Araştırma Makalesi – Original Article

A Technological Perspective on Traditional Medicine: Classification of Plant Species with Machine Learning*

Geleneksel Tıbba Teknolojik Bir Bakış: Bitki Türlerinin Makine Öğrenimi ile Sınıflandırılması

Fatma Söğüt[i], Bora Reşitoğlu[ii], Evrim Ersin Kangal[iii]

[i]Ass.Prof., Mersin University, Vocational School of Medical Services, Department of First Aid and Emergency https://orcid.org/0000-0002-1108-8947

[ii]Ass.Prof., Mersin University, Vocational School of Medical Services, Department of First Aid and Emergency https://orcid.org/0000-0003-2703-6831

[iii]Assoc.Prof., Mersin University, Computer Technology and Information Systems, School of Applied Technology and Management of Erdemli, https://orcid.org/0000-0001-5906-3143

ABSTRACT

Objective: The aim of this study is to determine the morphological characteristics of any plant; that is, to classify it with the method of image processing and machine learning by defining it with features such as leaf shape, color or odor.

Method: In this study, plant images obtained from an open access database called kaggle were used as a source for machine learning. After the image learning process, the leaf images of the plants were classified by the Convolutional Neural Network (CNN) method. To verify that the system was working, 100 images of leaves and flowers were taken for each of two different plants, and the number of statistical data was increased to 700 with the ImageData Generator algorithm.

Results: It was concluded that the system identified plants with 97% accuracy. The performance of the machine learning algorithm can also be understood from the confusion matrix. In the method followed in this study, diagonal elements 98 and 79 of the confusion matrix were obtained. This indicates that the method we applied is statistically significant.

Conclusion: Thanks to the algorithm used in this study, the identification of plants used in traditional and complementary medicine could be made with an accuracy of 97%. With this algorithm, plants containing harmful chemicals can be identified to the user and their use can be prevented. Transferring the algorithm from the computer system to mobile applications by covering more plant varieties will be a guide for future studies.

Keywords: Medicinal plants, Machine learning, Image processing, Traditional medicine.

ÖZET

Amaç: Bu çalışmanın amacı herhangi bir bitkiyi morfolojik özellikleri; yani yaprak biçimi, rengi ya da kokusu gibi özellikleriyle tanımlayarak, görüntü işleme ve makine öğrenmesi yöntemiyle sınıflandırmaktır.

Yöntem: Bu çalışmada kaggle adlı açık erişimli veri tabanından elde edilen bitki görüntüleri makine öğrenimi için kaynak olarak kullanıldı. Görüntü öğrenme işlemi yapıldıktan sonra bitkilerin yaprak görüntüleri Evrişimli Sinir Ağı (CNN) yöntemi ile sınıflandırıldı. Sisteminin çalıştığını doğrulamak için iki farklı bitkinin her biri için 100 adet yaprak ve çiçek görüntüsü alınarak Görüntü Veri Üreteci algoritması ile eldeki istatiksel verinin sayısı 700'e arttırıldı.

Bulgular: Sisteminin bitkileri % 97 doğrulukla tanımladığı sonucuna varılmıştır. Makine öğrenimi algoritmasının performansı karışıklık matrisinden de anlaşılabilir. Bu çalışmada izlenen yöntemde karışıklık matrisinin köşegenel elemanları 98 ve 79 elde edilmiştir. Bu da uyguladığımız metodun istatistiksel olarak anlamlı olduğunu belirtmektedir.

Sonuç: Bu çalışmada kullanılan algoritma sayesinde geleneksel ve tamamlayıcı tıpta kullanılan bitkilerin kimliklemesi %97 doğrulukla yapılabilmiştir. Bu algoritma ile içeriğinde zararlı kimyasalların olduğu bitkiler kullanıcısına tanımlanabilir ve kullanmaları engellenebilir. Algoritmanın daha fazla bitki çeşidini kapsayarak bilgisayar sisteminden mobil uygulamalara aktarılması ileriki çalışmalar için yol gösterici olacaktır.

Anahtar Kelimeler: Tıbbi bitkiler, Makine öğrenmesi, Görüntü işleme, Geleneksel tıp.

Mersin Üniversitesi Tıp Fakültesi Lokman Hekim Tıp Tarihi ve Folklorik Tıp Dergisi 2023;13(3):764-774
DOI: 10.31020/mutftd.1339794
e-ISSN: 1309-8004
Geliş Tarihi – Received: 08 August 2023; Kabul Tarihi - Accepted: 13 September 2023
İletişim - Correspondence Author: Fatma Söğüt <fatmasogut@gmail.com>

Introduction

The World Health Organization (WHO) explains traditional medicine as the full of knowledge, proficiencies and habits that can be described or not based on theories, beliefs and traditions of different cultures for the prevention, diagnosis, healing as well as treatment of physical and mental diseases. When auxiliary methods are applied to the patient more to modern medicine during the treatment, it is called "complementary medicine", and the application of other approaches instead of modern medicine is called "alternative medicine". In recent years, the definition of 'traditional and complementary medicine' has come to the fore, by concluding that there is no alternative to medicine, but only an alternative to treatment, based on the definition of WHO.[1,2]

Despite the prevalence and power of modern medical science, a large part of humanity either cannot access these applications or prefer not to apply them. Many people prefer practices called traditional, alternative and complementary medicine. The differentiation in the phenomenon of disease with globalization, the idea that naturalness is free from side effects, the inadequacy of medicine in the treatment of some diseases, the side effects of drugs, etc. For these reasons, people prefer these applications. In addition, the rate of use of these methods has been increasing over the years.[3]

The drugs used in the methods followed for the treatment of some medical diseases such as diabetes and cancer that human beings are exposed to cause difficulties in other ways.[4] Medicinal plants continue to be an alternative to modern drugs due to expensive health costs, especially for developing countries.[5]

Our knowledge of the components of medicinal plants plays a very important role in their usage in the medical field.[6] Although its roots date back to ages ago, traditional medicine is still one of the indispensable health methods of our lives in many parts of the world.[7] Due to the content of these plants, which are used in the treatment and prevention of various diseases, some are beneficial to human health and some are harmful. Although the use of medicinal plants is made consciously in some societies, depending on past experiences, it is known that in some places this awareness has not yet settled in social life.[8]

Today, people in various parts of the world still collect and use medicinal plants according to the season. WHO found that these countries solve 80% of their health problems with this method.[4,9] When using these plants, it is necessary to know the chemical substances in their content, because these plants collected from wild areas may contain all kinds of substances in their roots under the ground. In developed societies, these contents are also determined by using technological devices. However, identifying these plants collected in natural areas away from the laboratory, knowing how to use them and recording them also ensures the transfer of cultural heritage for generations. In this regard, importance should be given to the creation of documents for traditional medicinal Turkish plants. Therefore, the aim of this study is to develop a technological method that enables the identification of medicinal plants and gives information about their medical content. For the development of this method, the knowledge of the plants initially determined from random regions can be used. The results obtained with these plants will allow for the identification of a more comprehensive number of medicinal plant species for the future studies.

Material and Method

Machine learning (ML) is one of the knowledge derivation methods provided by data mining so that we can work on subjects that we do not always have the opportunity to work with direct method. For this, computers are trained through large datasets to produce a prediction for random events, and when the sensitivity of this prediction is at a satisfactory level, new information can be obtained for the event encountered for the

first time. In this context, artificial intelligence (AI), deep learning (DL), artificial neural- networks (ANN) and ML have become concepts that we often encounter in our daily lives.

In the supervised methods of ML algorithms, a connection is created between the data used as input and the output data with the help of classification and regression algorithms.[10] In these methods, some properties of the plant are used as input for the identification of the medicinal plant. The flower or scent of the plant is some of the features used to identify the plant. For the steps followed as methodology, it can be listed as taking the image of the plant, processing the image and applying the neural network algorithm. Identification and classification of the plant are the two most important steps to ultimately obtain a satisfactory degree of accuracy.

In this study, approximately 200 images belonging to two plant classes were processed to determine the characteristics. CNN based on color leaf shape was used to classify plants. Among the different types of neural networks the CNN is one of the most used method compared to other traditional methods, such as ANN. CNNs use rather little pre-processing as to other image classification algorithms.[11] The classification is adopted to identify the image and to give knowledge covered in the image. A CNN can be expressed more than on convolutional layers. It is a set of entirely related, grouped convolutional layers and pooling. The convolution layer in the sets takes the convolutional operations as the input. In the convolutional process, there are three components as the input image, kernel, output image. The aspect detector is called as the mask or kernel.[12] There are masks as 3x3, 5x5, and 7x7. 32 filters are made for the first con1, 64 filters are made for the second con2 and the 128 filters for the last con3. CNN identifies the images with higher accuracy. This process can be simply illustrated as in the ***Figure 1*.**

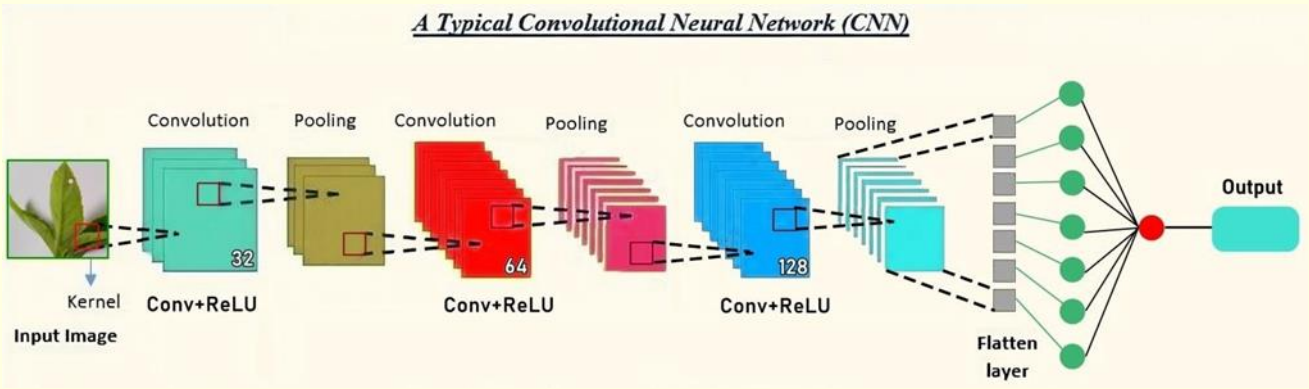

Figure 1. The process for the CNN method.

For other methods of image processing from plant leaves, detailed studies can be consulted.[13-17] The method followed here is also classified as supervised learning. In the model, the PC learns the dataset of images consisting of various leaves that are labeled and named. Features such as shape and color are taken as the basis in the leaf cluster. First, convolution filters are applied for feature extraction. Then, the image is reduced in size with max pooling and this process is repeated three times. Then, a color-based characterization of the image is made.

Dataset

There is no standard database for most medicinal plants. Many countries have their own databases. For this reason, the dataset used here was accessed from the open source Kaggle database, on July 28, 2023.18 The data belongs to two different plant species and consists of 24-bit color images. The corresponding image data requires approximately 17 GB space, and these images are of 3120x4160 pixels with a horizontal and vertical

resolution of 72 dpi. Initially, the algorithm developed via Google CodeLabs, which communicates with Kaggle, and than was put to the test for several image samples. With the performance achieved as a result of the test, these data was downloaded from the source site to a local PC for ML. Later, algorithm training was done with the help of training data. The PC provides the classification of test data in line with the knowledge gained by the network from this training. This process has been individually tested by each author on three different PCs.

The names and characteristics of the plants are given in ***Table 1.***

Table 1. Images and features of plants.

Plant Name	Image	Features
Bohera		Bohera of the family Combretaceae, reaching aheight of 15 to 22 m and a diameter of about one metre or more. The fruits are eatable, but are usually and widely utilized in particular medical intentions. The kernel can be eaten but has narcotic properties. The pulp of the plant is used in treating diarrhoea and leprosy. The oil extracted from theseeds is helpful as a hair tonic.[19]
Ochna		Species of this genus can be obtained in the Old World Tropics, especially in Africa. In Kenya root extracts are taken against kidney and stomachproblems.[20]

Analysis

A total of 200 images for two plants were used from Kaggle but later to obtain a better statistics, the image dataset was augmented in real time using the *ImageDataGenerator* class and the number of data was increased from 200 to 700, see ***Figure 2.***[21]

Image enlargement assists the CNN method. Image data augmentation creates modified versions of the data in the image set. Image data augmentation applies only to the train dataset and artificially expands the size of the training dataset, not the validation or test dataset.

In ***Figure 3***, we can see the visualization of some input images taken from the kaggle dataset by computer. Images are taken from different point of view and distances.

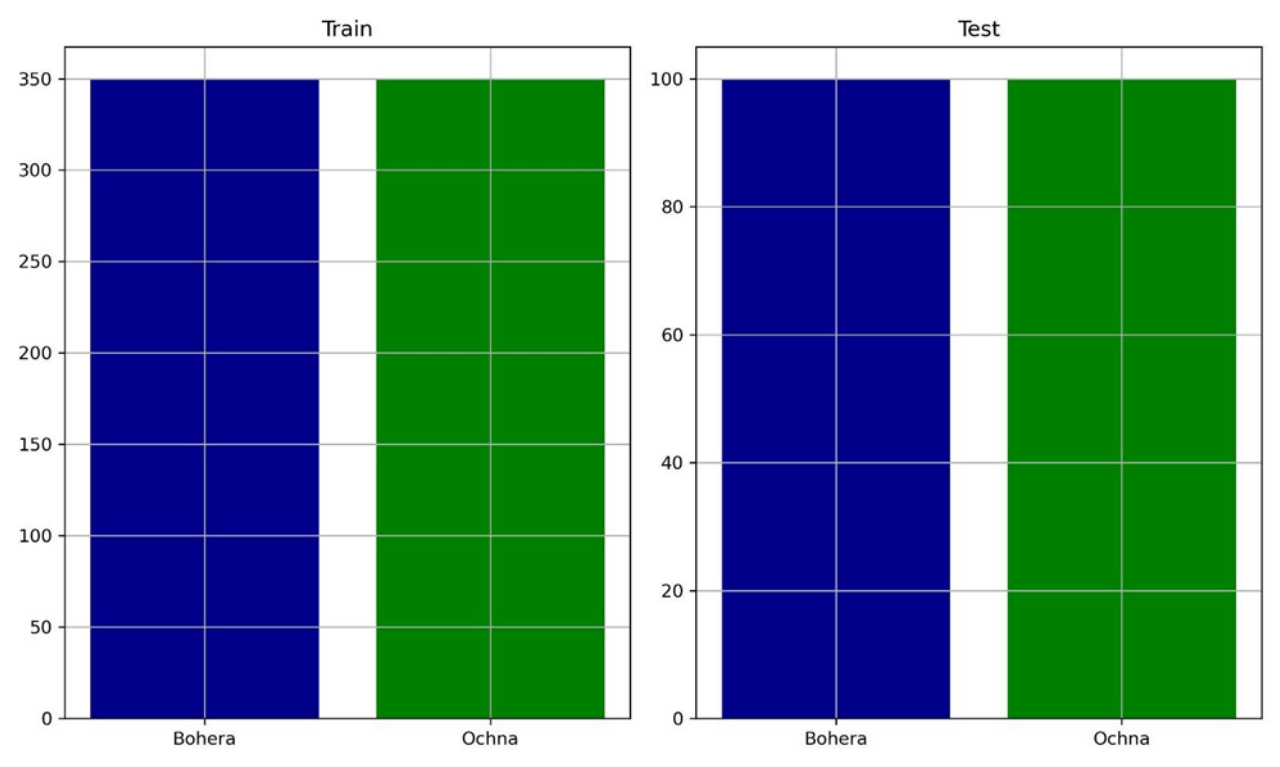

Figure 2. The data increased from 200 images to 700 images by using ImageDataGenerator.

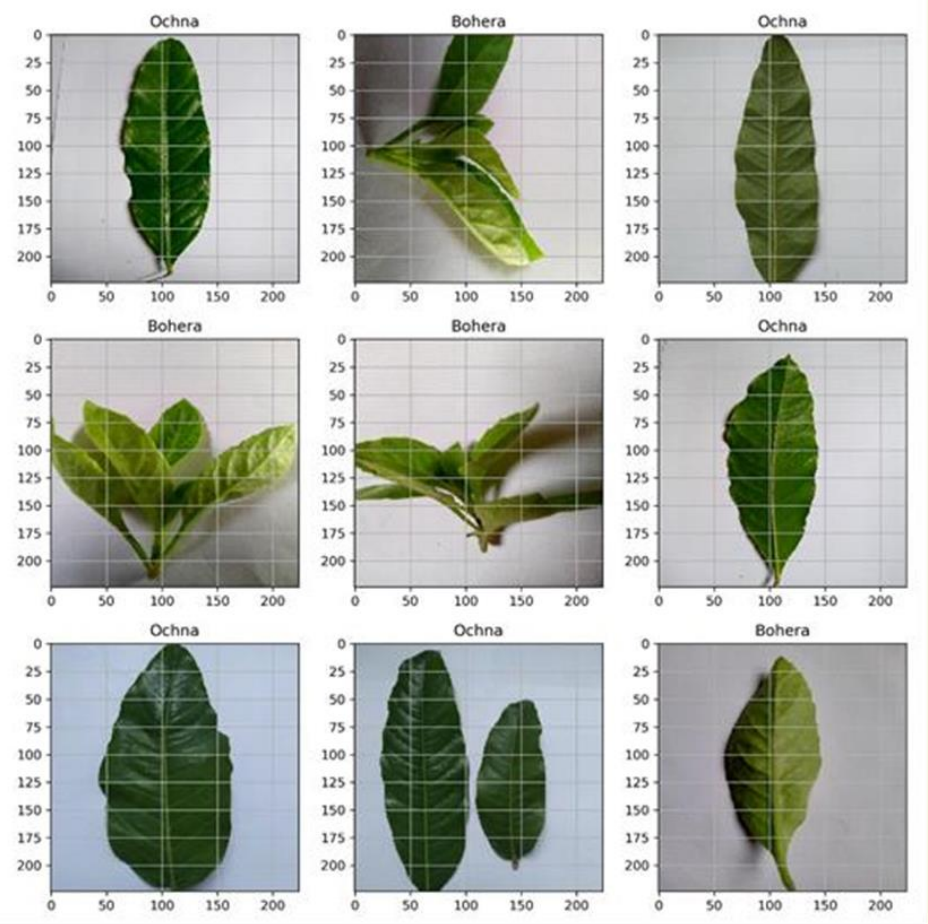

Figure 3. Some of the sample input images deriven from kaggle.

After the process of data reading of computer, the segmentation level that is seen as black image of leaf starts and the segmentation of image is converted to grey scale, finally by defining a threshold the edge lines of the plant leaf are obtained, see ***Figure 4.***

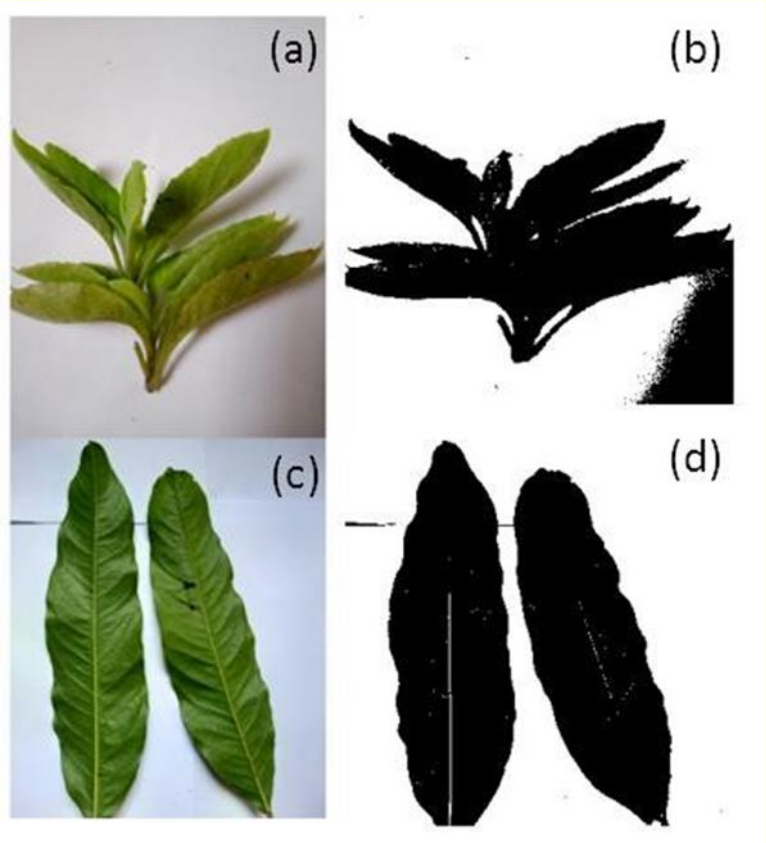

Figure 4. The edge lines of the Bohera and Ochna plant leafs and their images obtained by data segmentation are seen in (a), (b) and (c), (d), respectively.

Results

Classification of Dataset and Performance of the System

Classification of plant images is the process of comparing input images with test images. The accuracy of the classification depends on the number of images the computer is trained on. In this process, at least half of the available image data should be used by the system during the testing phase. The leaf image of the plant is entered into the system as training data. CNN, which consists of hidden layers, processes the image, compares it with the training data set and classifies the correct medicinal plant. Correct classification will help non-experts to identify medicinal plants correctly. The model trained for the neural network is tested on the test data to obtain the corresponding classification accuracy.

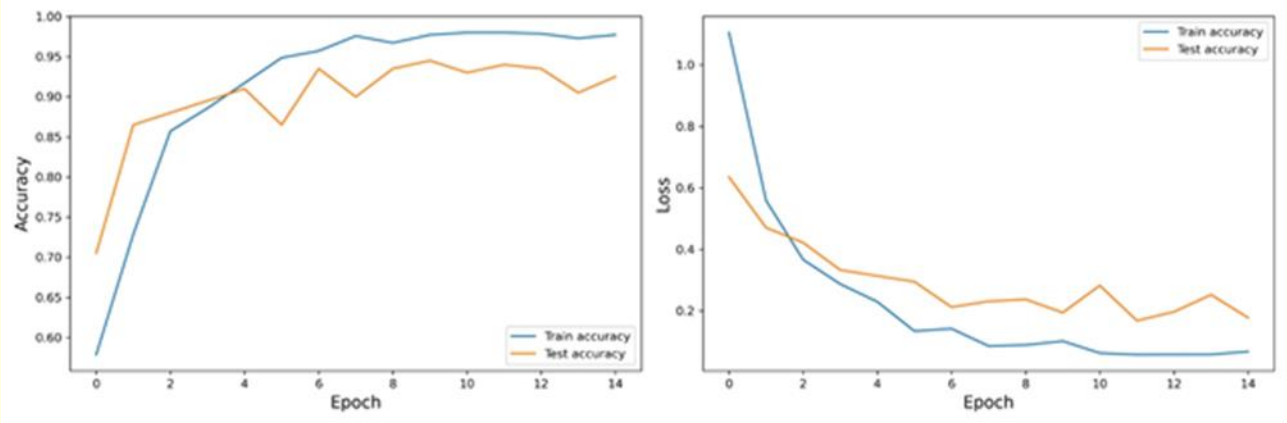

Figure 5. The accuracy and loss plots obtained from the data.

The left side of ***Figure 5*** shows the accuracy we get by training the images applying the CNN model. The right side of ***Figure 5*** shows the loss or the errors that appeared in the time of training. This plot shows that the loss decreases. Namely, there is a high accuracy at the training time. The accuracy rate is given in ***Table 2*****.**

Table 2. Classification Accuracies (100%)

Plant Name	Training Accuracy (%)	Test Accuracy (%)
Bohera	97	96
Ochna	97	96

Discussion

It is thought that medical plants are used by 80% of the world population, mostly in developing countries. The chemical structures present in plants play a role as part of the physiological functions of the living flora, and therefore it is believed that plants are highly compatible with the human body in the treatment of disease. When the primitive medicine literature is examined, it is mentioned that herbal medicines are used in the treatment of some diseases that develop due to aging, such as diabetic wounds, memory loss, osteoporosis, immune system disorders and liver disorders.[22]

In recent years, medical plants have been used even to prevent cancer.[23] Epidemiological studies show that cancer is mostly caused by diet and smoking. The first step in prevention of cancer is to stop its development. Some plants have been shown to be effective in stopping this development. For example, an extract of the bark of the plant Pygeum africanum (Prunus Africana) is used in Europe for the prevention and treatment of prostate disorders. It has been stated that this plant has an important effect on the regulation of prostate cancer both in vitro and in vivo, therefore it can be used in individuals who are more likely to develop prostate cancer.[24] In another study, the medicinal usefulness of the plant called Azadirachta indica (Family Meliaceae) Neem was examined. A large number of about 60 different types of biochemicals, including steroids, have been obtained from this plant. The anticancer properties of the plant, its protective and suppressive effects against various types of cancer have been investigated in many ways.[23] Also, Ngo et al. showed that the extract of the rosemary plant and some compounds present in this plant, carnosol, carnosic acid, etc., showed anticancer properties.[26] There are many other plant species with medicinal ingredients but we cannot mention them all here, for more details we refer the study by Sofowora et al.[27]

Today, some of the experiences we have gained from the internet and environmental factors can lead to information pollution. In particular, from a health point of view, information pollution about which plants are used in medical treatments and which are frequently encountered in nature is increasing rapidly. In

parallel with this, people lose their health in some cases. The developed algorithm will be of great benefit to people in recognizing the plant species they encounter. Thus, people will be prevented from harming their health by using the wrong plants.

Even an expert with sufficient knowledge of plant species may have difficulties in recognizing some plants from time to time. Because the visual illusions caused by the person and the illusions created by the environmental effects in the appearance of the plant can make it difficult to recognize the plant. Therefore, we can use a plant that contains harmful compounds for human health in a way that endanger our health. To avoid this, machine learning algorithms have been developed in plant recognition and classification in recent years. Thanks to these algorithms, we can more accurately identify the plants and their contents, about which we have random information.

In a study by Oppong et al., a computer vision algorithm was developed using CNNs and Log-Gabor filters to identify medicinal plants, based on some of their morphological features. In their study, they used 10 pre-trained networks, namely Alexnet, GoogLeNet, DenseNet201, Inceptionv3, Mobilenetv2, Restnet18, Resnet50, Resnet101, vgg16 and vgg19, and they obtained different degrees of success for them.[28] In a study conducted by Kayhan et al., five medicinal and aromatic plants (MAP) were classified using the Naive Bayes Classifier (NBC), Classification and Regression Tree (CART), KEN Nearest Neighbor (KNN) and Probabilistic Neural Network (PNN) methods.[29] Also, Zhang et al. introduced some widely known machine learning algorithms and presented how these algorithms are applied to the analysis of medicinal plants.[30]

In the present study, we have developed an algorithm that has an accuracy rate of 97%. The network was created with ImageDataGenerator, feature selection and CNN method.

In general, efficiency determiner of the ML methods is presented through the number of metric criteria.[31] The confusion matrix (CM) is an intuitive tool for measuring classifier performance with True and False values last of all classification. In ***Table 3***, we see the comprehensive description of CM. The rows represent the actual data and the columns stand for predicted data. The abbreviations TP, FN, FP and TN signify the True Positive (actual positivity in the data that predicted correctly as positive for the corresponding model), False Negative (actual negativity in the data that predicted correctly as negative for the corresponding model), False Positive (actual negativity exists in data, but the studied algorithm has estimated it as positive) and True Negative (actual positivity exists in data, but the model has predicted it as negative), respectively.

Table 3. The representation of CM.

	Actual Data	
Predicted data	**Positive (1)**	**Negative (0)**
Positive (1)	TP	FN
Negative (0)	FP	TN

In our study, the obtained results can be reported by the following CM (***Figure 6***).

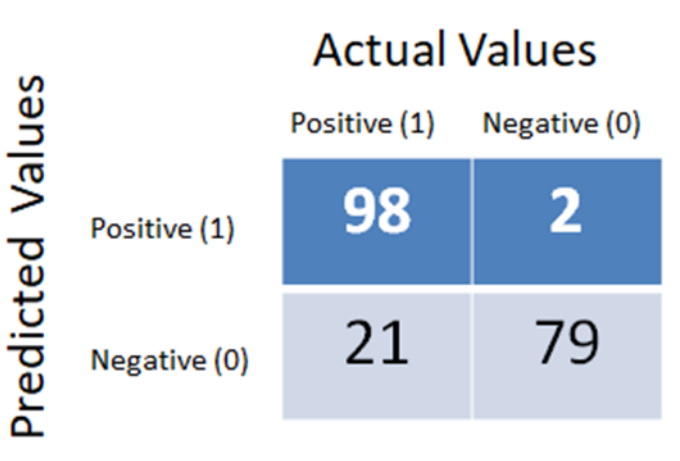

Figure 6. The CM obtained for our data. The TP value is 98, the FP value is 2, the TN is 79 and finally the FN is 21.

In order to obtain the standard deviation for the corresponding single-layer algorithm with 128 neurons, approximately 25% of the neurons were closed randomly with the dropout method and the algorithm was run by doing 10 repetitions. As a result, the average accuracy was obtained as 0.95. These results indicate that there is a standard deviation of 0.021 for the accuracy value, and the accuracy value can be written as 0.97±0.021.

Conclusion

The forms of treatment encountered in the field of modern health are mostly based on the use of extensive technological devices and expensive drugs. For this reason, complementary medicine in health (CM) is still a very good alternative approach for underdeveloped societies with limited economic conditions. Appropriate and correct use of complementary medicine is very important. For this, users should be informed about the appropriate CM methods developed and especially awareness should be created in this regard. In this regard, the World Health Organization (WHO) has also conducted a study and prepared a guideline for the correct use of traditional and complementary medicine.[32] In this sense, WHO explained to governments and non-governmental organizations how CM users could be improved in therapeutic choices regarding health care.

In the present study the training set accuracy of using the CNN approach is get 97%. This accuracy stends on the number of epochs that delivers. Also, the test data accuracies were at intervals 95% - 99% when tested on 700 images each of Bohera and Ochna rare plants. It is possible that the degree of accuracy to be obtained will decrease somewhat with the increase in the number of plants. However, this value will still be within acceptable limits. A successful machine learning methodology is promising for the future, both in recognizing and classifying plants as in our study, and in producing successful herbal formulas and using them in traditional medicine with the development of the learning methodology.

Performance analyzes can be evaluated by including more different types of plants in the created algorithm. Processing of a larger number of plant photos could not be done due to the limitations of our technological possibilities (Each picture consists of approximately 13MP and there are 700 images for the 2 species we have and there are 13 MPx700≈10 Billion Pixels worth of pictures to be examined.), it is possible to do this with more powerful machines. If the plant species that adversely affect the metric values in the CM are excluded from the study, then created algorithm makes less errors.

Acknowledgements

The author has no conflict of interest to declare.

Author Contributions

Fatma Söğüt: Idea/concept, design, data collection and processing, literature review, article writing, critical review, references and fundings.

Bora Reşitoğlu: Idea/concept, design, data collection and processing, literature review, article writing, critical review, references and fundings.

Evrim Ersin Kangal: Idea/concept, design, data collection and processing, literature review, article writing, critical review, references and fundings.

References

1. Ünal M, Dağdeviren HN. Geleneksel ve Tamamlayıcı Tıp Yöntemleri. Euras J Fam Med 2019;8(1):1-9.
2. Mollahaliloğlu S, et al. The New Period in Traditional and Complementary Medicine. Ankara Med J 2015;15(2):102-105.
3. Öztürk YE, Dömbekçi HA, Ünal SN. Geleneksel Tamamlayıcı ve Alternatif Tıp Kullanımı. Bütünleyici ve Anadolu Tıbbı Dergisi 2020;1(3):23–35.
4. Kigen GK, et al. Current trends of traditional herbal medicine practice in Kenya: a review. African Journal of Pharmacology and Therapeutics 2013;2(1):32-37.
5. Nimri LF, Meqdam MM, Alkofahi A. Antibacterial activity of Jordanian medicinal plants. Pharmaceutical biology 1999; 37(3):196-201.
6. IUCN (International Union for Conservation of Nature). Approaches to Conservation of Medicinal Plants and Traditional Knowledge. A Focus on the Chittagong Hill Tracts. Bangladesh Country Office. 2010, 40 P.
7. Joshi AR, Joshi K. Indigenous knowledge and uses of medicinal plants by local communities of the Kali Gandaki Watershed Area, Nepal. Journal of Ethno pharmacology 2000;73(1-2):175–183.
8. Anselem A. Herbs for healing pax herbals Edo State, Nigeria. 2004.
9. Dold AP, and Cocks ML. The trade in medicinal plants in the Eastern Cape Province, South Africa. South African Journal of Science 2002;98:589–598.
10. Donalek, C., Supervised and unsupervised learning, Astronomy Colloquia, California Institute of Technology, USA, 2011.
11. CNL [Internet, cited: 14.08.2023], Available from: https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/
12. Bektaş J, Bektaş Y, Kangal EE. Integrating a novel SRCRN network for segmentation with representative batch-mode experiments for detecting melanoma. Biomedical Signal Processing and Control 2022;71(B):103218
13. Tan JW, et al. Deep Learning for Plant Species Classification using Leaf Vein Morphometric. IEEE/ACMTransactionson Computational Biology and Bioinformatics 2020;17(1):82-90.
14. Cope JS, et al. The extraction of venation from leaf images by evolved vein classifiers and ant colony algorithms. In: Talon B, et al. editors. In International Conference on Advanced Concepts for Intelligent Vision Systems; 2010; Dec 13-16; Sydney, Aust. Berlin: Springer 2010.
15. Anami BS, Suvarna SN, Govardhan A. A combined color, texture and edge features based approach for identification and classification of Indian medicinal plants. International Journal of Computer Applications 2010;6(12):45-51.
16. Kadir A, et al. Neural network application on foliage plant identification. International Journal of Computer Applications 2011; 29(9):15-22.
17. Larese M, et al. Legume identification by leaf vein images classification. Progress in Pattern Recognition, Image Analysis. Computer Vision, and Applications 2010; 447-454.
18. KGL [Internet, cited: 14.08.2023], Available from: https://www.kaggle.com/code/codefantasy/identifying-plants-and-it-s-medicinal-properties
19. Bohera [Internet, cited: 14.08.2023], Available from: http://dhcrop.bsmrau.net/bohera/
20. Oc, [Internet, cited: 14.08.2023] Available from:https://pza.sanbi.org/ochna-serrulata
21. IDG,[Internet,cited:14.08.2023]Availablefrom:https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator
22. Kamboj VP. Herbal medicine. Current Science 2010;78(1):35–39.
23. Yasukawa K. Medicinal and Edible Plants as Cancer Preventive Agents. In: Vallisuta O, Olimat SM, editors. Drug Discovery Research in Pharmacognosy. Rijeka: In Tech; 2012. pp: 181-208.
24. Shenouda NS, et al. Phytosterol Pygeum africanum regulates prostate cancer in vitro and in vivo. Endocrine 2007;31(1): 72-81.
25. Paul R, Prasad M, Sah NK. Anticancer biology of Azadirachta indica L (neem): a mini review. Cancer Biol Ther 2011;12(6): 467-476.
26. Ngo SN, Williams DB, Head RJ. Rosemary and cancer prevention: preclinical perspectives. Crit. Rev. Food Sci. Nutr 2011;

51(10):946- 954.

27. Sofowora A, Ogunbodede E, Onayad, A. The role and place of medicinal plants in the strategies for diseaseprevention. African journal of traditional, complementary, and alternative medicines. AJTCAM 2013;10(5):210–229.

28. Oppong SO, et al. A Novel Computer Vision Model for Medicinal Plant Identification Using Log-Gabor Filters and Deep Learning Algorithms. Comput. Intell Neurosci 2022; Article ID 1189509.

29. Kayhan G, Ergün E. Medicinal and Aromatic Plants Identification Using Machine Learning Methods. BAJECE 2020; 8(1): 81-87.

30. Zhang Y, Wang Y. Recent trends of machine learning applied to multi-source data of medicinal plants. Journal of Pharmaceutical Analysis 2023; https://doi.org/10.1016/j.jpha.2023.07.012

31. Caruana R, Niculescu-Mizil A. Data mining in metric space: an empirical analysis of supervised learning performance criteria. Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle- WA, USA, 2004.

32. WHO. Guidelines on developing consumer information on proper use of traditional, complementary and alternative medicine, Italy, 2004.