THE AUTOMATIC IDENTIFICATION OF BUTTERFLY SPECIES USING DEEP LEARNING METHODOLOGIES

Capstone Project

Seda Emel TEK KARA

İSTANBUL, 2020

MEF UNIVERSITY

THE AUTOMATIC IDENTIFICATION OF BUTTERFLY SPECIES USING DEEP LEARNING METHODOLOGIES

Capstone Project

Seda Emel TEK KARA

Advisor: Asst. Prof. Dr. Tuna Çakar

İSTANBUL, 2020

MEF UNIVERSITY

Name of the project: The Automatic Identification of Butterfly Species Using Deep Learning Methodologies
Name/Last Name of the Student: Seda Emel Tek Kara
Date of Thesis Defense: 30/12/2020

I hereby state that the graduation project prepared by Seda Emel Tek Kara has been completed under my supervision. I accept this work as a “Graduation Project”.

30/12/2020
Asst. Prof. Dr. Tuna Çakar

I hereby state that I have examined this graduation project by Seda Emel Tek Kara which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

30/12/2020

Director of the
Graduate Program in
Information Technologies

We hereby state that we have held the graduation examination of Seda Emel Tek Kara and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Dr. Tuna Çakar	………………………..
2. …………………………..	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Seda Emel Tek Kara	30/12/2020	

EXECUTIVE SUMMARY

THE AUTOMATIC IDENTIFICATION OF
BUTTERFLY SPECIES USING DEEP LEARNING METHODOLOGIES

Seda Emel Tek Kara

Advisor: Asst. Prof. Dr. Tuna Çakar

DECEMBER, 2020, 24 Pages

Automatic identification of butterflies, especially at an expert level, is needed for important topics such as species conservation studies, minimizing the insect damage on plants in agriculture, and biodiversity conservation. An efficient and performing model which can define species even in small datasets may reduce the need for experts on the subject or reduce the time spent for identification. By the model proposed in this study, automatic taxonomic classification of butterflies was studied. Convolutional Neural Network (CNN) applications were applied on 7148 photographs of six butterfly species used in the study. 80 percent of the data set was reserved for training and 20 percent for testing, and the model was run with the relevant parameters. At the end of the study, an accuracy degree of 92.73% was obtained.

Key Words: butterfly classification, artificial intelligence, deep learning, species identification, convolutional neural network

ÖZET

DERİN ÖĞRENME METODOLOJİLERİNİ KULLANARAK
KELEBEK TÜRLERİNİN OTOMATİK OLARAK TANIMLANMASI

Seda Emel Tek Kara

Proje Danışmanı: Dr. Öğr. Üyesi Tuna Çakar

ARALIK, 2020, 24 Sayfa

Kelebeklerin, özellikle uzman seviyesinde, otomatik olarak tanımlanabilmeleri; tür koruma çalışmaları, tarım alanlarında bitkiye zararlarının aza indirilmesi, biyolojik çeşitliliğin korunması gibi alanlarda ihtiyaç duyulan bir konudur. Otomatik tür tayini, uzman ihtiyacını aza indirebilir ve tanımlama süresini kısaltabilir. Bu çalışmada, derin öğrenme metodlarından biri olan Evrişimsel Sinir Ağları (CNN) kullanılarak öne sürülen modelle, kelebeklerin otomatik olarak taksonomik sınıflandırılması konusunda çalışılmıştır. Çalışmada, altı kelebek türüne ait 7148 fotoğraf üzerinde CNN uygulamaları yapılmıştır. Oluşturulan model, altı konvolüsyon tabakası ile birlikte, konvolüsyon katmanlarının sonunda kullanılan ek katmanlar ve bunların içinde kullanılan farklı parametrelerden oluşmaktadır. Veri setinin yüzde 80'i eğitim, yüzde 20'si de test için ayrılmış ve model, ilgili parametrelerle beraber çalıştırılmıştır. Çalışma sonunda % 92.73'lik bir doğruluk derecesi elde edilmiştir.

Anahtar Kelimeler: kelebek tanımlama, yapay zeka, derin öğrenme, türlerin tanımlanması, evrişimsel sinir ağları

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Every step of the routine flow taking place in our daily life, such as the steps we take, the objects we see in the environment, the words we speak, is data in today's world. If we think more broadly, company accounts, financial statements, agricultural production, social media accounts, movies, hotel reservations, shopping, customer information stored in companies and many more are being a dataset. In fact, rather than the size of this data, the way it processes the data and the meaningful results to be obtained have become much more important in today's life.

Enabling computers to do things that people do instinctively, such as facial recognition, speech understanding, decision-making, classification, is the basis of artificial intelligence. The beginning of applications about artificial intelligence concept goes back to the 1950s. Today, by the help of artificial intelligence techniques like machine learning, deep learning and image processing, data have effective areas of use. In addition to its technical and practical applications in business areas such as banking, trade and communication, it also has applications for large-scale scientific and public problems. Research and developments in AI continue in the areas that are very important for our lives, such as effective use of world resources, conservation of biodiversity, and sustainable agriculture. As an example, to these works, Akçay et al. [1], Haghiabi et al. [2], and Santangeli et al. [3] studies can be given.

Akçay et al. [1] studied automatic identification of birds from the photographs of geographically tagged bird clusters. They did this work with deep learning approaches such as Convolutional Neural Network (CNN). The bird count was made from the photographs of these bird flocks with the location label and put it on the map. The model which is proposed and the findings of such studies in this context, can facilitate many bird counting activities. Studies in this area can provide significant benefits in counting birds at migration bottleneck points, aerial bird counting, and bird watching over power plants.

Water and the quality of it has a vital importance for the environment and human health, with its use in drinking, agriculture and industry. Haghiabi et al. [2] developed a model for automatically measuring the quality of water using deep learning algorithms over a sample river system. In his study, he tried to estimate water quality components using ANN (artificial neural network), GMDH (group method of data handling) and SVM (support vector machine) methods. The data set that is meant to be explained with water quality components are metrics such as calcium, chloride, magnesium, sodium, or pH in water. The SVM structure, one of the models they proposed to measure the quality of water in their study, has become the most successful model for practical purposes and has the most accuracy. The outputs of such studies could be that these types of models can be used in areas such as conservation projects, feasibility studies, installation of irrigation systems in agricultural fields, determination of agricultural crop patterns in the fields, establishment of water treatment systems for industries.

Biodiversity is melting day by day, especially in the last 50 years it has declined significantly. All fish, amphibian, reptile, bird, and mammal species in the world declined by 68% between 1970 and 2016 [4]. This is a really serious number. The loss of this diversity in the world means that not only non-human beings but also human life will be endangered. So, the situation is not just ethical. Threatened human health and food safety is one of the consequences of biodiversity loss. The issue of where and how we produce our food is very important in this respect. Improper agricultural practices have a catastrophic impact on both biodiversity and consequently food. Economic practices aiming to consume world resources constantly feed these negative results. Life and resources on Earth are not endless. Diversity should be monitored regularly and effectively. Strategic decisions, for example, regarding the expansion of agricultural lands or the use of existing fields, should be made according to the results of this monitor. Santangeli et al. [3] proposed an idea of getting drone images captured by a thermal sensor on it and an AI algorithm together to identify farmland bird's ground nests with an efficient and performing model. Their semi-automated work powered by deep learning and drone based thermal images is able to detect the nests of ground nesting birds in an agricultural area.

2. BACKGROUND

In this study, we used the pictures of butterfly species for classification by using Convolutional Neural Network (CNN) methods. This section gives some technical background information about the target group and the deep learning concepts regarding our study.

2.1. The insect group under study

Butterflies and moths are the insects which belong to the order of Lepidoptera. The group is regarded as having the most species diversity in the entire insect class after beetles (Coleoptera order) [5]. Butterflies and moths are one of the most popular and well-known groups among insects. Their species are flying towards the light both during the daytime and at night. The butterflies are mostly diurnal due to their daytime flying, on the other hand moths are also called night butterflies because they are generally more active at night.

Butterflies and moths are the animals which undergo metamorphosis. From the beginning of their life until they become an adult, they spend certain periods of their life in different life forms. The life of a Lepidoptera consists of four stages, respectively as egg, larva, pupa, and adult (Figure 1).

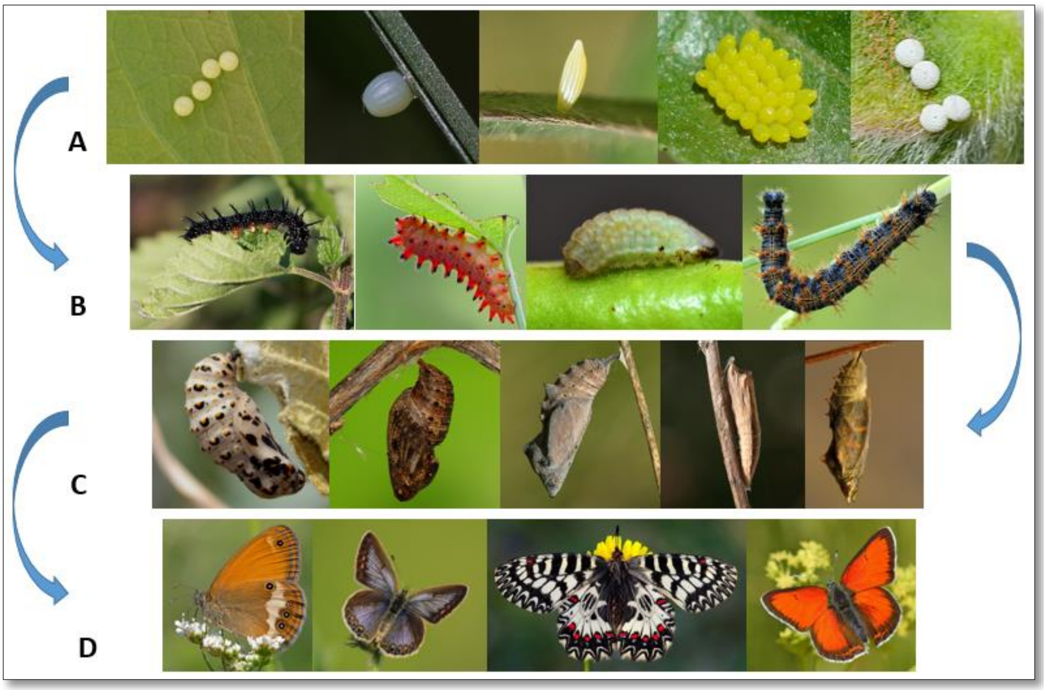

Figure 1: Life stages of Lepidoptera species; A: egg, B: larva, C: pupa, D: adult. Each photo belongs to different species

The egg is the first stage in which a butterfly starts its life, which is formed by the male and female individuals after mating and left on the appropriate plant by the female individual. The duration of the egg stage, the number and shape of the egg and the way it is laid differ from species to species. While some species lay their eggs in clusters on a leaf, some species lay a single egg on a single leaf and do this by applying the same to the other leaves of the plant.

Eggs, after spending some time, varying according to species, start to open and small sized caterpillars that have grown in eggs begin to hatch. With their emergence, the feeding processes also begin. In some species, the first feeding begins by eating the eggs they hatch. Except for a few examples, caterpillars are herbivorous. The caterpillars start feeding on the plant they are placed until they are fully grown. Some caterpillars feed on only one type of plant, while others can feed on many plant species. They feed on the plants solitarily or gregariously depending on the species.

Towards the end of the larval stage, the movements of caterpillars begin to decrease, and they stop feeding. This stage is now the caterpillar's preparation for pupa. Here the larva attaches itself to the leaf upside down and gradually turns itself to the pupa. Metamorphosis takes place inside the pupa. The time it takes to enter pupa (prepupa phase) varies from species to species. It can be rapid, or may take a couple of weeks, or maybe it takes more than a year. The duration of the pupa phase can also vary. When the adult form of the butterfly, which undergoes metamorphosis within the pupa, emerges from the head section of the cocoon, it gradually opens its wings as an adult butterfly and moves towards looking for food and mates.

Most butterflies and moths spend the whole winter in one life stage. Some species may overwinter in the egg state, while others in the form of larvae, pupae and even adults. For example, *Vanessa atalanta* and *Polygonia c-album* species spend the winter in adulthood by hiding in suitable places. Moreover, some butterflies migrate in their adult stages across a continent or even continents. For example, *Danaus plexibus* species migrate to the south in autumn, from the northeastern US and Canada, to central Mexico. In spring and summer, they moves back to North America. Another species which can be also seen in Turkey while their migration is *Vanessa cardui*. It has different migratory paths in the world. One of them is the way from northern Africa to Europe [6].

Butterflies have some essential roles in the ecosystem. Since they feed from the flowers during their adult stages, while visiting their food sources, they also help the plants to be pollinated. During their egg, larva, pupa, and adult stages, they become food for other animals such as birds, other insects and reptiles. Since most of them are herbivores in their larval period, some butterfly species consume crop plants in agricultural areas. However, in the ecological system both plants and insects act together. While butterfly larvae are feeding on plants, they also indirectly attract other insects that are beneficial for the plant. Thus, at the cost of becoming a prey, butterfly larvae contribute to reproduction of other insects and so the functioning of a healthy ecosystem. Moreover, some species in their larval period are in a directly useful position for agricultural lands by feeding on wild plants that are in competition with the plants planted in agricultural areas. Butterflies react quickly to changes in environmental conditions. In a healthy ecosystem, it is important that the diversity and density of butterflies be in a certain balance. Changes in these proportions tells us about the ecosystem. One way to understand a malfunctioning ecosystem is looking at butterflies. In other words, butterflies act as bioindicators for ecosystem health. Therefore, these insects should be monitored regularly and any information to be obtained about them is important.

According to an article published by Robbins and Opler in 1997 [7], there are approximately 17500 butterfly species in the world. This number is 700 for the United States and Canada, 500 in Europe, and around 380 for Turkey [8]. Morphological differences are mostly used to identify a butterfly. Additionally, molecular level tools (such as DNA barcoding) can be used.

With their colorful wing structures in adult stages, butterflies attract the attention of amateur naturalists and photographers. There are many country-specific websites whose members share their photographs and observations about butterflies. Apart from these, people take photographs and share them publically from their social media accounts. These pictures sometimes have a description about the species but mostly without a tag or a definition. In order to identify the species, it is necessary to go to the field for a certain period of time with people having experience and/ or to observe the insects using guidebooks. It is necessary to look for alternative ways due to the scarcity of experts who can define all the species, the availability of many undefined photographs, and the time spent for identification.

With the development of computer technologies, models are being developed that can enable butterflies to be classified according to their morphology, with the help of artificial intelligence algorithms. Butterfly photographs in the adult stage have been used in studies conducted on the subject so far. This study also used the pictures taken from the websites that have the records of butterfly observations in Turkey, and a model was developed that can automatically determine the species of butterflies from its images.

In the future, a well-prepared model can be applied to egg, larvae, or pupae pictures of butterflies (with sufficient data for training and testing), because the species are different from each other with also the morphology of these life stages.

2.2. Neural Networks

2.2.1. Biological Neural Networks

The nervous system of humans consists of a complex network system of nerve cells called neurons. The mission of this system is to receive information from the environment, process it and carry it to the target place in the body, which results in learning, reactions and other vital behaviors in life. Most neurons consist of the parts named as dendrite, cell body, axon, and axon terminal. There are different types of nerve cells. Figure 2-A shows a multipolar neuron (found in brain and spinal cord) and its sections. In Figure 2-B, a schematic view of a part of the network formed by this neuron with other cells can be seen. Here, the stimulus from the previous neuron (the leftmost part in the figure where the cell terminal is seen) is received through the dendrites of the other cell, transmitted to its cell body, where it undergoes some electrochemical changes. Then, the electrochemical signal is forwarded along the axon if it meets a certain threshold and reaches the axon terminal. Lastly, the message is delivered to the next junction part called synapse, which is responsible for transmitting the message from the axon terminal of the cell to the dendrite of the next nerve cell. In this way, as a network, the incoming message is processed and delivered to the target places where the message will be transmitted.

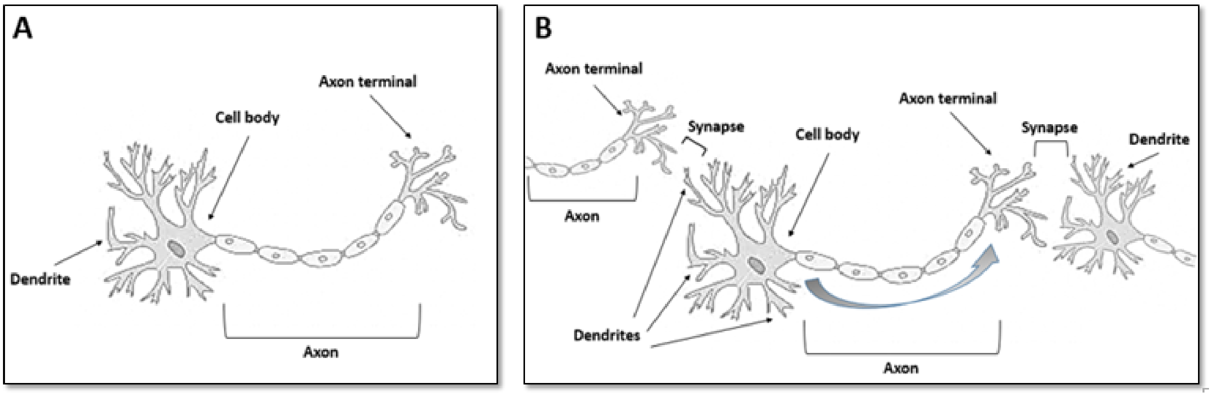

Figure 2: Generalized diagrams of biological neurons. A: A neuron with its main parts. B: Small part of a biological neural network composed of multiple neurons

2.2.2. Artificial Neural Networks

Artificial neural networks (ANN) are the system designed in analogy to biological neural networks by which learning processes occur. In the history, McCulloch and Pitts firstly presented the concept of artificial neurons in 1943. Rosenblatt developed the idea of *perceptron* in 1958. However, in 1969, Minsky and Papert showed the perceptron models' deficiencies and artificial neural network concepts lost the attention. In the 1980's, the subject gained interest again and after that lots of innovative research has been done about artificial neural networks [9, 10].

The working principle of an artificial neuron is similar to the biological nerve cell. The input goes into the artificial neuron, while it moves along the neuron, it passes through some processes –calculations, then leaves the cell and creates an output (Figure 3).

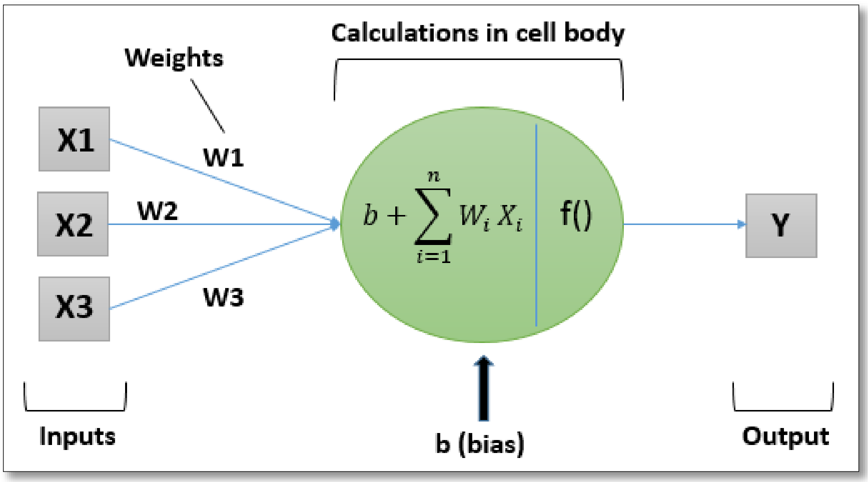

Figure 3: Diagram of an artificial neuron

An artificial neural network is composed of multiple neurons as layers. In Figure 4, the diagram of an artificial neural network can be seen. In the structure, there can be seen the parts of inputs, weights, neurons, and output. Inputs are the information about the data samples which the model will train to learn (such as images needed to be classified). Weights implies the importance of inputs. Learning processes are mainly done by summation and activation functions inside the artificial neurons. Summation function calculates the total inputs by using their weights. Activation function determines the output value by using the input calculations. There are different types of activation functions as linear, step, or non-linear (sigmoid) functions. The number of hidden layers can change in different models. When the hidden layers are added more, the model becomes deeper and more complex.

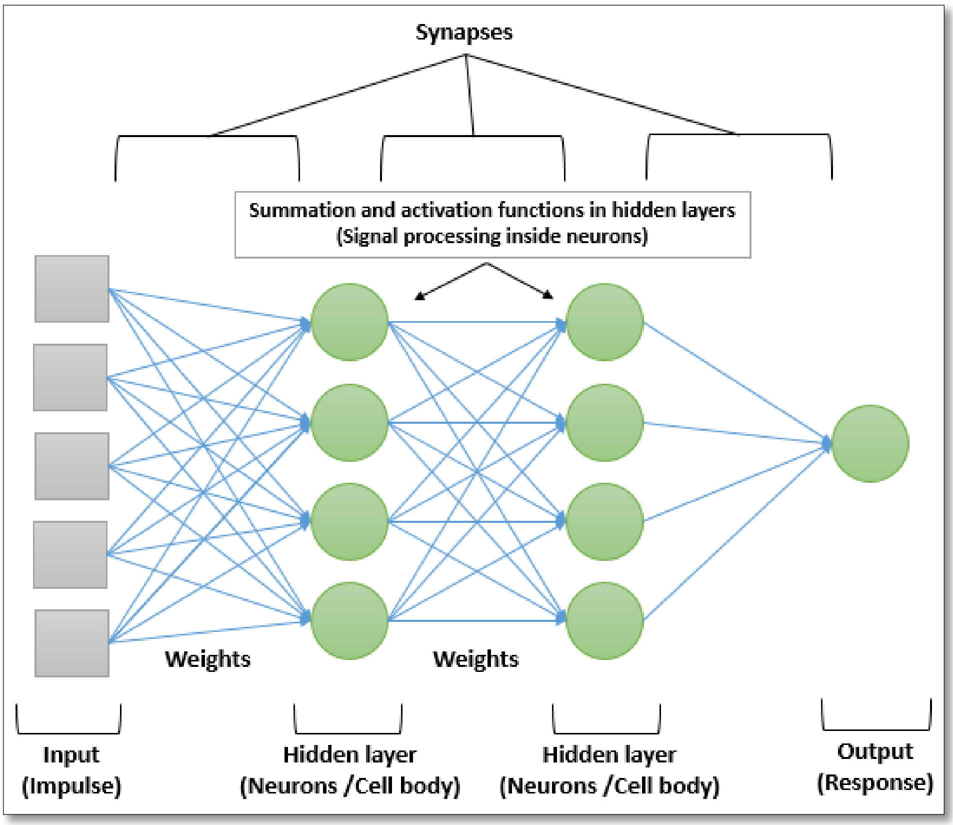

Figure 4: Artificial neural network model that mimics biological neural networks

2.3. Convolutional Neural Networks (CNN)

Convolutional neural network (CNN) is a structure designed to distinguish, label, classify and define the images given to it. In doing so, it basically uses people's sight and recall logic. We see the shape and color of an object when we look at it, recognize it from these properties in our next view and label it in our minds. CNN also works in a similar manner. For example, when someone shows us an apple and says its name, we see its shape, color, texture, branch connection on it, the seeds which are seen in the middle of the apple if it is cut. Thus, we combine all the visual data about the apple with the “apple” tag in our brain. We learn as much as we see that it is an apple. In our next meet with the apple, all of these features combine in our minds and we know that it is an apple whether the apple is on the table or on the tree. CNN is also doing this in a similar way. While recognizing an image, it processes it through several stages.

Basically, it is a neural network type which usually has parts of input, convolution, pooling, fully-connected, and output (Figure 5).

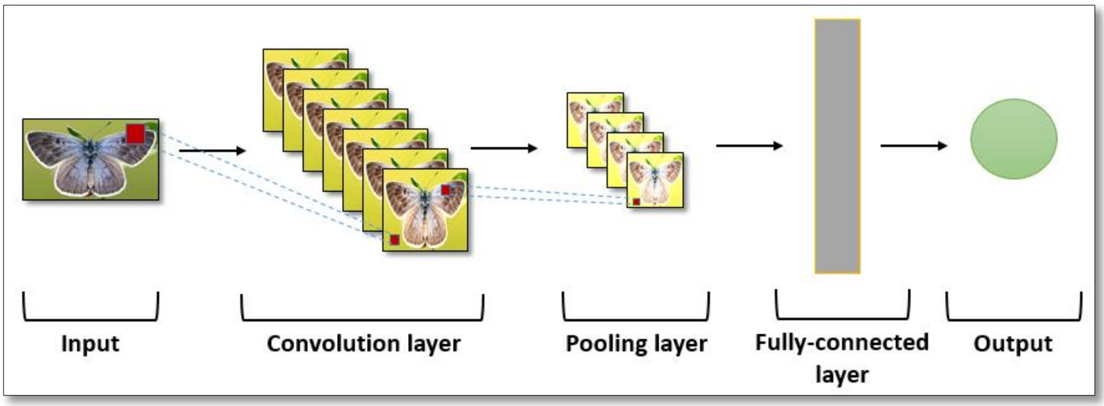

Figure 5: Convolutional neural network architecture

Convolution layer detects properties of images. This layer is the basic building block of CNN. Features are detected by filters which scan the input images in a specific size. Here the filter is put on the image and the values of the two matrices are multiplied by each other, the results are added together, and an output is obtained. The same process is repeated one pixel to the right above the filter image and then row by row. All results are stored in a matrix depending on how many times they went right and down. This output matrix, named Feature Map, gives us the properties of the image. Different filters can be used to extract different features of the same image. This means using different convolution layers. In the pooling layer, the convoluted images are downsampled, in which the maximum or average value of the view is taken. In this layer, the smaller outputs containing enough information for the neural network to make the right decision are used. Most used pooling type is maxpooling. Lastly in a fully-connected layer, optimizations are done. The last layer is the classical artificial neural network structure which is used for classification [11, 12].

A simple CNN design is to use several convolution layers with adding an activation function (e.g., ReLU) after each. Then pooling layers and fully-connected layers are added [11, 12].

There are some widely used architectures in the convolutional networks area. These are the architectures called LeNet, AlexNet, GoogLeNet, VGGNet, ResNet that perform well in different subjects [11, 12].

3. LITERATURE REVIEW

Image classification techniques are widely used in various areas such as face recognition, medical images for diagnosis, and reading handwriting. One other application is about biological species identification used for several disciplines like in agriculture, ecology, wildlife conservation, taxonomy and so on.

Miao et al. (2019) study deep learning methods to classify wildlife images collected from camera traps. They use CNN with VGG-16 and ResNet-50 algorithms and train the data for the images. A total of 111,467 images about 20 African wildlife species are used in their model. Accuracy of their model is 87.5% on average [13].

Cıbuk et al. (2019) propose a deep convolutional neural network (DCNN) based hybrid method including three stages. The two datasets used includes 17 categories of flowers with each including 80 images (Flower17) and 102 categories for 40-258 images for each (Flower102). In the first stage, they trained the datasets for feature extraction using AlexNet and VGG16 algorithms. Then in the next phase, features are concatenated from these two models and more efficient features are selected by the help of mRMR method. Lastly, these selected features became an input for label classification via SVM classifier and RBF kernel. Results of the model accuracies for Flower17 and Flower102 datasets are 96.39% and 95.70% [14].

A deep neural network architecture is built to classify plant seedling images by Alimboyong and Hernandez in 2019. They used 4,234 images belonging to 12 species. The model includes five convolutional layers and two fully connected layers. Training, validation, and testing splits are in the order of 70%, 20%, and 10%. The overall accuracy of the estimation is 90.15%. They also discuss about some future implementations of the same, different or greater amount of datasets with other architectures like AlexNet, VGG16, GoogLeNet, ResNet, or Transfer Learning [15].

Identification of plant diseases is also a studied area which is important in agriculture. Sardogan et al. (2018) mention about detection and classification of tomato leaf diseases. CNN (convolutional, pooling activation function, and fully connected layers) and LVQ (input, Kohonen (competition), and output layers) algorithm is used in their article. They use 500 images in four disease categories, in which 400 is for training and 100 is for testing. In their model, an average of 86% accuracy result is gained. However, it is said that because of the similar appearance of leaves with different diseases, some images are misclassified [16].

Valan et al. (2019) state about the taxonomy of some insect groups. In the research, automatic identification of the four groups of insects in an expert level is aimed. To the end, they study on basically four datasets of 884 face images of 11 families of dipteran (D1); 2936 dorsal images for 14 families of coleopteran (D2), 339 images of three species of a coleopteran genus (D3); and 3845 images of nine species of plecopteran larvae (D4). The authors use the VGG16 algorithm as CNN architecture for feature extraction. In the model five convolutional blocks are placed, in which there are two or three convolution layers and a MaxPooling layer per block. At the end of the fifth block, there are also three fully-connected layers. Overall accuracies of their model are in the order of 92%, 96%, 97%, and 98.6% for D1, D2, D3, and D4 [17].

Lim et al. (2018) study on a model for insect identification automatically. They used the Inception-v3 model on 30 forest insect species. The accuracy of their work is on the average of 94%. After that the authors mentioned about the application, they developed for insect classification [18].

Kasinathan et al. (2020) classify the images of two datasets which include nine and 24 insect classes. The nine-classed images are splitted into 162/63 images of train-test ratio, and the other insect dataset has 785/612 images of train-test ratio. As the model, Kasinathan et al. uses ANN, SVM, KNN, NB, and CNN models. Overall accuracies of their work are 91.5% and 90% for nine and 24 class insects [19].

Thenmozhi et al. (2019) propose an insect pest detection model that can be efficiently used for crop protection in agriculture. They trained their model on three datasets named NBAIR (40 classes of insects), Xie1 (24 classes), and Xie2 (40 classes). Then they compared their results with other deep learning models like AlexNet, ResNet, GoogLeNet, and VGGNet. Their proposed models achieved the highest accuracy for the NBAIR, Xie, and Xie2 datasets in the order of 96.75%, 97.47%, and 9.97% [20].

Zhu et al. (2016) make a combination of DCNN AlexNet and SVM models to classify lepidopteran species. They use 1301 images for 22 species. 80% of the dataset is split as training and the 20% is testing. The best accuracy of their model is 100% while classifying the test data. They concluded that when DCNN depth is increased, classification accuracy is not always high. However, their proposal may be a potential real time classifier model for lepidopteran identification [21].

Rodrigues et al. (2020) categorize butterfly species of 10 with 830 images. To this end, they use CNN methodology in which the sequential model is composed of five layers. Their study shows 90% accuracy with 80:20 ratio of train-test split [22].

Arzar et al. (2019) study about an image processing plus CNN (GoogleNet) model to identify butterfly species. They use 120 species with four species in their dataset and achieved 97.5% of overall accuracy in their results [23].

Zhao et al. (2019) state Faster R-CNN algorithm in their models for butterfly species recognition. 5695 images for 111 species are used in the dataset. Their study achieves 70.7% of stable accuracy. In their structure, the Faster R-CNN method has a great performance for classification speed [24].

For the purpose of automatic butterfly labeling, YOLO algorithm is used in the research done by Liang et al. (2020). The authors claim that their proposed model has a good solution for the problems in small sampling with a high accuracy in automatic detection and recognition of butterfly species. They studied on 94 species belonging to 11 genera with a total of 5695 images. The pictures are both specimen and ecological photography. In conclusion, they achieved the accuracy of 98.35% [25].

Theivaprakasham (2020) tries to develop a model having a high accuracy to identify butterfly species. For this purpose, they make trials on the dataset of 34024 images of 315 species. They also use augmentation techniques to amplify their dataset. The authors built 11 DCNN models using 11 architectures which are ResNet-18, ResNet-34, ResNet-50, ResNet-121, ResNet-152, Alex-Net, DenseNet-121, DenseNet-161, VGG-16, VGG-19, and SqueezeNet-v1.1. Among these models, ResNet-152 achieves a maximum top-1 accuracy (94.44%), top-3 accuracy (98.46%) and top-5 accuracy (99.09%) [26].

Almryad and Kutucu (2020) studies on three models of VGG16, VGG19, and ResNet50 and makes a comparison among these models by using a butterfly dataset having 17769 images with 10 genera. They achieved the highest accuracy by VGG16 model with 80% for train and test data [27].

4. MATERIALS AND METHODS

The stages which are applied in this study are shown in Figure 2, respectively.

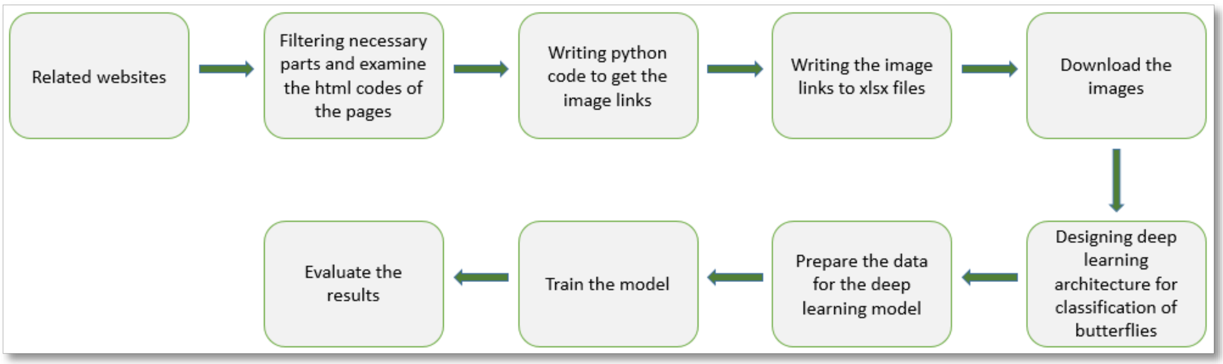
Related websites
Filtering necessary parts and examine the html codes of the pages
Writing python code to get the image links
Writing the image links to xlsx files
Download the images
Designing deep learning architecture for classification of butterflies
Prepare the data for the deep learning model
Train the model
Evaluate the results

Figure 6: The stages that the project goes through

4.1. Dataset

The fundamental brick of a classification architecture is a reasonable, reliable, and a clear dataset. When a model has a well prepared and balanced data, the accuracy changes in a good way.

Turkey has a high butterfly diversity of approximately 380 species. Apart from the researchers, lots of volunteer butterfly watchers wrote their observations into the web sites which are built only for this insect group. Observers in Turkey put their high quality species images into these sites with all other information like date, location, individual number, and share with the platform. These observations have a great importance for conservation purposes of these species.

In this project, we used the pictures in the websites of Butterflies of Turkey, Butterflies monitoring and photography society of Turkey, and Anonymous Butterflies of Turkey to create a dataset [28, 29, 30].

To download the images from the websites and give their names, we wrote python codes. Since there are very few photos in most of the species, we had to eliminate some of the species. We selected the first six species having the greatest number of images among the photos we have and worked on it. As a result, the average number of images per species was around 1000. We enumerated the species in the dataset systematically as can be seen in Table 1. Finally, our dataset includes 7148 photographs of 6 species in total.

Table 1: Species of butterflies which are used in this study and image numbering system

Image Name	Species Name	Species Image
from 100001.jpg to 12039.jpg	*Polyommatus icarus*	
from 200001.jpg to 21054.jpg	*Colias crocea*	
from 300001.jpg to 31026.jpg	*Glaucopsyche alexis*	
from 400001.jpg to 41020.jpg	*Callophrys rubi*	
from 500001.jpg to 51006.jpg	*Lycaena phlaeas*	
from 600001.jpg to 61003.jpg	*Melitaea didyma*	

4.2. Data Preparation

In this project, we created a dataset of 7148 images belonging to six butterfly species. All the data preparation and model codes are written in Python 3.7.7 language. First, we get the file names of the images from the operating system and put them into a variable. Image names in the dataset are prepared according to a rule that each photo name for each butterfly category starts with a specific digit like the ones which starts with 10k are for *Polyommatus icarus* species or 20k for *Colias crocea* (Table 1). Thus, we obtained the data labels by using this rule via separating the category labels from the file names. The butterfly images and the labels were stored in the pandas data frame. The column values of the dataframe and the value counts for each label can be seen in Figure 3.

Table 2: Image counts for each individual category

Labels	Value Counts
1	2039
2	1054
3	1026
4	1003
5	1006
6	1026

We converted label numbers to their species names as “1” for “*Polyommatus icarus*”, “2” for “*Colias crocea*”, “3” for “*Glaucopsyche alexis*”, “4” for “*Callophrys rubi*”, “5” for “*Lycaena phlaeas*”, and “6” for “*Melitaea didyma*”. That is, we converted data frame labels from integer to an object type.

We are now starting to create our x and y variables to fit into our model. Firstly, we made a numpy array named 'y' from the label's values. Data images have different resolutions and backgrounds from each other. We resized their dimensions to 256×256 pixels and turned them into a numpy array. This list became our X variable. According to these X and y values, we plotted some random images to look at the dataset as seen in Figure 3.

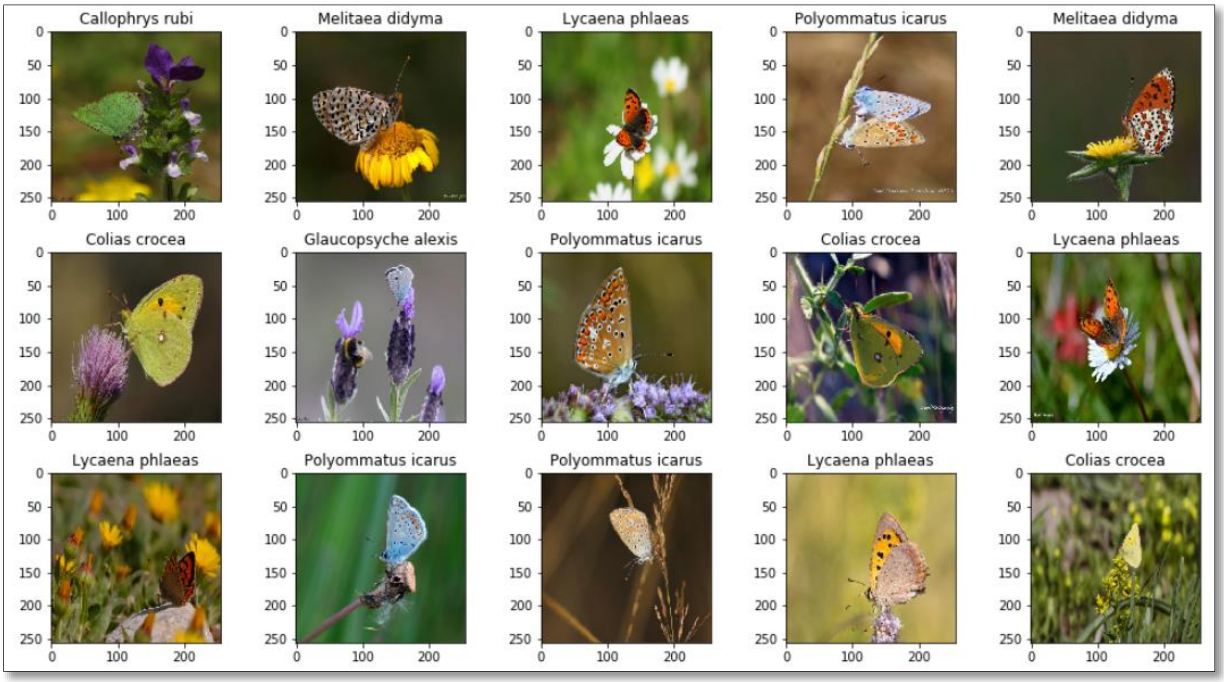

Figure 7: Randomly selected images and their labels from the dataset

To prepare the dimensions to fit into the classification model, we need some conversions on data. To this end, we converted categorical labels into numeric values again by using a dictionary and made a numpy array named 'Y' from the label's values. Then data was encoded by OneHotEncoder function, thus the new format of Y became a csr_matrix. On the other hand, list of X values was converted into a numpy array.

4.3. Model

After the data is prepared, the dataset (X and Y values) is split into training and test parts in the ratio of 80:20. Thus, four variables were constructed as “X_train”, “x_test”, “Y_train”, and “y_test”. Number of images in train and test data are 5718 and 1430.

In the modelling part, we used tensorflow 2.4.0 and keras 2.4.3. Sequential model was used with six convolution layers for feature learning and fully connected layers for classification of images. The structure which is used in this project is depicted in Figure 4.

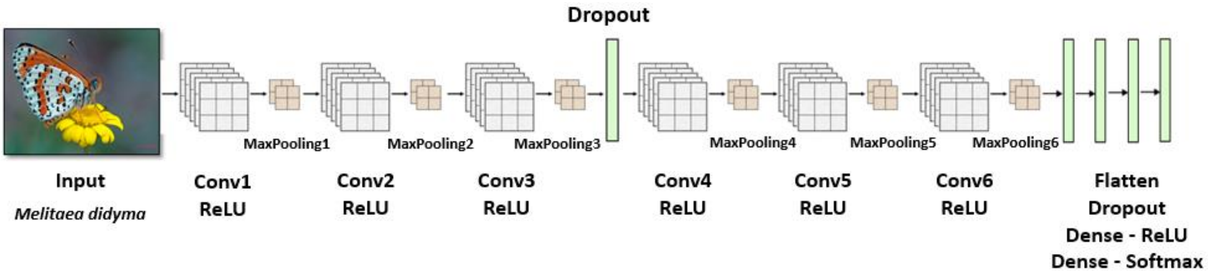

Figure 8: Convolutional neural network (CNN) model used for butterfly classification

As explained in the data preparation part, because the input images differ in size, they all equally were resized to 256×256 pixels and three-color layers (RGB). In the first layer, the images were convoluted by 32 filters in 4×4 kernel size. Because it is more efficient during calculations than sigmoid or tahn functions, here, ReLU is used as an activation function. Then the output of this layer is reduced by 2×2 pooling size. This step is regulating the size of the network and reducing the output size. In the second convolution layer, the same filters with the first one was used. In the third and fourth convolution layers, we used a kernel size of 64; and in the fifth and the sixth layers the size of the kernel was 128. Between 3rd and the 4th layers, we used a Dropout layer to reduce complexity and prevent overfitting. After the convolutional layers, we used fully connected layers. In the Flattened layer, matrices are converted from previous convolutional layers into a one-dimensional array. Then, to reduce complexity and prevent overfitting, we again used the Dropout layer, and then we used 256 neurons in the Dense layer. In the last layer, the output layer, we added six neurons (because of the six categories that we have) in the Dense layer with a softmax activation function. After that model object is created, it was compiled using a loss function (categorical_crossentropy), an optimizer (adam), and with an accuracy metric. Once we compiled the model, we converted Y train into a numpy array and fit the model with X_train and Y_train variables. Number of epochs was set to 100.

5. RESULTS AND DISCUSSION

The working time of the whole model took almost five hours. The final accuracy of the model was 92.73%. During 100 epochs, change of the accuracy and loss are depicted in Figure 5.

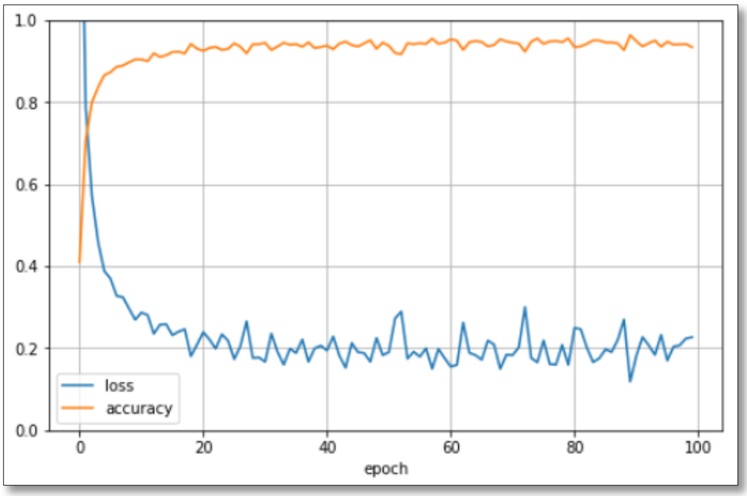

Figure 9: Epoch versus accuracy graph, in which loss and accuracy trends are shown

There were many trials and evaluations conducted before creating the final version of the model. We explained some of the earlier versions below. In all versions, activation functions and pooling sizes were the same, but layers were different in terms of their counts and parameters.

In the first version of the model, there are 6475 butterfly images in dataset. The dataset (X and Y values) is split into training and test parts in the ratio of 60:40. The model has three convolution layers. The images were equally resized to 128×128 pixels and three-color layers (RGB). In the first layer, the images were convoluted by 32 filters in 4×4 kernel size. In the second convolution layer, 64 filters in 4×4 kernel size were used. In the third convolution layers, we used kernel size of 128 and 4×4. In the fully connected layer, we used an extra Dense layer with 100 neurons. Once we compiled the model, fit the model with X_train and Y_train variables. Number of epochs was set to 30. The accuracy of this model was 74.32%.

In the second version of the model, we increased the sample size to 7148. Training and test parts of the dataset were split in the ratio of 80:20. The model has three convolution layers. The images were equally resized to 64×64 pixels and three-color layers (RGB). In the first layer, the images were convoluted by 32 filters in 4×4 kernel size. In the second convolution layer, 64 filters in 4×4 kernel size were used. In the third convolution layers, we used kernel size of 128 and 4×4. In the fully connected layer, we were used an extra Dense layer with 100 neurons. Once we compiled the model, fit the model with X_train and Y_train variables. Number of epochs was set to 100. The accuracy of this model was 78.32%.

In the third version of the model, there are 7148 butterfly images in dataset. The dataset (X and Y values) is split into training and test parts in the ratio of 80:20. The model has five convolution layers. The images were equally resized to 128×128 pixels and three-color layers (RGB). In the first two layers, the images were convoluted by 32 filters in 3×3 kernel size. In the third and the fourth convolution layers, 48 filters in 3×3 kernel size were used. In the fifth convolution layer, we used kernel size of 64 and 3×3. In the fully connected layer, this time we did not used an extra Dense layer. Once we compiled the model, fit the model with X_train and Y_train variables. Number of epochs was set to 100. The accuracy of this model was 89.93%.

In the fourth version of the model, there are 7148 butterfly images in dataset. The dataset (X and Y values) is split into training and test parts in the ratio of 80:20. The model has five convolution layers. The images were equally resized to 128×128 pixels and three-color layers (RGB). In the first three layers, the images were convoluted by 32 filters in 3×3 kernel size. Here, we used an extra Dropout layer with the value of 0.3 to prevent overfitting. In the fourth convolution layers, 48 filters in 3×3 kernel size were used. In the fifth convolution layer, we used kernel size of 80 and 3×3. In the fully connected layer, this time we did not used an extra Dropout layer. Once we compiled the model, fit the model with X_train and Y_train variables. Number of epochs was set to 100. The accuracy of this model was 90.56%.

In the fifth version of the model, there are 7148 butterfly images in the dataset. The dataset (X and Y values) is split into training and test parts in the ratio of 80:20. The model has six convolution layers. The images were equally resized to 256×256 pixels and three-color layers (RGB). In the first two layers, the images were convoluted by 32 filters in 3×3 kernel size. In the third convolution layers, 48 filters in 3×3 kernel size were used. Here, we used an extra Dropout layer with the value of 0.3 to prevent overfitting. In the fourth convolution layer, we used a kernel size of 64 and 3×3. In the fifth convolution layer, we used a kernel size of 80 and 3×3. In the sixth convolution layer, we used a kernel size of 126 and 3×3. In the fully connected layer, this time we did not use an extra Dropout layer. Once we compiled the model, fit the model with X_train and Y_train variables. This time, the number of epochs was set to 150. The accuracy of this model was 92.59%. The model and the accuracy value is shown in Figure 13.

In the sixth and final version of the model, there are 7148 butterfly images in the dataset. The dataset (X and Y values) is split into training and test parts in the ratio of 80:20. The model has six convolution layers. The images were equally resized to 256×256 pixels and three-color layers (RGB). In the first two layers, the images were convoluted by 32 filters in 3×3 kernel size. In the third convolution layers, 48 filters in 3×3 kernel size were used. Here, we used an extra Dropout layer with the value of 0.3 to prevent overfitting. In the fourth convolution layer, we used a kernel size of 64 and 3×3. In the fifth convolution layer, we used a kernel size of 80 and 3×3. In the sixth convolution layer, we used kernel size of 126 and 3×3. In the fully connected layer, this time we did not use an extra Dropout layer with 100 neurons. Once we compiled the model, fit the model with X_train and Y_train variables. This time, the number of epochs was set to back again 100. The accuracy of this model was 93.57%. This is the final version of the model. However, when we tried to run the model one more time, we got our final accuracy result of 92.73%.

We did not change any parameters after getting the final accuracy. As a future study, sample size can be increased by adding extra images, or by the augmentation technique (creating different representations of the same image such as by flipping in several directions). Or, the feature extraction part of CNN can be trained before by a different dataset, thus, an already trained model can continue from the learned parameters (which is the term called transfer learning). Additionally, maybe fully connected layers can be arranged in a different way both for performance, efficiency, and for accuracy of the classification.

CONCLUSION

Many models about automatic identification of butterflies have been tried or proposed in past literature studies and have been compared with each other. Automatic identification of butterflies, especially at an expert level, is needed for important topics such as species conservation studies, agriculture (insect damage on plants), and conservation of biodiversity. A model that can define species even in datasets containing a small number of samples minimizes the need for experts on the subject. By the model proposed in this study, the taxonomic classification of butterflies was studied. CNN applications were made on 7148 photographs of six butterfly species used in the study. 80 percent of the data set was reserved for training and 20 percent for testing, and the model was run with the relevant parameters. At the end of the study, an accuracy degree close to 93 percent was obtained.

In the future, studies can continue to increase this accuracy value by increasing the working performance of the model. For this, the number of layers and / or parameters can be changed, the sampling number can be increased, or techniques such as augmentation can be tried.

In this study, adult butterfly pictures taken in their natural area have been used to identify. The future model can also be applied to egg, larva, or pupa pictures of butterflies if they have enough pictures for training and testing.

Lastly, a web and/or mobile application can be developed for the use of amateur butterfly watchers and curious people, whether for citizen science, hobby purposes, or even for the scientific studies.

REFERENCES

[1] Akçay H. G., Kabasakal B., Aksu D., Demir N., Öz M., and Erdoğan A. (2020). Automated bird counting with deep learning for regional bird distribution mapping. Animals. 10(7), 1207.

[2] Haghiabi A. H., Nasrolahi A. H., and Parsaie A. (2018). Water quality prediction using machine learning methods. Water Quality Research Journal. 53(1): 3-13.

[3] Santangeli A., Chen Y., Kluen E., Chirumamilla R., Tiainen J., & Loehr J. (2020). Integrating drone-borne thermal imaging with artificial intelligence to locate bird nests on agricultural land. Scientific Reports. 10(1): 1-8.

[4] Almond R. E. A., Grooten M., and Peterson T. (2020). Living Planet Report 2020-Bending the curve of biodiversity loss.

[5] Stork N. E., (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?. Annual Review of Entomology. 63: 31-45.

[6] Talavera, G. and Vila, R. (2017). Discovery of mass migration and breeding of the painted lady butterfly Vanessa cardui in the Sub-Sahara: the Europe-Africa migration revisited. Biological Journal of the Linnean Society. 120(2): 274-285.

[7] Robbins R. K. and Opler P. A. (1997). Butterfly diversity and a preliminary comparison with bird and mammal diversity, pp. 69-82. In: M. Reaka-Kudla, D. E. Wilson, and E. O. Wilson [eds.], Biodiversity II: Understanding and protecting our biological resources. Joseph Henry Press, Washington, D.C.

[8] Karaçetin E. and Welch H. J. (2011). Red book of Butterflies in Turkey. Nature Conservation Center, Ankara, Turkey.

[9] Kröse B. and van der Smagt P., 1996. An introduction to neural networks. Amsterdam Press, Amsterdam, the Netherlands.

[10] Macukow B., 2016. Neural Networks - State of art, brief history, basic models and architecture. 15th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM), September 2016, Vilnius, Lithuania.

[11] https://rubikscode.net/2018/02/26/introduction-to-convolutional-neural-networks/ (Data accessed: December 2020).

[12] CS 230 - Deep Learning. https://stanford.edu/~shervine/teaching/cs-230/ (Data accessed: December 2020).

[13] Miao Z., Gaynor K. M., Wang J., Liu Z., Muellerklein O., Norouzzadeh M. S., McInturff A., Bowie R. C., Nathan R., Stella X. Y., and Getz W. M., 2019. Insights and approaches using deep learning to classify wildlife. Scientific Reports. 9(1): 1-9.

[14] Cıbuk M., Budak, U., Guo Y., Ince C. I., and Sengur A., 2019. Efficient deep features selections and classification for flower species recognition. Measurement. 137(2019): 7-13.

[15] Alimboyong C. and Hernandez, 2019. An improved deep neural network for classification of plant seedling images. 2019 IEEE 15th International Colloquium on Signal Processing & its Applications (CSPA 2019), 8-9 March 2019, Penang, Malaysia.

[16] Sardogan M., Tuncer A., and Ozen Y., 2018. Plant leaf disease detection and classification based on CNN with LVQ algorithm. IEEE 3rd International Conference on Computer Science and Engineering (UBMK), pp. 382-385.

[17] Valan M., Makonyi K., Maki A., Vondráček, D., and Ronquist F. 2019. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. Systematic Biology. 68(6): 876-895.

[18] Lim S., Kim S., Park S., and Kim D., 2018. Development of application for forest insect classification using CNN. 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). 18-21 November 2018, Singapore.

[19] Kasinathan T., Singaraju D., and Uyyala S. R., 2020. Insect classification and detection in field crops using modern machine learning techniques. Information Processing in Agriculture.

[20] Thenmozhi K. and Reddy U. S., 2019. Crop pest classification based on deep convolutional neural network and transfer learning. Computers and Electronics in Agriculture. 164, 104906.

[21] Zhu L. Q., Ma M. Y., Zhang Z., Zhang P. Y., Wu W., Wang D. D., Zhang D. X., Wang X. and Wang H. Y., 2016. Hybrid deep learning for automated lepidopteran insect image classification. Oriental Insects. 51(2): 79-91.

[22] Rodrigues R., Manjesh R., Sindhura P., Hegde S. N., and Sheethal A., 2020. Butterfly species identification using convolutional neural network. International Journal of Research in Engineering, Science and Management. 3(5): 245-146.

[23] Arzar N. N. K., Sabri N., and Johari N. F. M., Shari A. A., Noordin M. R. M., Ibrahim S., 2019. Butterfly species identification using convolutional neural network (CNN). 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2019), 29 June 2019, Selangor, Malaysia.

[24] Zhao R., Li C., Ye S., Fang X., 2019. Butterfly recognition based on Faster R-CNN. Journal of Physics: Conference Series, IOP Publishing. 1176(3): 32048.

[25] Liang B., Wu S., Xu K., and Hao J., 2020. Buttery detection and classification based on integrated YOLO algorithm. In: Genetic and Evolutionary Computing (Eds. Pan J. S., Lin J. W., Liang Y., Chu S. C.). ICGEC 2019, Advances in Intelligent Systems and Computing, Vol. 1107. Springer, Singapore.

[26] Theivaprakasham H., 2020. Identification of Indian butterflies using Deep Convolutional Neural Network. Journal of Asia-Pacific Entomology. 17(2): 1456-1462.

[27] Almryad A. A. and Kutucu H., 2020. Automatic identification for field butterflies by convolutional neural networks. Engineering Science and Technology, an International Journal. 23(1): 189-195.

[28] AdaMerOs Türkiye, 2020. Butterflies of Turkey. http://adamerkelebek.org (Data accessed: December 2020).

[29] Kelebek-Türk, 2020. Butterflies monitoring and photography society of Turkey. https://www.kelebek-turk.com/ (Data accessed: December 2020).

[30] TRAKEL, 2020. Anonymous Butterflies of Turkey. http://www.trakel.org/ (Data accessed: December 2020).