A CNN-Based Hybrid Approach to Classification of Raisin Grains

Esra Kavalcı Yılmaz[1,*], Taha Oğuz [2] and Kemal Adem [3]

[1,3]*Dept. of Computer Engineering, Sivas University of Science and Technology, Turkey*
[2]*Dept of Metallurgical and Materials Engineering, Sivas University of Science and Technology, Turkey*

[*]*(esra.kavalci@sivas.edu.tr) Email of the corresponding author*

Abstract **–** Raisin Grains, which are an important food source thanks to their rich carbohydrates, potassium and iron contents, are also beneficial for many health problems. When the classification of the type and quality of raisin grains is done with traditional methods, it can be easily affected by the psychological and physiological condition of the specialist who performs the operation. For this reason, it is important to realize systems based on machine learning methods in order to obtain more successful and reliable results. In this study, we focused on CNN-based hybrid machine learning methods for the classification of 2 different types of raisin grains. Evaluations were made using 5 different machine learning methods: KNN, Ridge Classifier, XGBoost, SVC and LDA. In order to evaluate the CNN-based hybrid model, raisin grains were first classified by the classical method using these machine learning methods. Then, classification operations were performed using CNN + Machine Learning methods and compared with the results obtained with classical machine learning. As a result of the study, when the results obtained with the hybrid model proposed in the study were compared with the results obtained with the classical methods, it was seen that the hybrid model increased the success compared to the classical machine learning methods.

Keywords – Raisin Grains, Deep Learning, Machine Learning, Classification

I. INTRODUCTION

Raisin grains are a nutritious food item containing potassium, fiber and iron as well as rich carbohydrate content. Although it contains a high amount of sugar, it has beneficial effects on human health. Although larger studies are needed, current studies show that raisin grains may be beneficial for oral health, colon and intestinal function, cancer and Alzheimer's diseases. In addition, raisin grains provide a better diet quality by reducing appetite [1]. When the production activities of this product, which is stated to have many benefits, are examined, Turkey is one of the most suitable regions for grape cultivation due to its geographical location and climate characteristics. As a result of this, it is seen that it is in the first place in the world grape production. Grapes, which have various usage areas, are used in fields such as table, drying, wine, etc. [2].

There are many traditional method applications for the evaluation of foods. However, these applications are both time-consuming and costly. In addition, since it is based on human power, it is easily affected by human conditions such as fatigue, psychological mood, and individual experience. This makes operations inconsistent and inefficient. Considering all these reasons, it was necessary to develop artificial intelligence-based applications in order to obtain more efficient and successful results in a shorter time and with less cost. There are not many studies in the literature on

raisin grains classification, but a few recent examples are briefly mentioned below.

Karimi et al. [3] proposed a system for quality and purity classification from raisin grains images. In their study, a total of 146 features were created using 4 different methods: Gray Level Co-occurrence Matrix (GLCM), Gray Level Run-Length Matrix (GLRM), Local Binary Pattern (LBP) and Principal Component Analysis (PCA). Then, using these features, classification processes were carried out with Artificial Neural Networks (ANN) and Support Vector Machines (SVM) algorithms. When the results are compared, it has been observed that the SVM model performs classification more efficiently and accurately (averagely 92.71%). Çınar et al. [4] took raisin grains images and extracted 7 morphological features from these images and thus created a new data set. Then, using this data set they created, they evaluated the classification performances of various machine learning techniques (Logistic Regression-LR, Multilayer Perceptron-MLP and SVM). In the study, the best performance was obtained with SVM with an accuracy value of 86.44%. Kılıçarslan [5] proposed a hybrid model using Rotation Forest (RO) and Stacked Auto Encoder (SAE) methods to classify raisin grains correctly. With the model he proposed, he achieved the highest success values (91.50%) in the literature.

As in raisin grains, machine learning methods are used for quality and classification evaluation in many different foods. Köklü et al. [6] obtained a total of 898 images of 7 different date fruit species. Afterwards, they created a data set by extracting 34 features with image processing techniques. In the last stage of the study, firstly, classification operations were performed using LR and ANN. Then, they worked with the new model they obtained by combining ANN and LR models. They achieved 92.8% success with this hybrid model (ANN-LR). As a result of the study, they concluded that machine learning methods can be successfully applied in the classification of date species. Tütüncü et al. [7], on the other hand, tried to determine whether the mushrooms were poisonous by using their physical and morphological features. In this study, which they carried out using 22 features and 4 different machine learning methods (Naive Bayes, Decision Tree, Vector Machine and AdaBoost), they reached the conclusion that it is possible to distinguish whether the mushrooms are poisonous or not from their physical properties. In another study, Köklü et al. [8] carried out classification processes using the morphological features obtained from pumpkin seed images. They used 5 different machine learning methods (LR, MLP, SVM, Random Forest (RF) and k-Nearest Neighbour (k-NN)) for classification operations. As an outcome of the study, the highest accuracy was calculated as 88.64% with SVM.

In the remainder of the article, Section-2 Material provides information about the dataset used in this study. Section-3 Methods provides information about the deep learning and machine learning algorithms used in the study. Section-4 experimental results and discussion shows the results of all analyzes and discussions. The conclusion is given in Section-5 and summarizes the entire work.

II. Materials and Method

In this part of the study, the data set, CNN model and machine learning methods used in the study are explained in detail.

A. Raisin Dataset

The dataset used in the study was created by Çınar et al. in 2020 and uploaded to the UCI Machine Learning Repository [4]. The dataset consists of a total of 900 data, including 450 Besni and 450 Keçimen species. After the images of the raisin grains were taken, it was created by removing 7 morphological features. In the last case, the dataset consists of 900 samples and 7 features. The attributes of the data set and the descriptions of these attributes are shared in Table 1.

Table 1. Dataset attributes and descriptions

Attribute	Explanation
Area	# pixels within the borders of the raisin
Perimeter	Calculated from the distance between raisin borders and the pixels of surrounding
MajorAxisLength	Long axis length of the raisin
MinorAxisLength	Short axis length of the raisin

Eccentricity	The measure of the ellipse that has the same moments as the raisins
ConvexArea	Count the pixels of the smallest convex raisin skin region
Extent	Ratio of raisin region to total pixels in bounding box
Class	Besni-Kecimen

B. *Convolutional Neural Networks (CNN)*

CNN is a neural network that works similarly to neurons in the human brain. It has an architecture consisting of at least one convolution layer, normalization layer, activation layer, pooling layer, fully connected layer and softmax layer [9]. In the convolution layer, features are extracted, passed through the activation function and stored in a feature map. The dimensionality of the feature map is then reduced by applying a pooling layer to prevent overfitting. In the next step, fully connected layer, back propagation is used and training is performed. Although the CNN model was first developed for computer vision applications, it is frequently used in one dimension thanks to its success [10]. The one-dimensional CNN model used in this study is shared in Figure 1.

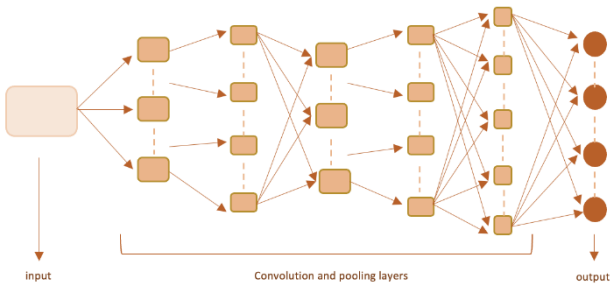

Fig. 1 CNN model used in the study

C. *Ridge Classifier*

If the independent variables are directly related to each other, results inconsistent with reality can be obtained with the linear classification model. In such cases, the Ridge Classifier is used to prevent overfitting of linear regression. Ridge Regression offers more reliable results by reducing the standard error and adding a deviation certain degree to the obtained result [11]. Ridge Classifier also shows high performance in processing small datasets [12].

D. *Extreme Gradient Boosting (XGBoost)*

It is an algorithm based on Gradient Boosting Decision Trees (GBDT) that can be used in both regression and classification operations. The algorithm basic principle is to evaluate the errors of many weak classifiers and then to reduce the error rate by iteration. In short, it is a situation where many unsuccessful learners come together to form a strong learner with higher performance and computational speed [13].

E. *K-Nearest Neighbours (KNN)*

K-Nearest Neighbor Algorithm is a classification method that performs classification operations by taking the k nearest data as a reference to classify data of unknown class. Each new sample is processed by comparing it with k existing samples using the specified distance function [14]. The distance formula used plays an active role in classification success. The most commonly used theorem in measuring neighborhood distance is Euclid. Manhattan, Minkowski, and Mahalanobis are also other commonly used distance measurements for distance calculation. When evaluated in the general context, Euclidean is used in univariate data sets when features have equal weights, Manhattan is used for high-dimensional data, and Mahalanobis is used in multidimensional spaces when there is a relation between the features of the data set. Minkowski is a generalization of Euclid and Manhattan. [15].

It is seen that the KNN algorithm provides better performance when compared to other algorithms using the Euclidean distance criterion such as the Bayesian algorithm [16].

F. *Support Vector Classifier (SVC)*

Support Vector Classifier, one of the supervised learning methods, is a method that learns from examples and analyses data at the same time. This method was first described by Vapnik et al. suggested by. It took the final version of Cortes and Vapnik, which are currently used, in 1995. In 1997, Vapnik et al. Its use has increased with the developing information technology [17]. The SVC model is shown in Figure 2. As can be seen in the figure, classification processes are carried out after the decision boundary determines the most appropriate boundary decision for each class.

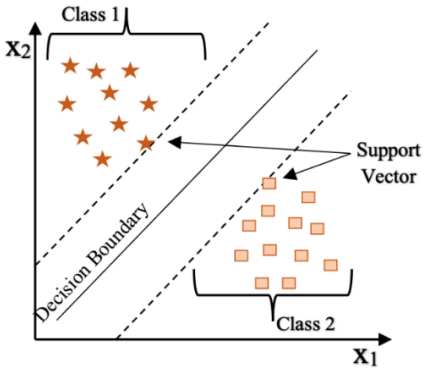

Fig. 2 SVC model

One of the most important features affecting SVM performance is the kernel parameter. The use of different kernels may cause different classification of SVM. Linear, Polynomial, Sigmoid, and Radial Basis are some of the most commonly used kernel types. [18]. The functions used in the use of these kernels are shared in Table 2.

Table 2. Kernel function formulas

Kernel Types	Function
Linear	$F(\alpha x,\alpha y)=\alpha x.\alpha y$
Polynomial	$F(\alpha x,\alpha y)=((\alpha x.\alpha y)+c)^{d}$
Sigmoid	$F(\alpha x,\alpha y)=\tanh((\alpha x.\alpha y)+c)$
Radial Based	$F(\alpha x,\alpha y)=\exp(-\lvert\alpha x-\alpha y\rvert^{2})$

This method is generally used in analysis of regression, face, behavior or pattern recognition, image, video or text classification, data mining, quality control and economics, genetics, biology or bioinformatics applications [19].

G. *Linear Discriminant Analysis (LDA)*

Discriminant analysis is one of the methods that perform classification operations using statistical features. It uses training data to find distinctive features. Discriminant functions define the boundaries between different classes and the estimation space. The classifier distinguishes between classes based on the prediction data [20]. Feature vectors are provided by the classifier. In this analysis, the classifier is faster to train and simpler to implement. One of the discriminant analysis type, "linear discriminant analysis" linear separation function assumes multivariate normal density for each group with general estimation of the covariance matrix. The discriminant function is obtained in such a way as to maximize the distinction between objects. For this purpose,

$$(X^{(-1)}\,A-\beta I)^{v}=0 \quad (1)$$

equation is examined. Here X is the within-group square sum matrix; A is the square matrix between groups. Solving equation (1) means finding the eigenvalues and eigenvectors of W−1B. The β values obtained from here are the eigenvalues; V stands for eigenvectors [21].

III. RESULTS

In the first stage of the study, 5 different machine learning methods, namely KNN, SVM, LDA, XGBoost and Ridge Classifier, were used and the results were compared. In order for these algorithms to be used to provide the best performance, certain hyper parameters have been adjusted using grid search. The parameters examined for each algorithm are shared in Table 3.

Table 3. Hyperparameters used in machine learning algorithms

ML Algorithm	Parameters	Value Range
KNN	n_neighbors	5,7,9,11,13,15
	algorithm	'auto', 'kd_tree', 'brute', 'ball_tree'

	weights	,‘distance’ ,‘uniform’
SVC	kernel	‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’
LDA	solver	‘svd’, ‘lsqr’, ‘eigen’
XGB	booster	‘gbtree’, ‘gblinear’, ‘dart’
	gamma	0,0.1,0.2
Ridge	solver	‘auto’, ‘cholesky’, ‘sparse_cg’, ‘sag’, ‘svd’ ‘saga’, ‘lbfgs’, ‘lsqr’

For KNN in the parameters examined, n_neighbors=9, algorithm=auto, weights=uniform; kernel=linear for SVC; solver=svd for LDA; It was seen that the best results were obtained with booster=gbtree, gamma=0 for XGB and solvent=auto for Ridge. Afterwards, classification processes were carried out using these determined parameters.

During the classification processes, cross validation was applied so that k=5. The data set was separated into 5 seperations and 4 seperations were used for training, while 1 seperation was used for testing.

In the second part of the study, first of all, training was carried out with the CNN model. The details of the CNN model used are as shown in Figure 3.

Layer (type)	Output Shape	Param #
Conv1D	(None, 5, 32)	128
BatchNormalization	(None, 5, 32)	128
Conv1D	(None, 3, 32)	3104
BatchNormalization	(None, 3, 32)	128
Conv1D	(None, 1, 64)	6208
BatchNormalization	(None, 1, 64)	256
Flatten	(None, 64)	0
Dense	(None, 1)	65

Fig. 3 CNN model used in the study

Then, the outputs obtained from the flatten layer of the CNN model were transmitted as input to the machine learning algorithms and classification processes were carried out. The block diagram of the model realized in the study is shared in Figure 4.

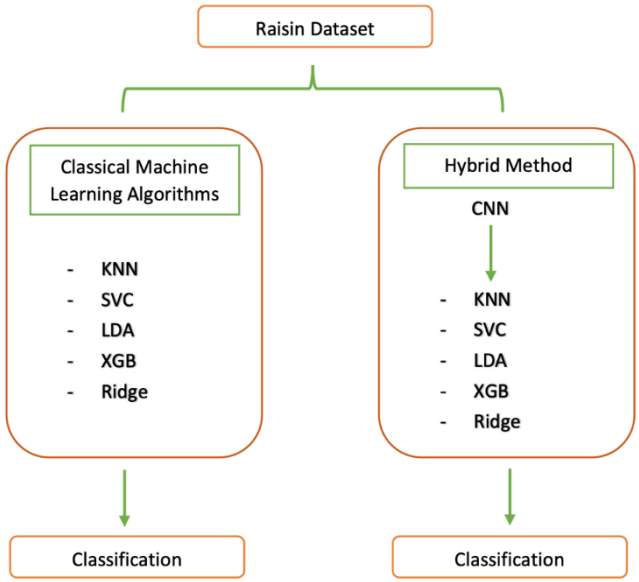

Fig. 4 Block diagram of the realized model

During the training of the CNN model, 'binary_crossentropy' was used as the loss function and 'adam' was used as the optimizer. The CNN model was trained separately for epoch 50 and 100 values, and these results were analyzed comparatively.

The Accuracy and F1 Score values obtained as a result of all these processes are shared in Table 4 in a comparative way.

When the table is examined in detail, it is seen that the model, which is recommended to evaluate machine learning algorithms separately, successfully increases the accuracy value. It is seen that the highest accuracy value was calculated as 87.5% with the CNN (100Epoch) + SVC hybrid model.

IV. DISCUSSION

In the study, the number of inputs was increased by using CNN, and thus, it is seen that the hybrid study in the form of CNN + ML increased the success. In the hybrid model, two different processes, 50 and 100 epochs, were carried out. When evaluated within the scope of epoch, the best results were obtained in studies with 100 epochs. Among the machine learning methods, it was determined that the best result was obtained with SVC. In the CNN + SVC study, the effect of the kernel variable on the result was examined and the best result was obtained with the 'linear kernel'. For this reason, classification processes were carried out using the 'linear kernel' in the study.

Table 4. Study results

Methods	Number of Input	ML Algorithms	Accuracy	F1 Score
None	7	KNN	83.22	83.61
		SVC	85.44	83.43
		LDA	85.77	86.39
		XGB	**86.88**	**86.22**
		Ridge	85.77	87.05
CNN (50 epoch)	64	KNN	86.8	88.75
		SVC	**87.36**	**87.71**
		LDA	87.3	87.2
		XGB	85.69	85.88
		Ridge	**87.36**	**87.71**
CNN (100 epoch)	64	KNN	85.97	88.37
		SVC	**87.5**	**88.37**
		LDA	87.08	87.2
		XGB	85.83	86.54

Ridge	86.94	87.71

V. CONCLUSION

Accurate classification of agricultural products increases productivity in production, while enabling experts to use their time more efficiently. In this study, a comparative analysis with classical machine learning algorithms was carried out using a hybrid model such as CNN + Machine Learning Algorithms in order to increase the accuracy in the classification of agricultural products. In order to increase the input size in the study, firstly the data was trained in the CNN model, and then the outputs of the flatten layer of the CNN model (64 features) were transmitted as input to the machine learning algorithms. As a result of the study, it was observed that the success rates generally increased after using CNN. As a result, it can be said that the hybrid method used in the study can be used as a decision support system in the classification of agricultural products.

It is thought that the success can be increased by better training the deep learning model by using a higher dimensional data set in future studies. In addition, it is thought that higher success can be achieved by using different deep learning models.

REFERENCES

[1] Olmo-Cunillera, A., Escobar-Avello, D., Pérez, A. J., Marhuenda-Muñoz, M., Lamuela-Raventós, R. M., & Vallverdú-Queralt, A. (2019). Is Eating Raisins Healthy? Nutrients, 12(1), 54. MDPI AG. Retrieved from http://dx.doi.org/10.3390/nu12010054.

[2] Yeğenoğlu, E. D., Aydın, Ş., Arık, C., Gevrekçi, Y. & Aşık, M. (2016). Üzümde Çeşitliliğin Belirlenmesinde Morfolojik Farklılıkların Kullanılması. Soma Meslek Yüksekokulu Teknik Bilimler Dergisi, 2 (22), 13-20. Retrieved from https://dergipark.org.tr/tr/pub/somatbd/issue/26739/281565.

[3] Navab Karimi, Ramin Ranjbarzadeh Kondrood, Tohid Alizadeh,An intelligent system for quality measurement of Golden Bleached raisins using two comparative machine learning algorithms,Measurement,Volume 107,2017,Pages 68-76,ISSN 0263-2241,https://doi.org/10.1016/j.measurement.2017.05.009.

[4] Çınar, İ., Koklu, M. & Taşdemir, P. D. Ş. (2020). Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods. Gazi Mühendislik Bilimleri Dergisi, 6 (3), 200-209. Retrieved from https://dergipark.org.tr/en/pub/gmbd/issue/58697/777134.

[5] Kılıçarslan, S. (2022). Kurum Üzüm Tanelerinin Sınıflandırılması İçin Hibrit Bir Yaklaşım. Mühendislik Bilimleri ve Araştırmaları Dergisi, 4 (1), 62-71. DOI: 10.46387/bjesr.1084590.

[6] Koklu, M., Kursun, R., Taspinar, Y. S., and Cinar, I. (2021) "Classification of Date Fruits into Genetic Varieties Using Image Analysis", Mathematical Problems in Engineering, vol. 2021, pp. 1- 13, 2021.

[7] K. Tutuncu, I. Cinar, R. Kursun and M. Koklu, "Edible and Poisonous Mushrooms Classification by Machine Learning Algorithms," 2022 11th Mediterranean Conference on Embedded Computing (MECO), 2022, pp. 1-4, doi: 10.1109/MECO55406.2022.9797212

[8] Koklu, M., Sarigil, S. & Ozbek, O. The use of machine learning methods in classification of pumpkin seeds (Cucurbita pepo L.). Genet Resour Crop Evol 68, 2713–2726 (2021). https://doi.org/10.1007/s10722-021-01226-0.

[9] Serhat Kilicarslan, Kemal Adem, Mete Celik, Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network, Medical Hypotheses, Volume 137, 2020, 109577, ISSN 0306-9877, https://doi.org/10.1016/j.mehy.2020.109577.

[10] F. C. Zegarra, J. Vargas-Machuca and A. M. Coronado, "Comparison of CNN and CNN-LSTM Architectures for Tool Wear Estimation," 2021 IEEE Engineering International Research Conference (EIRCON), 2021, pp. 1-4, doi: 10.1109/EIRCON52903.2021.9613659.

[11] D. Li, Q. Ge, P. Zhang, Y. Xing, Z. Yang and W. Nai, "Ridge Regression with High Order Truncated Gradient Descent Method," 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 2020, pp. 252-255, doi: 10.1109/IHMSC49165.2020.00063.

[12] A. Kumar, P. K. Das, R. K. Mallick and P. Nayak, "Islanding Detection of Micro-grid using Ridge Regression," 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), Keonjhar, India, 2020, pp. 1-5, doi: 10.1109/CISPSSE49931.2020.9212236.

[13] Ramazan Katırcı, Esra Kavalcı Yılmaz, Oğuz Kaynar, Metin Zontul, Automated evaluation of Cr-III coated parts using Mask RCNN and ML methods, Surface and Coatings Technology, Volume 422, 2021, 127571, ISSN 0257-8972, https://doi.org/10.1016/j.surfcoat.2021.127571.

[14] O. Altay, "Performance of different KNN models in prediction english language readability," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, pp. 1-5, doi: 10.1109/ICMI55296.2022.9873670.

[15] V. Kalra, I. Kashyap and H. Kaur, "Effect of Distance Measures on K-Nearest Neighbour Classifier," 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2022, pp. 1-7, doi: 10.1109/ICCSEA54677.2022.9936314.

[16] Aman Kataria and M.D. Singh, "A Review of Data Classification Using K-Nearest Neighbour Algorithm", International Journal of Emerging Technology and Advance Engineering., vol. 3, no. 6, 2013.
[17] Bilişik, M.T. 2011. Destek Vektör Makinesi, Çoklu Regresyon ve Doğrusal Olmayan Programlama ile Perakendecilik Sektöründe Gelir Yönetimi İçin Dinamik Fiyatlandırma. İstanbul Kültür Üniversitesi, XI. Üretim Araştırmaları Sempozyumu, 23-24 Haziran, İstanbul, 785-799.
[18] W. An-na, Z. Yue, H. Yun-tao and L. I. Yun-lu, "A novel construction of SVM compound kernel function," 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM), Harbin, China, 2010, pp. 1462-1465, doi: 10.1109/ICLSIM.2010.5461210
[19] http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/2109 (Last Access: 16.01.2023)
[20] Cao, J. and Sanders, D.B., (1996). Multivariate discriminant analysis of the electromyographic interference pattern: statistical approach to discrimination among controls, myopathies and neuropathies. Medical and Biological Engineering and Computing Vol.34, 5, pp:369-374.
[21] Tatsuoka, M.M., (1971). Multivariate Analysis: Techniques for Educational and Psychological Research. New York: John Wiley & Sons, Inc., pp:159-162.