International Journal of Computer Science & Communications
Vol. 9, No. 1, 2024, pp. 1-13

http://ijcsc.ielas.org/index.php/ijcsc
ISSN: 2545-4153

CLASSIFICATION OF PUMPKIN SEEDS USING MACHINE LEARNING TECHNIQUES

Mehr Ali Qasimi
Selçuk University Institute of Science and Technology, BilişimTechnologjıleırMühendisliği, Konya, Türkiye
E-mail: q.mehrali@yahoo.com

Received: 2024-05-13 Accepted: 2024-06-10 Published online: 2024-06-13

Abstract

Accurate and effective seed classification techniques are crucial for seed quality control and crop production optimization, as the need for healthy, high-quality seeds in agriculture continues to rise. With their high oil content and excellent nutritional value, pumpkin seeds are one of the main oil crops. A key component of precision breeding and variety enhancement is the identification and gathering of various pumpkin germplasm resources. Due to its sufficient amounts of protein, fat, carbohydrates, and minerals, pumpkin seeds are eaten raw, roasted, marinated, and sweetened as a dessert around the world. Thus, "UrğüpSivrisi" and "Çerçevelik," the two most significant and high-quality varieties of pumpkin seeds, which are often grown in Turkey's Ügrüp and Karacaören region, were the subject of this study. Nevertheless, measurements of 2500 morphological seeds of both types were achievable through the use of threshold approaches in their gray and binary forms. In order to identify the most effective technique for categorizing pumpkin seed varieties, all the data were modeled using six different machine learning techniques that took morphological features into account: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), kNearest Neighbor (k-NN), Decision Tree Classifier (DT), and Naive Bayes Algorithm (NV). A total of 87.06 percent for LR, 88 percent for SVM, 88.2 percent for RF, 87 percent for k-NN, 87 percent for DT and 86 for NV were the classifiers' accuracy rates. The results have demonstrated that the proposed Random Forest classification Algorithm achieved a satisfactory overall accuracy of 88.2.

Keywords: Machine Learning, Pumpkin seeds, Support Vector Machine Classification, Logistic Regression, Random Forest Classification and other Classification Algorithms.

I. INTRODUCTION

Our work is in the Pumpkin Seeds Dataset, where we classified pumpkin seeds using several machine learning classification algorithms. Cucurbit crops, such as pumpkins, can be planted in a multitude of ways and using a diverse range of resources. It will significantly boost the edible oil industry and raise the added value of pumpkin if high-yielding pumpkin germplasm resources are unearthed. Furthermore, the great nutritional value and significant research value of pumpkin seeds can be attributed to their abundance in proteins, amino acids, and trace elements [1]. The Cucurbitaceae family includes pumpkin. Many cultivars are grown all over the world, but Cucurbita moschata, C. maxima, C. pepo, and C. mixta are the most commercially significant species [2].

The most significant cultivar in Tropical America, namely in Mexico and Central America, is C. moschata. Conversely, C. maxima is thought to be one of the oldest varieties, having been grown in South America from pre-Columbian times [3]. C. pepo, which produces 41612 tons of pumpkin annually, is the most popular species of pumpkin farmed in Turkey [4]. Turkey is a nation located in the Mediterranean region's northern shore. In the Mediterranean region, agriculture is the main user of fresh water. Given its location in the Middle East, Turkey possesses vital and scarce water resources[5]. Even though pumpkin is grown all across the country, Nevşehir, which lies in the middle of Cappadocia—one of the seven places included on the World Heritage List—produces over 31% of all pumpkin imported into Turkey. Pumpkin seeds are said to be beneficial to human health since they provide calcium, potassium, phosphorus, magnesium, iron, and zinc [6]. In addition to 37% of carbohydrates, 35% to 40% of fat, and protein. Pumpkins come in a variety of varieties, and "Ügrgüp Sivrisi" is one of these species. A kind of pumpkin seed known as ÜrgüpSivrisi has a long, white, extremely bright, thin, and nearly indiscernible shell with a pointy tip[3]. "Çerçevelik" is another variety of pumpkin seeds. It's a specific species called "Topak" that's grown in Nevsehir, Karacaören, Turkey.

As technology advanced, machine learning techniques were beginning to be applied globally, including in the agricultural sector. The researcher has used classification algorithms techniques to classify pumpkin seed using six different algorithms of machine learning such as Logistic Regression, Support Vector Machine Algorithm (SVM), Decision Tree, Random Forest, K Nearest Neighbors (KNN) and also Naive Bayes, Each algorithm had different outputs, accuracies and scores. The best algorithm in this study the researcher has implemented and found best accuracy was Random Forest algorithm for classification of pumpkin seeds dataset. From 24 distinct cotton types that were collected at various stages of growth using NB, J48 (C4.5), and MLP machine learning techniques. Tenfold cross-validation of the model was done to assess accuracy. The results indicated that the accuracy of the J48 (C4.5) and MLP models was 98.78 percent, while the accuracy of the NB model was 94.22 percent. Nevertheless, it was found that MLP required more processing time than J48; as a result, J48 was regarded as the most appropriate model. The results of the SVM and MLP methods showed that the accuracy rates of these models were 86.8% and 94.5 percent, respectively. With the data utilized in this investigation, it was discovered that the model built using MLP was more effective. Most machine learning studies involving agricultural products included algorithms like SVM, MLP, and k-NN. This study used gray form and binary form threshold methods to turn hitherto unstudied properties of two different pumpkin seed kinds, "ÜrgüpSivrisi" and "Çerçevelik," into measurable forms, which were then modeled using five different machine learning techniques. The goal of this research was to pave the way for future investigations that would use these models [9].

II. MATERIAL and METHOD

The study employed a dataset of pumpkin seeds that were acquired and processed in accordance with prior guidelines. The main research process of this study included five parts: data collection, data analysis, construction of classification model, transfer study and visualization [10]. In summary, the study employed two varieties of pumpkin seeds from Turkish cultivars, "Nevşehirçerçevelisi" and "Ürgüpsivrisi," (as shown in figure 1) both of which are members of the C. pepo species.

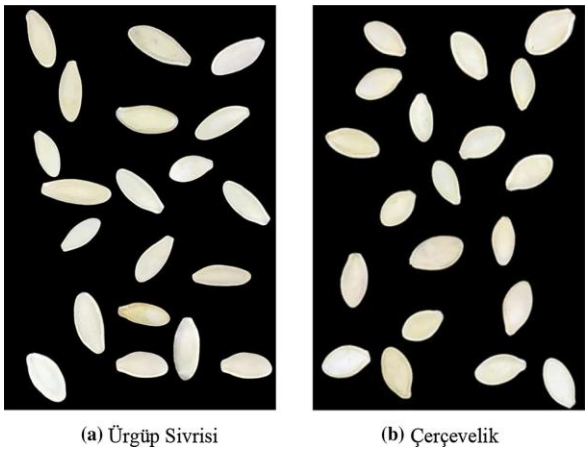


Figure 1: Core type of pumpkin seed varieties [9].

This dataset had 2500 rows or data and Thirteen features total—one was used as an output feature which contained categorical data andthe remaining twelve characteristics were used as input features which contained numerical data. The dataset data were consist of 52% "Çerçevelisi" and 48% "Ürgüp sivrisi," data (shown in figure 2).

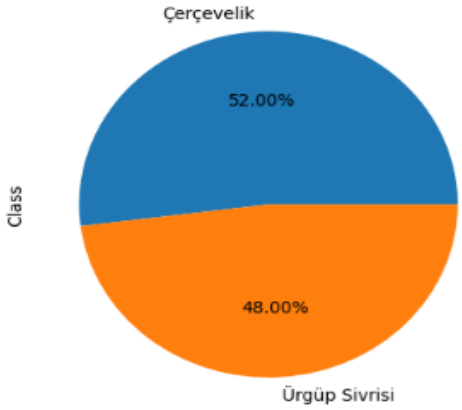


Figure 2: showing the percentage of each pumpkin seeds class

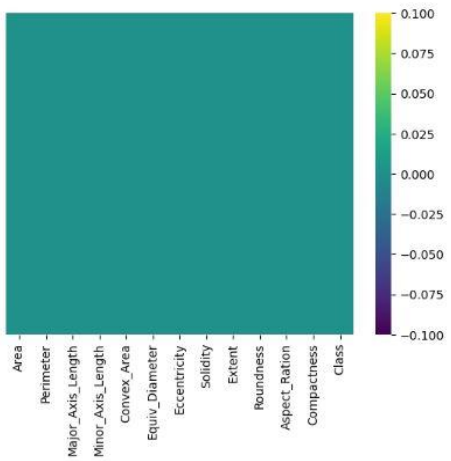
Description of each of the attributes in the dataset is shown in Table 1.

Table 1
The most effective morphological features and explanations used in feature extraction

No	Name	Explanation
1	Area (A)	It gave the number of pixels within the borders of a pumpkin seed
2	Perimeter (p)	It gave the circumference in pixels of a pumpkin seed
3	Major Axis Length (Maj.AL)	It gave the circumference in pixels of a pumpkin seed
4	Minor Axis Length (Min.AL)	It gave the small axis distance of a pumpkin seed
5	Eccentricity (e)	It gave the eccentricity of a pumpkin seed
6	Convex Area (CA)	It gave the number of pixels of the smallest convex shell at the region formed by the pumpkin seed
7	Extent (E)	It returned the ratio of a pumpkin seed area to the bounding box pixels
8	Equiv Diameter (ED)	It was formed by multiplying the area of the pumpkin seed by four and dividing by the number pi, and taking the square root
9	Compactness (C)	It proportioned the area of the pumpkin seed relative to the area of the circle with the same circumference
10	Solidity (s)	It considered the convex and convex condition of the pumpkin seeds
11	Roundness (r)	It measured the ovality of pumpkin seeds without considering its distortion of the edges
12	Aspect Ratio (AR)	It gave the aspect ratio of the pumpkin seeds

In this experiment, we used the Machine Learning implemented in Python, with the aim to build a model that will be applied further to make prediction and analysis of a test dataset. At first we have analyzed the data to found the null values (as show in figure 3), duplicate values, unnamed columns, normalization, correlation between features (as shown in figure 4) and also we have plotted confusion matrix of each algorithm. We have implemented six different classification algorithms on this dataset and every algorithms had different outputs and accuracy, Random forest had the most high accuracy among six classification algorithms.

International Journal of Computer Science & Communications
Vol. 9, No. 1, 2024, pp. 1-13
http://ijcsc.ielas.org/index.php/ijcsc
ISSN: 2545-4153



Figiure 3: Counting null values of each feature

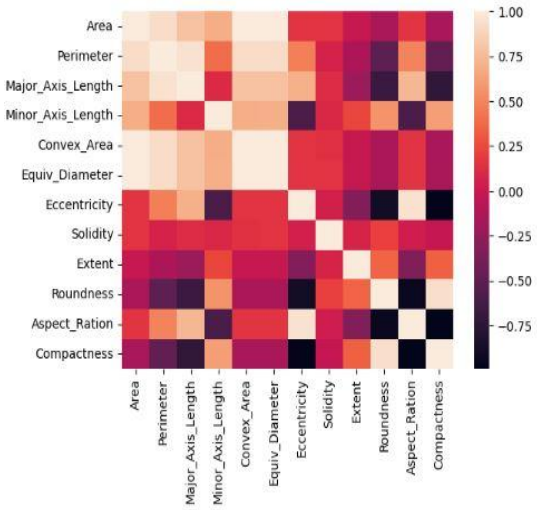


Figure 4: Visualizing the correlation matrix of the numerical columns

Turkey produced over 365 thousand tons of pumpkins annually, compared to the 13 to 15.5 million tons produced worldwide [11]. The confectionery pumpkin types were included in all the pumpkin kinds produced in Turkey. The districts of Cappadocia, Ürgüp in Nevşehir and Tekirdağ, and Kırklareli in Thrace were the growing grounds for pumpkin seeds used in confections [12]. The dataset included two distinct varieties of commercial pumpkin seeds, identified as Çerçevelik and Ürgüp Sivrisi, respectively, from the Karacaören and Ürgüp regions of Nevşehir. In our nation, the most widely cultivated pumpkin seeds are Çerçevelik and Ürgüp sivrisi. Subgroups of these two types comprise other cultivars that are cultivated.

It's crucial to classify these two varieties of pumpkin seeds, particularly for the seed industry. Table-2 displayed the distributions of various pumpkin seeds within the sample.

Table 2

Distribution of Ürgüp Sivrisi and Çerçevelik type pumpkin seeds in the dataset

AD	PIECE
Çerçevelik	1300
Ürgüp Sivrisi	1200
TOTAL	2500

The average, standard deviation, maximum, and minimum statistical values of the two types of pumpkin seeds were mentioned in Table 3. A general inference could be made about pumpkin seeds from the table[9].

Table 3

The statistical distribution of Ürgüp Sivrisi and Çerçevelik pumpkin seed varieties

NO	FEATURE	MIN	MEAN	MAX	Std.dev
1	Area (A)	47,939.0	80,658.220	136,574.0	13,664.510
2	Perimeter (p)	868.485	1,130.279	1,559.45	109.256
3	Major axis length (Maj.AL)	320.844	456.601	661.911	56.235
4	Minor axis length (Min.AL)	152.171	225.794	305.818	23.297
5	Eccentricity (e)	0.492	0.860	0.948	0.045
6	Convex area (CA)	48,366.0	81,508.084	138,384.0	13,764.092
7	Extent (E)	0.467	0.693	0.829	0.060
8	Equiv. diameter (ED)	247.058	319.334	417.002	26.891
9	Solidity (s)	0.918	0.989	0.994	0.003
10	Aspect ratio (AR)	1.148	2.041	3.144	0.315
11	Roundness (r)	0.554	0.791	0.939	0.055
12	Compactness (C)	0.560	0.704	0.904	0.053

Models for numerous problems were made possible by machine learning. A few performance metrics were used to ensure that the model used on the classifiers worked well. By using these performance measures, the algorithm's success was evaluated instead of the model's. Assuming that the values of a confusion matrix of size n x n connected to a classifier represented the number of classes, it displayed both the estimated and actual classification results [13]. A confusion matrix was the name of the performance set that was used to assess how well the categorization models worked. Figure 5 had discussed the confusion matrix's structure, which was applied to the study's classification of pumpkin seeds. There are four parameters in the confusion matrix, as seen in Figure 5. These are defined as;

Tp: Çerçevelik was estimated, and the result obtained was Çerçevelik, Fp: Ürgüp Sivrisi was estimated and the result obtained was Çerçevelik, Fn: Çerçevelik was predicted, and the result obtained was Ürgüp Sivrisi, Tn: Ürgüp Sivrisi was estimated and the resultobtained was Ürgüp Sivrisi.

Confusion matrix		Predicted	
		Cercevelik	Urgup Sivrisi
Actual	Cercevelik	*tp*	*fp*
	Urgup sivrisi	*fn*	*tn*

Figure 5: The confusion matrix used in the classification of pumpkin seed grains

Development of models

The most crucial aspect of machine learning techniques was the model generation process, wherein various approaches took into account the nature of the problem and the properties of the dataset. These techniques comprised regression, clustering, and classification. Additionally, the study's dataset comprised two classes of goal variables and numerical input variables. These characteristics of the dataset led to the decision that using classification techniques in the study was appropriate. By examining the patterns in the training set of data, classification algorithms trained the model. This allowed it to classify the test data in a highly accurate way that had never been seen before. In this study, the Python 3 programming language was used to model pumpkin seed kernels using the following classifiers: Naive Bais (NV), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and K Nearest Neighbors (k-NN). These are machine learning techniques that are commonly applied to challenges involving classification. Other machine learning techniques have also been tested in addition to these, however the techniques employed in the study produced more significant classification results than the others [9].

Logistic regression (LR)

To describe the fitness between dependent and independent variables with the fewest number of variables possible, a model utilizing logistic regression analysis was developed [14]. The identification of which individual belonging to which population was accomplished by drawing a regression curve [15]. Eq. 1 was utilized to compute the curve. The optimization process during classification in this study was carried out using the Newton approach with the assistance of LR.

$$\Phi(Z) = 1/E^{-Z} \quad (ONE)$$

Support vector machine (SVM)

The classical Support Vector Machines (SVMs) are non-probabilistic, binary classifiers that aim at sending the dividing hyper plane which separates both classes of the training set with the maximum margin. Then, the predicted label of a new, unseen data point is determined based on which side of the hyper plane it falls. In order to statistically distinguish two classes on the multidimensional plane, support vector machines predicted an appropriate hyper plane function [16]. The sigmoid hyper plane function was found in this investigation, and the gamma value of "1/feature number" was accepted. SVMs were initially designed for binary (two-class) problems. When dealing with multiple classes, an appropriate multi-class method is needed. Techniques such as 'one against one' and the 'one against the rest' are in frequent use for the multi-class problems [17].

Random forest (RF)

The combination of predictors that obtained the greatest vote from all tree estimators was taken into consideration by the random forest classifier to classify a large number of random samples that were sampled independently of the input vector [17]. The study's forest, which included 100 trees, was found to have that many trees. Nonetheless, the entropy was used to calculate the acquisition of information. The random forest classifier consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector. The random forest classifier used for this study consists of using randomly selected features or a combination of features at each node to grow a tree. Bagging, a method to generate a training dataset by randomly drawing with replacement N examples, where N is the size of the original training set was used for each feature/feature combination selected. Any examples (pixels) are classified by taking the most popular voted class from all the tree predictors in the forest. Design of a decision tree required the choice of an attribute selection measure and a pruning method. There are many approaches to the selection of attributes used for decision tree induction and most approaches assign a quality measure directly to the attribute. The most frequently used attribute selection measures in decision tree induction are the Information Gain Ratio criterion and the Gini Index. The random forest classifier uses the Gini Index as an attribute selection measure, which measures the impurity of an attribute with respect to the classes [17].

K Nearest neighbor (k-NN)

The nearest k points in the same space with each piece of data in the training set were found using the k-NN, or k-nearest neighbor, technique, which typically took the Euclidean distance into account. Based on the Euclidean distance values, the test data that entered the model belonged to the same class as the lowest one [18]. The k value in this investigation was five. In K nearest neighbors ("KNN"), the objective is to classify a given new, unseen data point by looking at K given data points in the training set, which are closest in input or feature space. Therefore, in order to find the K nearest neighbors of the new data point, we have to use a distance metric.

Naive Bayes (NV)

Naive Bayes classifiers require a small number of data points to be trained, can deal with high-dimensional data points, and are fast and highly scalable. Moreover, they are a popular model for applications such as spam filtering, text categorization, and automatic medical diagnosis used this algorithm to combine factors to evaluate the trust value and calculate the final quantitative trust of the Agricultural product [18].

Decision Tree (DT)

A normal tree includes root, branches and leaves. The same structure is followed in Decision Tree. It contains root node, branches, and leaf nodes. Testing an attribute is on every internal node, the outcome of the test is on branch and class label as a result is on leaf node. A root node is parent of all nodes and as the name suggests it is the topmost node in Tree [19]. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continues value). As decision trees mimic the human level thinking so it’s so simple to grab the data and make some good interpretations. The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf. Decision Tree is similar to the human decision-making process and so that it is easy to understand. It can solve in both situations whether one has discrete or continuous data as input [20]. The example of Decision Tree is as follow [19].

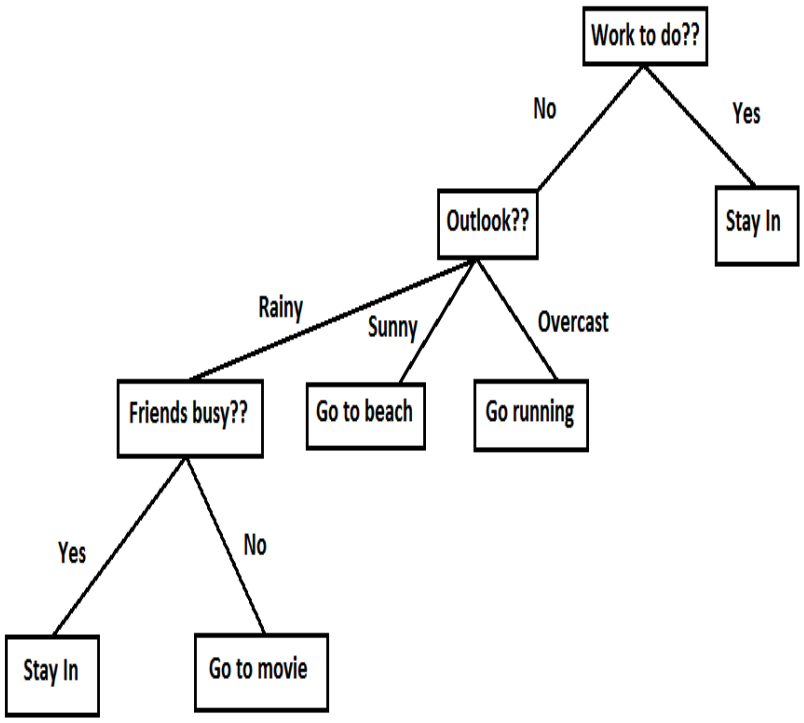


Figure 6: Example of Decision Tree on what to do when different situations occur in weather.

Table 4

Confusion matrix values of classifiers

Algorithms	**Confusion Matrix**
LR	355 32 65 298
SVM	358 29 61 302
KNN	354 33 67 269
Decision Tree	347 40 61 302
Naive Bayes	347 40 68 295
Random Forest	354 33 55 308

Table 5

Performance measurement results of the models

Measure	LR	SVM	KNN	DT	NVB	RF
Accuracy	87	88	86.6	86.5	86	88.26
precision	85	85	84	85	84	87
Recall	92	93	91	90	90	91
F1 Score	88	89	88	87	87	89

III. CONCLUSION

Two varieties of pumpkin seeds that have been tested on a total of 2500 samples were frequently bred in Turkey for trade purposes. Table 5 displayed the assessment findings that were taken from Table 4's confusion matrix. Table 5 illustrates that the success rate of the models produced exceeded 85% in terms of accuracy. The Random Forest model's accuracy value of 88.26 percent demonstrated its superiority over other models. For the Random forest model, the success rate was 91% when taking the precision value into account. This evaluation found that compared to the other models, the Random forest model produced results that were more accurate. However, the results indicated that the Random forest model was more effective than the other models in terms of specificity and F1-score evaluations. The study's pumpkin seeds, Çerçevelik and Ürgüp Sivrisi, were among the commercial items for which the research findings were anticipated to improve the quality of production. Twelve morphological characteristics of pumpkin seeds were found in this investigation. Six distinct machine learning techniques (LR, SVM, RF, NVB, DT, and k-NN) were utilized to assess the identified features for the two distinct pumpkin seed classes (Çerçevelik and Ürgüp Sivrisi). The models yielded accuracy ratings of 87.9 percent, 88.26 percent, 86.5 percent, 86.6 percent, and 86 percent, in that order. But according to the literature analysis, the model would perform better if morphological characteristics like texture, color, and professional judgment were included to the independent ones.

Simultaneously, the goal of this study was to serve as a source of inspiration, particularly for the other exporting textile and culinary items. In the near future, quality measurements have been carried out in consultation with experts. It was anticipated that the automation of the unmanned factories, whose classifications depended on the quality of products, would expand with the entrance of Industry 5.0 into daily life.

REFERENCES

[1] M. Batool, U. Roobab, U. Farooq, U. Farooq, S. Selim, and S. A. Ibrahim, "Nutritional Value, Phytochemical Potential, and Therapeutic Benefits of Pumpkin (Cucurbita sp.)," vol. 11, no. 11, 2022, doi: 10.3390/plants11111394.

[2] L.-G. Saucedo-Herna, María Jesús, Herrero-Martínez, José Manuel, Ramis-Ramos, Guillermo, Jorge-Rodríguez, Elisa, Simó-Alfonso, Ernesto F., "Classification of pumpkin seed oils according to their species and genetic variety by attenuated total reflection fourier-transform infrared spectroscopy," vol. 59, no. 8, pp. 4125-4129, 2011, doi: 10.1021/jf104278g.

[3] N. Aktaş, T. Uzlaşır, and Y. E. Tunçil, "Pre-roasting treatments significantly impact thermal and kinetic characteristics of pumpkin seed oil," vol. 669, pp. 109-115, 2018, doi: 10.1016/j.tca.2018.09.012.

[4] I. İli, T. Biyokütle, P. Ve, E. Eşdeğeri, E. Kuş, and I. Üniversitesi, Iğdır İli Tarımsal Biyokütle Potansiyeli ve Enerji Eşdeğeri. [Online]. Available:https://www.researchgate.net/publication/319702317.

[5] Y. Kuslu, U. Sahin, F. M. Kiziloglu, and S. Memis, Fruit yield and quality, and irrigation water use efficiency of summer squash drip-irrigated with different irrigation quantities in a semi-arid agricultural area, vol. 13, no. 11, pp. 2518-2526, 2014, doi: 10.1016/S2095-3119(13)60611-5.

[6] D. Peričin, L. Radulović, S. Trivić, and E. Dimić, "Evaluation of solubility of pumpkin seed globulins by response surface method," vol. 84, no. 4, pp. 591-584, 2008, doi: 10.1016/j.jfoodeng.2007.07.002.

[7] B. Demir, I. Eski, Z. A. Kus, and S. ErcislI, "Prediction of physical parameters of pumpkin seeds using neural network," vol. 45, no. 1, pp. 22-27, ErcislI, Sezai, doi: 10.15835/nbha45110429.

[8] K. S. Jamuna, S. Karpagavalli, M. S. Vijaya, P. Revathi, S. Gokilavani, and E. Madhiya, "ACE 2010 - 2010 International Conference on Advances in Computer Engineering," Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques, pp. 312-315, 2010, doi: 10.1109/ACE.2010.71.

[9] M. Koklu, S. Sarigil, and O. Ozbek, "The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.)," vol. 68, no. 7, 2713-2726, 2021, doi: 10.1007/s10722-021-01226-0.

[10] X. Li et al., "Classification of multi-year and multi-variety pumpkin seeds using hyperspectral imaging technology and three-dimensional convolutional neural network," vol. 19, no. 1, 2023, doi: 10.1186/s13007-023-01057-3.

[11] J. E., W. E., G. C., and C. RodriguezSao, "Weed and Pest Control - Conventional and New Challenges," Companion Planting and Insect Pest Control 2013, doi: 10.5772/55044.

[12] K. Kökten, M. Kaplan, S. Seydoşoğlu, H. Tutar, and H. Tutar, "Bingöl Koşullarında Bazı Burçak (Vicia ervilia (L.) Willd) Genotiplerinin Tohum Verimi ve Kalite Özelliklerinin Belirlenmesi," vol. 56, no. 1, pp. 31-40, 2019 doi: 10.20289/zfdergi.409921.

[13] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix*," 1971 Doi: https://doi.org/10.3758/BF03213026

[14] M. J. L. F. Cruyff, U. Böckenholt, P. G. M. van der Heijden, and L. E. Frank, "A Review of Regression Procedures for Randomized Response Data, Including Univariate and Multivariate Logistic Regression, the Proportional Odds Model and Item Response Model, and Self-Protective Responses," vol. 34, pp. 287-315, 2016, doi: 10.1016/bs.host.2016.01.016.

[15] B. Kalantar, B. Pradhan, S. Amir Naghibi, A. Motevalli, and A. Motevalli, "Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN)," vol. 9, no. 1, pp. 49-69, 2018 Doi: 10.1080/19475705.2017.1407368.

[16] Kavzoğlu T and Ç. I, pp. 73-82, 2010 Doi: https://doi.org/10.17475/kastorman.289762

[17] M. Pal, "Random forest classifier for remote sensing classification," vol. 26, no. 1, pp. 217-222, 2005, doi: 10.1080/01431160412331269698.

[18] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," vol. 4, no. 3, pp. 161-175, 2018, doi: 10.1016/j.dcan.2017.10.002.

[19] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74-78, 2018/10// 2018 Doi: 10.26438/ijcse/v6i10.7478.

[20] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," vol. 6, no. 10, pp. 74-78, 2018 Doi: 10.26438/ijcse/v6i10.7478.