data. When the synthetic data creation methods were evaluated among themselves, it was observed that both methods were successful. In addition to the success parameters, it was seen that the nSM method was more successful in the selection of the most effective features.

Conclusion. In this study, chickpea species estimation and productivity per decare estimation were made by using the data set from Harran University Faculty of Agriculture. The estimation process was made using machine learning methods, DT, SVM and KNN. When the synthetic data, which was created for the evaluation to be more sensitive and acceptable, was used, the results obtained showed that all three methods were successful and could be used for this area. The success rate was 100% for classification in seed species estimation, and the rmse value was 0.16 for productivity per decare. The machine learning algorithms presented here can assist agronomists in any field of agriculture. In addition, this study can provide support to agriculturalists in terms of chickpea seed species estimation and productivity.

Acknowledgement. The authors would like to thank Assoc. Prof. Abdullah Kahriman for helping to acquire data sets.

REFERENCES

[1] Eyidogan F., M. T. Öz (2007) Effect of salinity on antioxidant responses of chickpea seedlings, Acta Physiol. Plant., **29**, 485–493.

[2] Shiferaw B., H. Teklewold (2007) Structure and functioning of chickpea markets in Ethiopia: Evidence based on analyses of value chains linking smallholders and markets, IPMS Working Paper 6. ILRI (International Livestock Research Institute), Nairobi, Kenya. 63 pp.

[3] Erman M., V. Çiftçi, H. G. Geçit (1997) A Research on Relations among the Characters and Path Coefficient Analysis in Chickpea (*Cicer ariefinum* L.), Journal of Agricultural Sciences, **3**(03), 43–46.
[4] Belay A. J., A. O. Salau, M. Ashagrie, M. B. Haile (2022) Development of a chickpea disease detection and classification model using deep learning, Informatics in Medicine Unlocked, 100970.
[5] Khatri A., S. Agrawal, J. M. Chatterjee (2022) Wheat Seed Classification: Utilizing Ensemble Machine Learning Approach, Scientific Programming, `https://doi.org/10.1155/2022/2626868`.
[6] Ayele N. A., H. K. Tamiru (2020) Developing Classification Model for Chickpea Types using Machine Learning Algorithms, International Journal of Innovative Technology and Exploring Engineering, **10**(1), 5–11.
[7] Bayar G. (2021) Increasing measurement accuracy of a chickpea pile weight estimation tool using Moore-neighbor tracing algorithm in sphericity calculation, Journal of Food Measurement and Characterization, **15**(1), 296–308.
[8] Kosmowski F., T. Worku (2018) Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia, PloS one, **13**(3), e0193620.
[9] Kirtis A., M. Aasim, R. Katirci (2022) Application of artificial neural network and machine learning algorithms for modeling the in vitro regeneration of chickpea (*Cicer arietinum* L.), Plant Cell, Tissue and Organ Culture, **150**(1), 141—152.
[10] Taheri-Garavand A., A. Nasiri, D. Fanourakis, S. Fatahi, M. Omid et al. (2021) Automated in situ seed variety identification via deep learning: a case study in chickpea, Plants, **10**(7), 1406.
[11] Sinha R., V. Irulappan, B. S. Patil, P. C. O. Reddy, V. Ramegowda et al. (2021) Low soil moisture predisposes field-grown chickpea plants to dry root rot disease: evidence from simulation modeling and correlation analysis, Scientific reports, **11**(1), 1–12.
[12] Azimi S., T. Kaur, T. K. Gandhi (2020) Water stress identification in chickpea plant shoot images using deep learning. In: 2020 IEEE 17th India Council International Conference (INDICON), 1—7.
[13] Kini A. S., K. V. Prema, S. N. Pai (2021) Intelligent classification model for Indian chickpea. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1847–1852.
[14] Saha D., A. Manickavasagan (2022) Chickpea varietal classification using deep convolutional neural networks with transfer learning, Journal of Food Process Engineering, **45**(3), e13975.
[15] Pourdarbani R., S. Sabzi (2021) Detection of different chickpea varieties using a computer vision system based on computational intelligence, Agricultural Mechanization, **5**(1), 21–31.
[16] Podolny B. S., V. V. Gursky, M. G. Samsonova (2020) A machine-learning analysis of flowering gene expression in the CDC frontier chickpea cultivar, Biophysics, **65**(2), 225–236.
[17] Pourdarbani R., S. Sabzi, D. Kalantari, J. L. Hernández-Hernández, J. I. Arribas (2020) A computer vision system based on majority-voting ensemble neural network for the automatic classification of three chickpea varieties, Foods, **9**(2), 113.

[18] Candan F., Y. Markushin, G. Ozbay (2022) Uptake and Presence Evaluation of Nanoparticles in *Cicer arietinum* L. by Infrared Spectroscopy and Machine Learning Techniques, Plants, **11**(12), 1569.
[19] Das K., R. N. Behera (2017) A survey on machine learning: concept, algorithms and applications, International Journal of Innovative Research in Computer and Communication Engineering, **5**(2), 1301–1309.
[20] Karadağ K. (2020) Semen Quality Estimation with Machine Learning Methods, European Journal of Science and Technology, **18**, 306–311.

Electrical and Electronics Engineering Department
Faculty of Engineering
Harran University
63000 Haliliye, Şanlıurfa, Turkey
e-mail: k.karadag@harran.edu.tr
firatksknbck63@gmail.com