**Butterfly Classification Using Machine Techniques**

**Abstract**: In this study, the best model for an Android application was identified by training and testing on a butterfly dataset, which allowed us to compare classic machine learning, deep learning, and transfer learning approaches. The application may determine a butterfly's category by either taking a real-time image of it or selecting one from a collection of pictures. 15 unique species were chosen from the dataset, which included a range of butterfly orientations, camera angles, lengths, occlusions, and backdrop complexity. We discovered an unbalanced class distribution among the 15 selected classes when we studied the dataset, which caused overfitting. To avoid data scarcity and minimize overfitting, the suggested system performs data augmentation. The data models' accuracy is also increased by using the enriched dataset.

**Keywords**: Classification of butterflies, Deep Learning, Transfer learning, data augmentation.

**Introduction**

Butterflies belong to the macrolepidopteran cladeRhopalocera, which also contains moths, in the orderLepidoptera. Adult butterflies have huge, brilliantly colored wings and a fluttering, noticeable flight. The group includes the Papilionoidea superfamily, which includes at least one former group, the skippers (originally the superfamily"Hesperioidea"), as well as the moth-butterflies, according to current research (formerly the superfamily "Hedyloidea"). Butterfly fossils date from around 56 million years ago, during the Paleocene epoch. There are numerous varieties of butterflies. Systematic butterfly names of the world is a reference work that lists the names of 17 families, 47subfamilies, 1690 genera, and 15,141 species of butterflies found worldwide. There are 12 families, 33subfamilies, 434 genera, and 2153 species of Chinesebutterflies to be found in this region. Non-native species such as butterflies have a high ornamental value, but they also play a critical role in maintaining the system's overall stability.



Fig 1. Different types of butterflies

As a result of their short lifespan and ability to respond quickly to changes in the environment, butterflies are particularly effective at detecting subtle ecosystem changes in the environment. Butterflies are extremely useful to scientists who are studying global climate change.... Furthermore, the larvae of most butterfly species are agricultural and forestry pests, which have a direct impact on the living conditions of humans and animals, as well as the availability of food. Research on automatic identification of butterfly species is extremely important not only in the field of species identification, but also in practical work such as environmental protection, pest control in agriculture and forests, and border quarantine, among other things. A fine-grained image classification problem such as butterfly classification is more difficult to solve than a more general/generic image classification problem.

In this study, we concentrate on butterfly classification in specimen photographs. The ecological images feature a complicated background with varying scales and illuminations, and the specimen images have a simple background. The classification of butterflies based on specimen pictures is a complex process to do. The authors use deep reinforcement learning to present a deep-learning-based technique for fine-grained butterfly classification in specimen photos. The Deep Forest and ResNet algorithms were used extensively in this investigation, which resulted in an accuracy of 91 percent.

This paper has been broken into six pieces for your convenience. In Section I, we provide an overview of the subject matter. Section II provides an overview of the research methodology. A further discussion of deep learning-based classifiers, their performances, and datasets was also included in this section. Section III contains the results. Section IV presents a summary of our findings and conclusions from our research. The topic of future work is also included in this section.

**Motivation:**

There are a wide range of kinds of blossoms, and there are an expected 250,000 types of blossoms around the world. Most of individuals regularly stroll by blossoms without having the option to distinguish them. They counsel botanical specialists, go through blossom books, or use watchword look through on the Web to get familiar with the names of these sprouts. By ordering bloom pictures, it is feasible to quickly and basically decide blossom names. It is especially significant given the boundless utilization of portable computerized cameras in this day and age. By adding a discourse to a bloom picture and transferring it to a framework that classifies blossom photos, individuals can be aided their mission to distinguish blossoms. These errands take a great deal of time, and the outcomes are as yet erroneous, particularly when there are a few animal varieties to consider on the grounds that blossom pictures have complex foundations.

Utilizing AI, specialists have as of late exhibited various triumphs in different points in the field of PC vision, including object location, picture division, and picture grouping, among others.

Inspiration is to make a bloom grouping framework with at most exactness and with less computational and tedious technique.

Following the extraction of bloom picture include vectors, the proposed approach arranges the pictures utilizing the Help Vector Machine (SVM), and Stochastic Inclination Plunge (SGD) with most exactness. The order approach is assessed using five different bloom classifications. Because of leading this assessment, the outcomes show that the proposed approach is fit for characterizing the blossom name with a serious level of exactness. A blossom order arrangement of this sort can be applied in different certifiable circumstances. For instance, it tends to be utilized as an intelligent instructive apparatus to further develop learning techniques for both youthful and elderly folks individuals, as well concerning grown-ups.

**Main Contributions and Objectives:**

Dataset collection: Sai Keerthi Potlapally& Shivani Mamindla

Dataset loading and preprocessing coding: Siva Sai Reddy & Deekshith Koride

Algorithm implementation coding: Sai Keerthi Potlapally, Shivani mamindla, Siva Sai Reddy and Deekshith Koride

Documentation: Sai Keerthi Potlapally and Siva sai Reddy

Powerpoint: Shivani Mamindla, Deekshith Koride

Objectives: Objective of this project is to achieve better accuracy with less training images.

**Related Works:**

Zhu et al ordered lepidopteran bug pictures, an integrative district coordinated, and

the double intricate discrete wavelet approach was introduced. They tried their strategy on a

assortment of 100+ lepidopterous bugs from 18 genera, and the acknowledgment precision was found

to be 84.47%. Silva et al. needed to see which highlight determination methodologies and classifiers were

best for recognizing bumble bee subspecies among the seven element pickers and grouping

techniques accessible. In their analyses, they found that the best blend was the

gullible Bayes classifier and the common data extraction of elements.

Wen et al. integrated neighborhood and worldwide data for bug order;

scientists fostered a model that integrates K closest neighbor characterization (KNNC),

normal densities predicated successive order model (NDLC), insignificant level least

straight grouping calculation (MLSLC), the closest mean classifier (NMC), and choice

tree (DT). Their exploratory outcomes showed a 86.6% grouping rate when surveyed on

pictures taken during genuine field catching for preparing.

Kaya et al. proposed two different neighborhood paired designs (LPB) descriptors for identifying

unique surfaces in pictures. The first depends on the distance between the consecutive neighbors

of a middle pixel. Conversely, the second depends on the focal pixel boundary deciding the

neighbors in a similar direction. They utilized lab based photographs of 140 morphos got

in Van, Turkey to assess their descriptors for recognizing butterfly species. The fake brain

network has the best precision of 90.71% in arranging butterflies.

Faithpraise et al. removed underlying data and utilized ANN to build an

programmed recognition calculation for copepod species. They surveyed a general exactness of

93.13 % utilizing seven copepod characteristics from 240 reference photographs. Xie et al. utilized progressed

different undertaking scanty portrayal and various part learning draws near; extra

endeavors were towards a characterization technique to order butterfly symbolism. They applied

the applied way to deal with the preliminary on 24 predominant cultivating frameworks species and differentiated

it to some more up to date draws near.

Feng et al, depending on fold highlights of butterfly photographs, further developed a bug species

acknowledgment and recovering strategy. Their recuperation strategy is based upon the CBIR structure, which likewise doesn't need a proper affirmation yet furnishes clients with a scope of

matches. Abeysinghe et al. utilized multi-facet Siamese organizations to fabricate a totally

automated framework for distinguishing snake species. Despite the fact that the first snake dataset

is restricted, they could achieve reliable outcomes. Alsing et al. made real CNN foundation in light of area variation to recognize post-it areas and afterward changed over these

calculations to be utilized on cell phones and tablets. The quickest RCNN ResNet50 structure

had the most elevated Guide (mean normal accuracy) of 99.33%, however it took 20,018 milliseconds

to derive.

**Data Description:**

Datasets

Preparation, followed by testing There are 75 butterfly species included in this validation data set. All of the photographs are in jpg format and measure 224 x 224 x 3 pixels. 9285 photos have been divided into 75 subdirectories, one for each species in the train set. 375 photos have been divided into 75 subdirectories, with 5 test photographs per species in the test set. The valid set consists of 375 photos that have been partitioned into 750 subdirectories, with 5 validation images per species in each of these subdirectories. Each row in the butterflies.csv file represents one of the images in the dataset. The butterflies.csv file is divided into three columns, each of which has 10,035 rows. File paths, label names, and data set names are the columns to be filled in. The relative path to the image is stored in the file paths column. The species label linked with the image file is stored in the labels column.



Fig 2. Dataset Sample

Data Preparation

It is critical for a Machine Learning Engineer to spend time preprocessing or purifying data before constructing a model from scratch, and the vast majority of Machine Learning Engineers devote a significant amount of time and effort to this portion of their job. A few examples of data pre-processing techniques include outlier detection and treatment, missing value treatment, and the elimination of undesirable or noisy data, to name a few. Before photos are utilized for model training and inference, they must first be formatted. Image preprocessing is the process of formatting images before they are used for model training and inference. Among the things you can do is resize images, rotate them, and make color changes, to name a few.

Diagram

Description automatically generated

Fig 3. Image processing

Modeling

The texture descriptors extracted from the butterfly photos serve as the basis for the method of classification method we will use in this project for classifying images is machine techniques. Our approach may be broken down into four major steps, which can be separated as follows:

Diagram

Description automatically generated

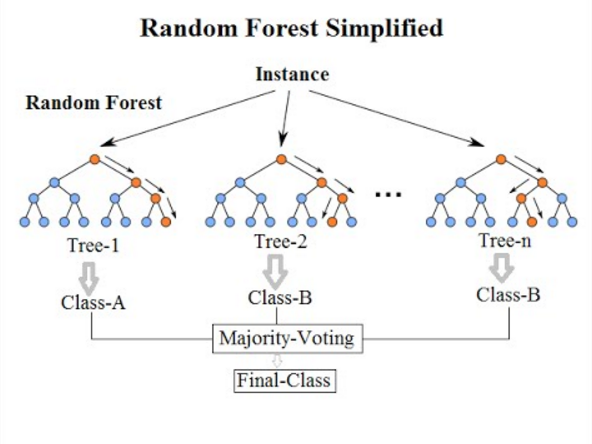
**Proposed Framework:**

Logistic Regression:

Logistic regression would be another method in machine learning adapted from the field of statistics. Logistic regression predicts the probability of two possible consequences for supervised learning. Supervised learning, it is an extension of the linear regression model. The weight interpretation in logistic regression varies from its weight representation in linear regression since the consequence in logistic regression is a possibility of 0 and 1. The weights no longer uniformly impact the expectation. The weighted sum is converted to a possibility even by logistic function. Therefore, we need to reformulate the equation for the interpretation so that only the linear term is on the right side of the formula.

Random Forest Classifiers

The Random Forest Classifier is an ensemble method using a collection of decision trees. They are commonly applied on weak features to reduce the problem that decision trees have in over-fitting. It uses bagging or a random selection of training examples together with a random selection of attributes for each tree. Ensembling the trees reduces the resulting error improving performance using the meta-model.



Support Vector Machines

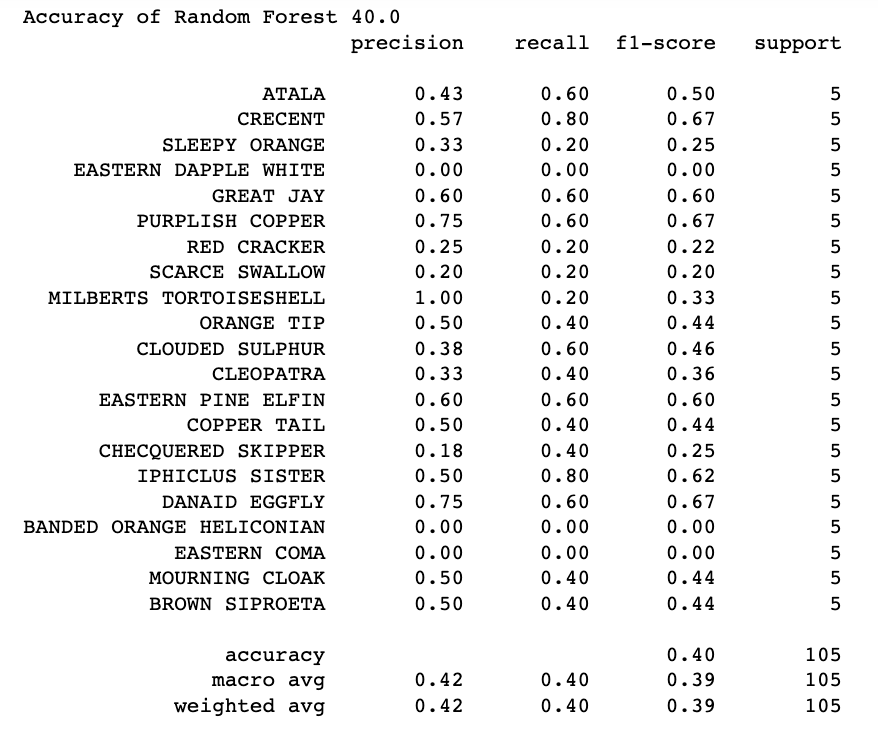
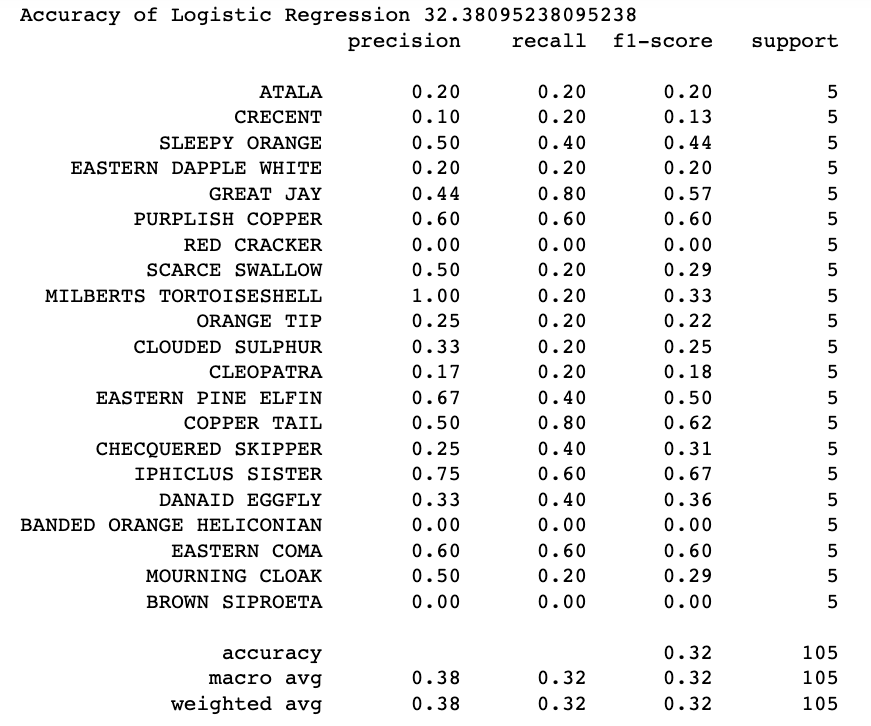
Support Vector Machine is a supervised learning method focused on the philosophy of probability, informatics, and computational geometry. It is used to discover hyper-plane or maximal width, distinguishing two types. The goal is to find the width margin which minimizes the classifier's generalization error. It requires a central collection of points helping to define boundaries. The line-forming data points are called support vectors since they support the boundary. One kernel could be used to differentiate the information sequentially.

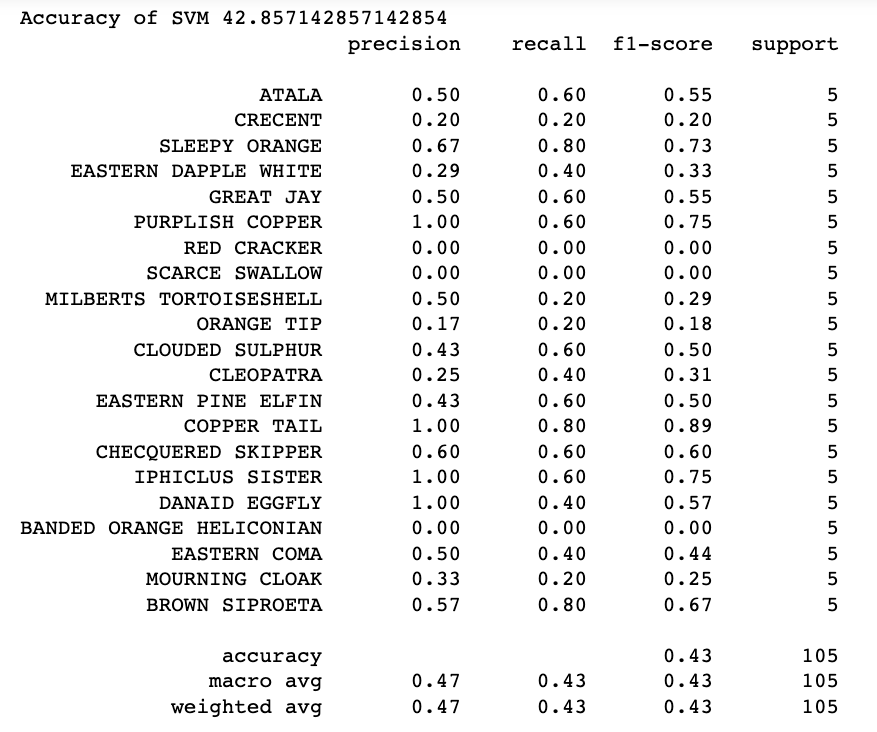
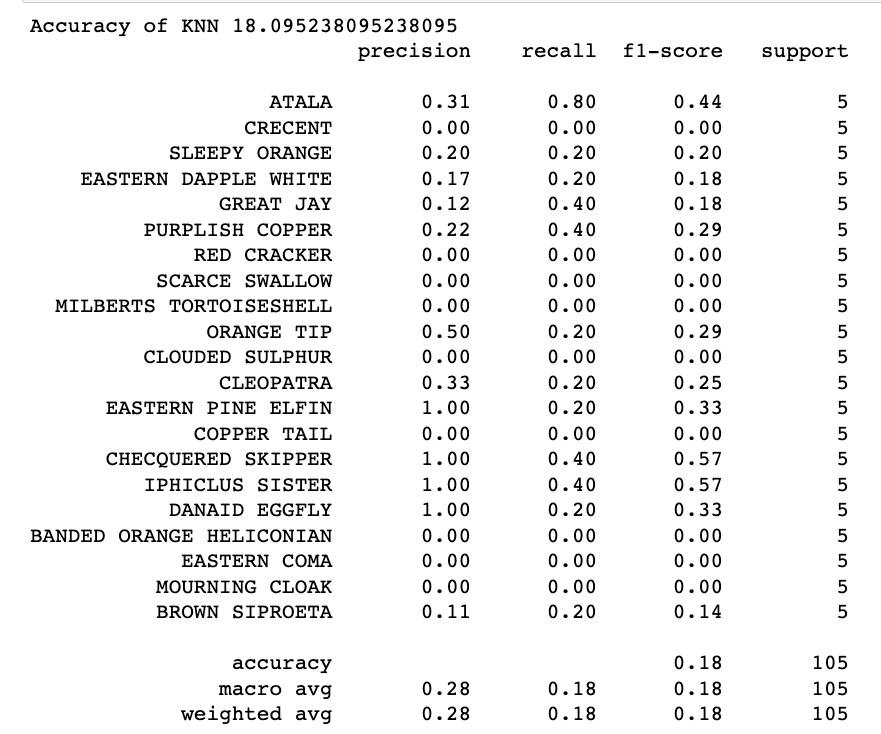
K-nearest neighbour

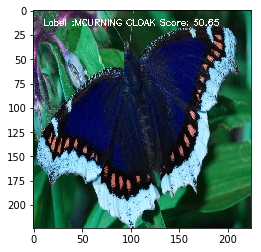
The easiest categorization in image space is the nearest neighbor classification. Under such a system an image is identified in the training set while allocating it to the nearest level mark in the learning set, where distance is computed in image space. The KNN algorithm has been one of the methods that can effectively be used in supervised learning for either regression or classification. KNN is called memory-based or incompetent learning since it preserves the data from the training instances. The entire training subset is "memorized" and when it is appropriate to identify unlabeled instance records, the input properties of the new unlabeled records are compared with the complete training collection to find the nearest match.

Validation Method

These are the key metrics in a classification challenge that are discussed in the classification report. Precision, recall, f1-score, and support will all be available for each class. When we say recall, we mean "how many elements of this class you find in relation to the total number of elements of this class." The precision will be defined as "the number of people accurately identified among that class." It is the harmonic mean of precision and recall that is used to calculate the f1-score. The support for a given class is the number of instances of that class in your dataset (for example, you have 37.5K instances of class 0 and 37.5K instances of class 1, which is a very well-balanced dataset).

**Results**





Annotated test images

Conclusion

In this research, a field-based dataset was built utilizing butterfly photos collected from nature that were categorized by expert entomologists and used to create the dataset. Different machine learning techniques are used to classify the input photos of butterflies, which eliminates the need for any feature extraction methods The outcomes of four distinct algorithms are compared and evaluated, and the results of the experiments are compared and evaluated.

Despite the fact that the photos have several flaws, such as the position of the butterflies, the shooting angle, the distance between the butterflies, occlusion, and backdrop complexity, roughly 50 percent success was achieved in the SVM model for both test and training data. Both test and training data were successfully used in the SVM model, which obtained a 35 percent success rate.

Our next research will include the development of a mobile application that will make use of the pre-trained model that was proposed for this study.

**References**

Arzar, N.N.K.; Sabri, N.; Johari, N.F.M.; Shari, A.A.; Noordin, M.R.M.; Ibrahim, S. Butterfly species identification using Convolutional Neural Network (CNN). In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 29 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 221–224.

. Li, C.; Zhang, D.; Du, S.; Zhu, Z.; Jia, S.; Qu, Y. A butterfly detection algorithm based on transfer learning and deformable convolution deep learning. Acta Autom. Sin. 2019, 45, 1772–178

Wang, W.; Zhang, J.; Wang, F. Attention bilinear pooling for fine-grained classification. Symmetry 2019, 11, 1033.

Andrian, R.; Maharani, D.; Muhammad, M.A.; Junaidi, A. Butterfly identification using gray level co-occurrence matrix (glcm) extraction feature and k-nearest neighbor (knn) classification. Regist. J. Ilm. Teknol. Sist. Inf. 2019, 6, 11–21.

Shou, J.; Zhou, Y.; Li, Y. Systematic Butterffly Names of the World; Shaanxi Science and Technology Press: Xi’an, China, 2006

Wang, W.; Zhang, J.; Wang, F. Attention bilinear pooling for fine-grained classification. Symmetry 2019, 11, 1033.

Andrian, R.; Maharani, D.; Muhammad, M.A.; Junaidi, A. Butterfly identification using gray level co-occurrence matrix (glcm) extraction feature and k-nearest neighbor (knn) classification. Regist. J. Ilm. Teknol. Sist. Inf. 2019, 6, 11–21.

Li, F.; Xiong, Y. Automatic identification of butterfly species based on HoMSC and GLCMoIB. Vis. Comput. 2018, 34, 1525–1533

Xue, A.; Li, F.; Xiong, Y. Automatic identification of butterfly species based on gray-level co-occurrence matrix features of image block. J. Shanghai Jiaotong Univ. 2019, 24, 220–225.

Kartika, D.S.Y.; Herumurti, D.; Yuniarti, A. Local binary pattern method and feature shape extraction for detecting butterfly image.

Kartika, D.S.Y.; Herumurti, D.; Yuniarti, A. Butterfly image classification using color quantization method on hsv color space and local binary pattern. IPTEK J. Proc. Ser. 2018, 78–82.

Pinzari M., Santonico, M., Pennazza, M. et al., "Chemically mediated species recognition in two sympatric Grayling butterflies: Hipparchia fagi and Hipparchia Hermione (Lepidoptera: Nymphalidae, Satyrinae)", PloS ONE Accelerating the publication of peer-reviewed science, 13(6): 1-14 (2018).

Feng, L., Bhanu, B., Heraty, J., "A software system for automated identification and retrieval of moth images based on wing attributes", Pattern Recognition 51: 225-241 (2016).

Zhou, A.; Ma, P.; Xi, T.; Wang, J.; Feng, J.; Shao, Z.; Tao, Y.; Yao, Q. Automatic identification of butterfly specimen images at the family level based on deep learning method. Acta Entomol. Sin. 2017, 60, 1339–1348.

Arzar, N.N.K.; Sabri, N.; Johari, N.F.M.; Shari, A.A.; Noordin, M.R.M.; Ibrahim, S. Butterfly species identification using Convolutional Neural Network (CNN). In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 29 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 221–224.

. Li, C.; Zhang, D.; Du, S.; Zhu, Z.; Jia, S.; Qu, Y. A butterfly detection algorithm based on transfer learning and deformable convolution deep learning. Acta Autom. Sin. 2019, 45, 1772–178

Zhou, A.; Ma, P.; Xi, T.; Wang, J.; Feng, J.; Shao, Z.; Tao, Y.; Yao, Q. Automatic identification of butterfly specimen images at the family level based on deep learning method. Acta Entomol. Sin. 2017, 60, 1339–1348.

**GITHUB LINK:**

<https://github.com/KeerthiPotlapally/MachinelearningProject>