VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Project Report on

Patang Abhidhani - A comprehensive survey on skipper butterflies for Lepidopterists and butterfly enthusiasts

In partial fulfillment of the Third Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2020-21

Submitted by
Abhijit Thikekar - D12A-68
Aniket Pawar - D12A-49
Saurav Telge - D12A-67

Project Mentor
Dr. Sharmila Sengupta

(2020-21)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
Department of Computer Engineering

Certificate

This is to certify that Abhijit Thikekar, Aniket Pawar, Saurav Telge of Third Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the mini project on "**Patang Abhidhani - A comprehensive survey on skipper butterflies for Lepidopterists and butterfly enthusiasts**" as a part of their coursework of Mini Project for Semester-VI under the guidance of their mentor ***Prof. Dr. Sharmila Sengupta*** in the year 2020-21.

This mini project report entitled ***Patang Abhidhani - A comprehensive survey on skipper butterflies for Lepidopterists and butterfly enthusiasts*** by ***Abhijit Thikekar, Aniket Pawar, Saurav Telge*** is approved for the degree of **T.E Computer Engineering.**

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date: 15/05/2021

Project Guide: Internal and External
Dr. Sharmila Sengupta (internal)
Mr. Nikhil Bhopale (external)

Industrial project certificate:

GREEN WORKS TRUST

For a greener and sustainable future

To,

Dr. Nupur Giri
Head of Department of Computer Engineering and Professor,
Vivekanand Education Society's Institute of Technology,
Chembur, Mumbai- 400071.

Date: 27/11/2020
Ref No: 027/2020

Subject – Collaboration letter

Respected Dr. Nupur Giri,

This letter is to confirm the collaboration between *Green Works Trust* (*GWT*) and *Vivekanand Education Society's Institute of Technology* (*VESIT*) students on a project to develop AI to study the Hesperiidae (Skipper) Family of Butterflies.

Our common goals and objectives were identified as:

Goals

1. A cross-platform Website for Hesperiidae (Skipper) Family of Butterflies.
2. Detailed analysis of the Butterflies in a visual format based on the distribution of the Skipper Family of butterflies.
3. Image Recognition to identify the Butterflies species of Hesperiidae integrated in the website.

Objectives

1. Extracting data from various resources such as books, websites, social media and journal papers.
2. Collecting information about the genus species and subspecies present in Hesperiidae Family of butterflies.
3. Generating heatmaps and graphical representations for better understanding.
4. AI model for Skipper family recognition and predicting the present and future diversity of the Hesperiidae family of butterflies.

GREEN WORKS TRUST
For a greener and sustainable future

This project will immensely contribute to the ongoing research by lepidopterists on Hesperiidae Butterflies and will also benefit the students to get a different perspective of the technologies they will use as well as increasing awareness for the ecosystem.

Realtime use of this project:

This project should benefit researchers and other stakeholders. It should not be limited till end of the academic year. Hence to ensure its continuity, this project will be handed over to *Green Works Trust* with all necessary permissions procured / received for the project. *Green Works Trust* or associate organisations can freely use this app for awareness and conservation purpose in future.

Details of the team involved:

Mentors:

- Dr. Sharmila Sengupta
- Mrs. Priya R.L.

Team Members:

- Yash Mate (B.E.)
- Gaurav Tirodkar (B.E.)
- Neelam Somai (B.E.)
- Gayatri Patil (B.E.)
- Saurav Telge (T.E.)
- Abhijit Thikekar (T.E.)
- Aniket Pawar (T.E.)

Thank you and regards,

Nikhil Bhopale
(Founder and Managing Trustee, Green Works Trust)

Mini Project Report Approval
For
T. E (Computer Engineering)

This mini project report entitled **Patang Abhidhani - A comprehensive survey on skipper butterflies for Lepidopterists and butterfly enthusiasts** by Abhijit Thikekar, Aniket Pawar, Saurav Telge is approved for the degree of T.E Computer Engineering.

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place: Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)
Abhijit Pradeep Thikekar - 68
(Name of student and Roll No.)

(Signature)
Saurav Sunil Telge - 67
(Name of student and Roll No.)

(Signature)
Aniket Ashok Pawar - 49
(Name of student and Roll No.)

(Signature)

(Name of student and Roll No.)

Date:15/05/2021

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Dr. Sharmila Sengupta** (Project Guide) for her kind help and valuable advice during the development of the project synopsis and for her guidance and suggestions.

We are deeply indebted to the Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support, and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Computer Engineering Department

COURSE OUTCOMES FOR T.E Mini Project

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge, and skills to analyzing user-submitted the project.
CO 2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources, and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety, and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents, and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced, and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment, and platforms for creating innovative solutions for the project.

Abstract

Skippers (Family: Hesperiidae) representing a fifth of the world's butterfly fauna with ca. 4000 species under 567 genera. The need to keep track of butterfly hotspots has become increasingly important due to climate change, which could account for their population decline. Also, very few researchers have been involved in studying skippers because they are so similar to moths. It's also difficult to distinguish and categorize them due to their similarity.

The objective of this project is to provide one authentic source of knowledge about the skipper butterfly in India to new aspiring researchers. This has been done through a comprehensive website- "Patang Abhidhani" which will be described in the following sections.

There is a lack of an integrated and regularly maintained data dictionary with skipper butterfly-related information. The data gathering process was very exhaustive as information was not readily available and reliable sources too were short of data. Hence data preprocessing involved cleaning and analyzing the data before applying an image classification model. This will allow the Lepidopterists and butterfly enthusiasts to upload relevant images to the website and automatically populate the map with information about skipper butterflies. The result will be a fully integrated web application along with machine learning features.

INDEX

Chapter 1: Introduction

1.1 Introduction:

The biodiversity available in India is quite vivid. It is home to a plethora of animals, plants, insects, and birds. Due to its tropical location, organisms find it easy to flourish in India. However, due to deforestation and excessive farming, and urbanization, the green cover of the nation has been steadily decreasing. As a result, most species are losing their habitats.[1] Butterflies are especially affected by this.

Due to their sensitivity to minor changes to the environment and decreasing availability of flowers for pollen, the populations have been seen migrating to different places. Thus, they act as indicators of environmental change. They are also an integral part of our ecosystem. They are needed to pollinate plants to acquire fruits and vegetables. Butterflies are some of the most biodiverse insects on the planet.

However, in recent years, pollution has gripped our planet leading to a decrease in variety and a count of various fauna. One such affected avian species is butterflies. Hence, the preservation and categorization of butterflies have become even more important. Among all the butterfly species available in India, the one most neglected is Hesperiidae (Skippers).

Skippers are a family of Lepidoptera (moths and butterflies) named Hesperiidae. Since they are diurnal, they are often referred to as butterflies. They were historically classified as a separate superfamily, Hesperioidea; however, the most recent taxonomy positions them in the superfamily Papilionoidea, the Butterfly superfamily, confirming their status as Butterflies. Their fast, darting flight habits have earned them the name. The antenna tips of the majority have been modified into narrow, hook-like system ions. Furthermore, most skippers lack the wing-coupling structure found in most moths. They can be found worldwide, but the Neotropical regions of Central and South America have the most variety.

The taxonomic status and the phylogenetic position of skippers (Hesperiidae) within Lepidoptera remains a controversial issue. Hesperides' knowledge, geographical distribution, immature stages, larval foodplants, and phylogenetic relationships remain poor compared to the other butterfly families. Since skippers have many similarities with moths, very few researchers have been interested in studying them. Also, due to their similarities, it is difficult to identify and

properly categorize them. However, all subspecies of skippers do not look like moths. Looking at all the members of its family, we also notice some of its sub-species to show beautiful colors, which also make them hard to differentiate. Due to the wide range of features that these butterflies possess, many people find it difficult to categorize them. [12]

The website, “Patanga Abhidani” would exhibit detailed information including gallery, heatmap, etc as well as have an option to upload images on Skipper butterflies. Therefore, enthusiasts and common people can further research this family. The user would upload images and details of the butterfly, which would then be sent to the machine learning model to classify it. If verified, the data is then sent to a database, and the results are reflected in the heatmap, which also shows the locations of the butterflies.

Using this data, naturalists, photographers, and enthusiasts can figure out where exactly skipper butterflies are prevalent to figure out changes in their living patterns and gather more data regarding particular sub-species. This also allows us to record various instances of a butterfly species existing at different places based on which we can better document and understand the species.

1.2 Motivation:

The primary reason for research on this topic was that Skipper butterflies aren’t well documented in India. Due to rapid urbanization, butterfly counts have been receding and it is important to track the migration patterns of these butterflies to prevent them from going extinct. The proposed system would accomplish this by gathering user submitted data and verifying its validity. This would enable new researchers to get information about skippers without having to go through the plethora of websites online. It would also work as a good platform for users to share photographs clicked by them and hence increase the total data available online.

1.3 Problem definition:

The primary objective of our project is to document skipper species of butterflies that exist in India and to create a user-friendly method for common people and researchers to access the data regarding the same. This will be done using a web application with heatmaps and an image recognition model for verification purposes. The goals of this project are:

1) Species data dictionary:

To create a dictionary documenting all species of skipper butterflies available in India based on user content. All the data in this dictionary would be acquired from the users. Any user would be able to access this data to research the butterfly species or better understand their habitats or acquire their location or hotspots.

2) Species identification:

The system would possess the ability to automatically recognize skipper butterflies from images uploaded by users. The purpose of this is to remove the human verification aspect from the process. Image verification can be done using an ML model so that incorrect data isn't added to the database.

3) Interactive web application:

The whole project would be packaged as a web application that would contain an interactive heat map that users can see to understand the species density at various places in India. It would also provide an upload feature so that user-generated content can be added to the database and hence increase the scale of the project.

1.4 Existing systems:

Currently, web applications to document species of butterflies exist. However, these websites cover all the species of butterflies and aren't focused on just skipper butterflies. Also, these websites do not possess the ability to automatically verify if the butterfly in the image is of the said species or not. As a result, these websites are run by professionals and experts in the field of Entomology and Lepidopterology i.e people who are trained at recognizing species of butterfly species.

1.5 Lacunas in the current system:

1) Non-existent automated models

The major shortcoming of today's systems is that there isn't a system that can automatically detect a butterfly species. The current systems rely on manual verification of the butterfly species by an expert or trust-based systems where what enthusiasts and photographers say is considered valid.

2) Image data unavailability

The second issue is the unavailability of species-specific data. For the process of detection to be automated, the model would need many images. Although butterfly data is available in plenty, images of skipper butterflies found in India are quite a few and far apart. As a result, the database regarding this is quite small.

3) No platform for normal people to get their photos verified

The websites related to butterflies that exist today do not allow any common person to upload images to their database. You need to contact the administrators and show credentials for your images to be featured on their website. The other alternative is Facebook groups which are a lot more chaotic and disarrayed.

1.6 Relevance of the project:

The system will be run by the non-profit Greenworks trust with whom we are affiliated. The system would be open for all to access and add content to it. It would cover the entire nation, which will in turn help people to understand the migration pattern of the species and to understand their concentrations, and figure out their habitats based on it.

This project will be useful for:

1) Lepidopterists
2) Butterfly enthusiasts
3) Nature photographers
4) Naturalists
5) Scientists and researchers

6) Entomologists

1.7 Methodology employed for deployment:

Step 1: Searching for reliable data sources

Step 2: Extracting data from relevant sources

Step 3: Dataset preparation- Cleaning the data to fill missing values and acquire latitude and longitude for each data element.

Step 4: Apply machine learning models classify images as skipper and non-skipper

Step 5: Train on the acquired dataset to generate a deployable ML model

Step 6: Test the model and acquire various performance metrics to further improve the model

Step 7: Generate a heatmap using a database and integrate it with the machine learning model

Step 8: Create a web application to display data and act as the frontend for user interaction.

Step 9: Test the system for bugs and errors.

Chapter 2: Literature survey

A. Brief overview of Literature Survey:

Various butterfly recognition methods have been developed before with varying degrees of success. Each method is specialized for a different purpose. Ecological research on the Hesperiidae family has also been conducted which shines some light on the best methods to identify and categorize skippers as a whole. The most relevant papers and patents pertaining to our project have been surveyed and their observations are listed below.

B. Related works:

Butterfly documentation systems similar to our system already exist. They attempt to encompass all butterfly species while our system focuses only on one specific butterfly family.

1) ifoundbutterflies.org:

This is the largest butterfly database in India. It covers all the species of butterflies found in India. It is an online peer-reviewed resource designed to disseminate comprehensive information on various aspects of the biology of Indian butterflies, encourage their observation to study their natural history and ecology, gather population and distributional data in a centralized database, and spread awareness about their conservation.

2.1 Research Papers Referred:

1) Identification of Indian butterflies using Deep Convolutional Neural Network[1]

Abstract: The common butterfly distinguishing strategy depends on the diverse characters possessed by butterflies, specifically wing-venation, shading, shape, designs. However, this analysis can prove to be quite costly and tedious. To overcome above-mentioned problems, a butterfly identification system is proposed which has been trained to immediately identify the butterfly species with high precision. The authors propose Deep Convolutional Neural Network (D-CNN) butterfly classifier models utilizing eleven pre-prepared designs specifically ResNet-18, ResNet-34, ResNet-50, ResNet-121, ResNet-152, Alex-Net, DenseNet-121,

DenseNet-161, VGG-16, VGG-19 and SqueezeNet-v1.1 The diverse model's characterization results accomplished a greatest top-1 accuracy(94.44%), top-3 accuracy(98.46%) and top-5 accuracy(99.09%) utilizing ResNet-152 model. The outcomes recommend that models can be decisively used to distinguish butterflies in India.

Inference: This study shows that butterfly classification using D-CNN is possible using pre-trained networks as the basis. It severely reduces the manpower required in the butterfly recognition process and also classifies them with high accuracy.

2) Investigations of butterfly species identification from images in natural environments.[2]

Abstract: Faster R-CNN is the bleeding edge of neural networks. However, when applied to butterflies, the accuracy wasn't good enough and the process was highly time-consuming. The authors propose a new partition and augmentation technique for highly unbalanced and skewed datasets. They found out that RetinaNet is an efficient pre-trained network that can efficiently classify butterflies with a high degree of accuracy (80% MAP). This is the best alternative that they found to Recurrent CNNs.

Inference: R-CNNs are not the best networks to be used in classifying butterflies as they are not built for it. Instead, pre-trained networks show promise in this field and can be used in conjunction with added layers to exhibit high accuracy and at the same time keep training time within acceptable limits.

3) A study on butterfly diversity in Singur, West Bengal. [3]

Abstract: A study was conducted between March 2015 and November 2016 to analyze the butterfly diversity in West Bengal, India. Through this survey/study, the authors found out that there were a total of 69 species of butterflies belonging to 5 families in the state. Nymphalidae was the most dominant family with 22 species, followed by Lycaenidae. 12 species of the Hesperiidae family were also found along with a few butterflies of Pieridae and Papilionidae families. 5 of the 69 species are protected under the Indian wildlife protection act.

Inference: The butterfly diversity in west Bengal is quite expansive with a large variety of species spanning 5 families. However, Hesperiidae (Skipper) butterflies are not the most dominant family here.

4)Ten genes and two topologies: an exploration of higher relationships in skipper butterflies (Hesperiidae) [4]

Abstract: Although multiple attempts have been made to deduce the genealogical relationships of various skipper butterflies, there persist some uncertainties regarding their deep clade relationships. A recent genealogical analysis using 30% of the known genera belonging to Hesperiidae reconstructed the higher-level relationships using a rich sampling of genetic markers (mitochondrial and nuclear). The study found that 2 contrasting topologies exist among skippers however neither is strongly supported and hence it was concluded that there is insufficient evidence to resolve these higher-level relationships. Nevertheless, contemplating morphological characters, the authors propose that one of the topologies is more probable.

Inference: It is difficult to distinguish between skippers based on their genetic makeup and hence no physical features can be deduced to accurately distinguish between every species belonging to the Hesperiidae family. As a result, we cannot rely on only physical features of the butterflies like their size to uniquely identify them.

5) Fine-Grained Butterfly Classification in Ecological Images Using Squeeze-And-Excitation and Spatial Attention Modules. [5]

Abstract: Butterfly classification is a fine-grained classification issue that is considerably more complex than normal image classification. Butterfly photos are of 2 types: Ecological and specimen. Most photos of butterflies are captured in their natural habitats (ecological) with a lot of noise present in the image due to it being captured outdoors. However, current studies are conducted using specimen samples (indoor). As a result, models trained using this data do not scale well when compared to outdoor images. Hence, the authors suggest a new classification network consisting of dilated residual networks, squeeze-and-excitation (SE) module, and spatial attention (SA) module. The SA module works on long-range dependencies in the image and the SE module enhances useful information while getting rid of useless information. This approach leads to a higher F1 score than normal CNN models.

Inference: Detecting butterflies in ecological images (images clicked outdoors) is difficult due to the amount of data present in a normal image. Instead, the images can be preprocessed using SA and SE to extract useful information from them, getting rid of background noise in the process.

6) Butterfly Species Recognition Using Artificial Neural Network.[6]

Abstract: The authors explore the automation of butterfly recognition using artificial intelligence. The patterns on butterfly wings are used as the primary parameter for determining the species of the butterfly. A local binary pattern descriptor is used after removing the background of the image to acquire a histogram of the image information which is computed. After this, an artificial neural network can be used to classify the images. The initial study was done using 2 butterfly species and an accuracy of around 90% was achieved for one species while 100% for the other.

Inference: Through this paper, it is clear that automation of the butterfly recognition process is entirely possible using artificial neural nets. Also, binary classification using this method can yield promising results. As our model is also binary, it sets a good precedent for training models.

2.2 Patent search:

Butterfly recognition systems exist all over the world and have been implemented in various projects. Some of them have been patented for commercial use. A few of the patents related to our project are:

1) Butterfly identification network construction method and apparatus and computer device and storage medium

Applicants: Ping An Tech Shenzhen Co Ltd [CN]

Inventors: Liu Aozhi [CN]; Wang Jianzong [CN]; Xia Zimin [CN]; Xiao Jing [CN]

Link:https://worldwide.espacenet.com/patent/search/family/064818940/publication/WO2020006881A1?q=pn%3DWO2020006881A1

Methodology used:

i) Resampling an original butterfly image to obtain a target butterfly image for training

ii) If the loss function value of the first convolutional layer of the capsule network is less than the threshold then move on to the next then move to the next layer or else, use backpropagation to update the previous capsule neurons.

iii) This is done using till the loss is less than a threshold value.

iv)This is done so that the same butterfly species from different angles of view can be distinguished, and the accuracy of butterfly identification by the capsule network is improved.

2) A butterfly identification method based on a neural network and a related device

Applicants: Ping An Tech Shenzhen Co Ltd [CN]

Inventors: Wang Jianzong; Wang Yiwen,Zhang Shuang

Link:https://worldwide.espacenet.com/patent/search/family/066925947/publication/CN109886295A?q=pn%3DCN109886295A

Methodology used:

i) Collecting large amounts of butterfly images corresponding to each butterfly species that is to be studied and classifying them and storing them accordingly.

ii) The images are input into the target detection YOLO algorithm network. The YOLO network is trained which produces a feature map of the butterflies in a 5*5 grid.

iii) Obtain an image that is to be recognized using the YOLO network. Pass it through the network and acquire S×S (B×5+C) dimension data of the image to be identified, where B is a feature map and C is predicting the number of butterfly species.

iv)Determine the butterfly type and location information of the grid of the image to be identified according to the S×S (B×5+C) dimension data, and display the butterfly category to the user.

3) A butterfly automatic classification method based on depth learning

Applicants: Univ Shantou

Inventors: Fan Zhun; Huang Longtao; Lu Jiewei; Mo Jiajie; Wu Yuming; Zhu Guijie

Link:https://worldwide.espacenet.com/patent/search/family/065404506/publication/CN109376765A?q=pn%3DCN109376765A

Methodology used:

i) a photo of a butterfly in a natural environment is collected; part of the collected butterfly photos are labeled by manual labeling, and a picture library with labeling box and classification label is constructed as a training sample set

ii) Faster-RCNN algorithm is used to train a convolutional neural network for butterfly location in photographs. A deep convolutional neural network algorithm is used to train a convolution neural network for butterfly classification.

iii) For recognizing butterflies in a natural environment, the butterflies are located using the CNN, and then it is classified using the Faster-RCNN algorithm efficiently and accurately.

2.3 Inferences drawn

From the papers studied above, we found out that butterfly recognition is a complex process due to the minute interspecies variations. The Hesperiidae family of butterflies do have a lot of species under it and their genetic makeup and physical fordimensions are not sufficient to categorize them properly. Instead, we have to rely on the colors and patterns on their wings and hence, the model has to be completely image data-oriented. All the required features would be extracted from images and the model would be trained based on that data. R-CNN and basic CNNs are not efficient or accurate enough to categorize them. Hence, we would need deep neural networks and would have to rely on transfer learning techniques using previously trained networks like GoogleNet, RESNET, VGG-16, etc. The images need to be processed in such a way that background data is as muted as possible so that only the important features of the butterfly can be extracted.

2.4 Comparison with the current system

Current systems	Our system
Do no specifically focus on the Hesperiidae family	The sole purpose is to document and display data regarding the Hesperiidae family
Do not have automatic butterfly recognition systems deployed on them	Will have methods to verify if a butterfly is of skipper species or not

User data is not accepted	It has a feature that allows users to upload images to the database
Human verification for each image necessary	Human verification for each image is not necessary. Only edge cases require human verification

Chapter 3: Requirement Gathering for the Proposed System

3.1 Definition of requirement gathering

Requirements Gathering is an exploratory process that generates a list of requirements (functional, system, technical) from the various stakeholders/entities involved (customers, users, vendors, IT staff, etc.) that will be used as the base for defining formal requirements definition. It consists of interviews, discussions, observations, etc.

3.2 Functional requirements:

1. **Fully functional website:**

   A fully functional website consisting of all the proposed features is listed below. The website should be deployable with front-end as well as backend services in place. It should have a database connected to it to make fetch and push queries.

2. **A dedicated page to show area-wise details of the butterfly population (i.e. map):**

   The website should have a section that works as an interactive map. In this, all the skipper butterflies in the database would be displayed as location markers. On clicking any one of them, more details about the marker would be shown as the name of species, where it was clicked, and by whom it was clicked along with the photo.

3. **Functionality to upload photos of butterflies if found in the wild to improve the database:**

   The application is user-centric. Hence, users should be able to add/upload new images to the database. An upload feature should exist which will allow users to do the same.

4. **Ability to classify an image as skipper and non-skipper:**

   An ML model should be integrated with the website which can classify user-uploaded images. This would help us get rid of human requirements and make the system entirely automated.

5. **Live updation of the Database according to the newly available datasets:**

   Once an image is verified to be a skipper, it should be added to the database instantly along with its metadata like location, name, species, etc.

6. **Provision to update new images of butterflies on the webpage:**

   Once a user has uploaded new images to the database and it has been verified, changes should be reflected in the map as well. The map should be dynamically updated and every new entry in the database should also be visible on the map instantaneously.

7. **Option to filter butterflies:**

   The system should have a feature that allows people to filter the list of butterflies using multiple parameters. The user can filter the list using species, place, and date.

8. **User login:**

   Each user should have the option to signup/login to their account so that they can upload images to the database and see what was uploaded by them. This helps maintain authenticity so that only the person who clicks the photo gets credits for it.

3.3 Non-functional requirements:

1. **Correctness of data:**

   The data that is being uploaded by the user has to be accurate. This is the responsibility of the ML model. It should be able to verify with reasonable accuracy that the data uploaded by the user is correct.

2. **Performance of the system:**

   The user should receive a quick response from the system. As it is a website, it should be highly responsive and users should get responses to their queries as soon as possible.

3. **Availability of the system:**

   The system should be accessible on any device of the user's choice.

4. **Reliability of the system:**

   The system should run without failure for a given period under predefined conditions.

5. **Ease of access:**

   The system should be easy to use so that even the less technologically savvy people can use it.

3.4 Hardware and software requirements:

Hardware requirements:

1. *User system*: As the end product is a web application, the requirements of the user system are quite low. The user system can be something as simple as a mobile phone. Any device connected to the internet with the ability to run a web browser can run the web application
2. *Admin system*: As an ML model needs to be trained using acquired data, the admin system needs to be powerful. It also needs to be able to host the website hence a server is required with the following specifications:
   a. 8 GB RAM and above
   b. Processor: at least 4 cores with 8 virtualization threads and a base clock of 2.4GHz
   c. Graphics card: 4 GB VRAM
   d. OS: Windows/Linux

Software requirements:

1. Google colab:

   To train models, powerful systems are required. A good alternative to this is google colab which is an online python compiler made and run by Google. Google colab provides an online platform for running python programs and training ML models. It also has all major python libraries preinstalled which are required during training so that we can reference any of them without having to download them. It runs on Google's cloud platform and uses its underlying hardware which is considerably more powerful than normal personal computers. The model training can proceed at a much faster pace.

2. Tensorflow and Keras:

   TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in ML, and developers easily build and deploy ML-powered applications. It is a library that can be imported in python. It contains a lot of predefined functions which makes creating and modifying deep learning neural network models easy. Keras is an API designed for reducing cognitive load by offering consistent & simple APIs. It helps minimize the number of actions to be performed to achieve a particularly common use case task and also provides robust debugging and error handling features. It is used by TensorFlow as the backend.

3. Scikit Learn:

   Scikit Learn is another python library used for machine learning training and testing. It provides a range of prebuilt supervised learning and unsupervised learning algorithms via a consistent interface. It is open-source and is considered one of the most robust libraries for machine learning training.

4. Flask:

   Flask is a web development framework that primarily runs on python. It is a lightweight framework that is used to create interactive websites in a short period. The main benefit of using flask is the ease of integration with machine learning models. As flask also runs on python, it can make the same calls as the machine learning model. As a result, there is no need to have an API between the website and the ML model. They can be seamlessly integrated with each other.

5. MySQL and XAMPP:

   MySQL is used as the database service to store the user-generated data as well as all the images and their metadata which is displayed on the website. It is used along with XAMPP which functions as the backend server. It is written in PHP

6. Facebook Graph API:

   This software API is required to gather data from Facebook groups and accounts. It is used in scraping Facebook pages.

7. Selenium:

   Selenium and chromium drivers are used to web scrape training images. To train the model, a vast amount of data was required and hence, selenium was used to scrape images off of Google, Flickr, Twitter, etc

3.5 Constraints:

1. The neural network is a classification model and hence, it doesn't always categorize each image perfectly. Hence, sometimes non-skipper images slip into the database.
2. There is no method to verify if the image was taken at the said place as the system trusts the user to input true values.
3. If population density in a place is too high, some of the points at a given location may overlap which might cause users to believe that there are fewer specimens.
4. The system needs to be trained again on new data to increase the accuracy.

Chapter 4: Proposed Design

4.1. System Design / Conceptual Design (Architectural) :

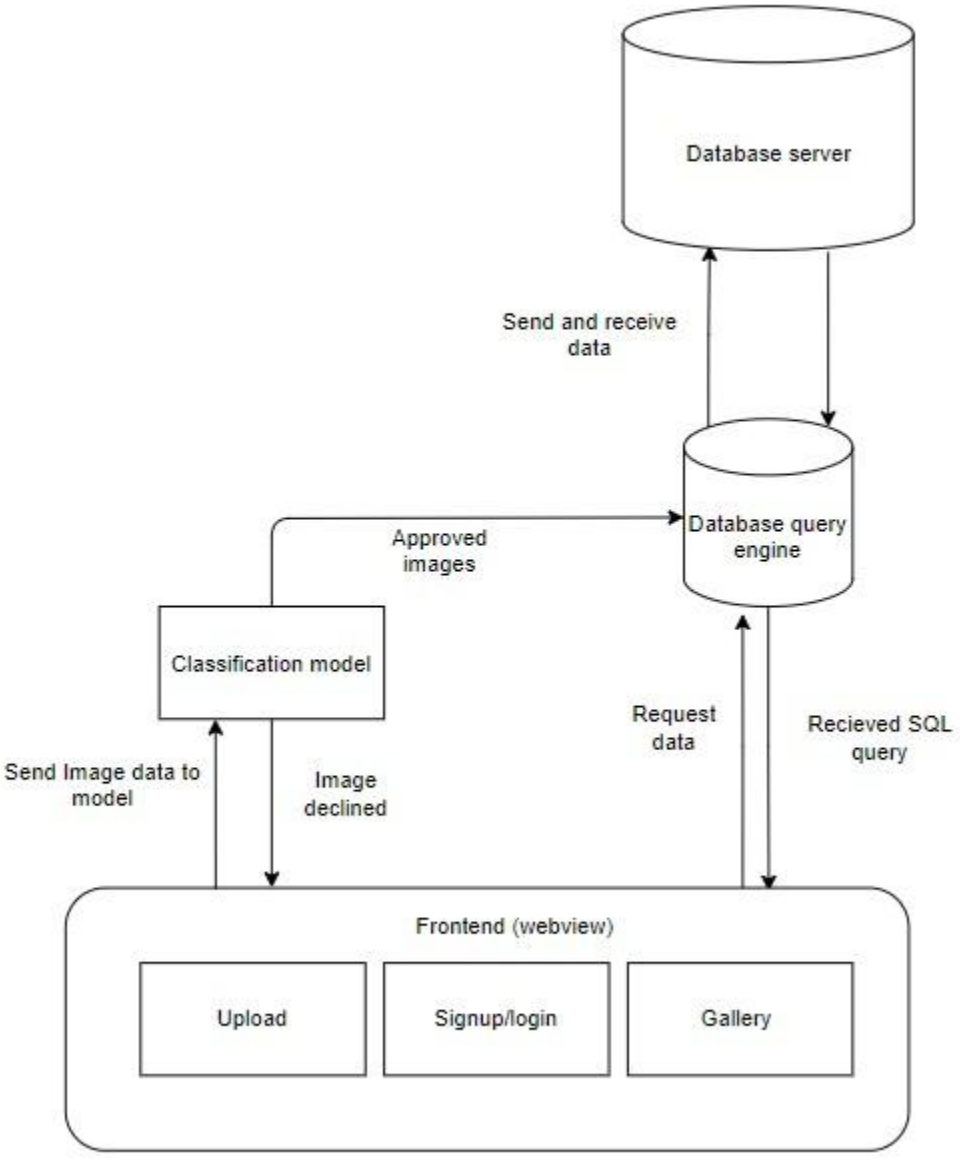


Fig 4.1. The architecture of the system

4.2. Modular diagram representation of the proposed system :

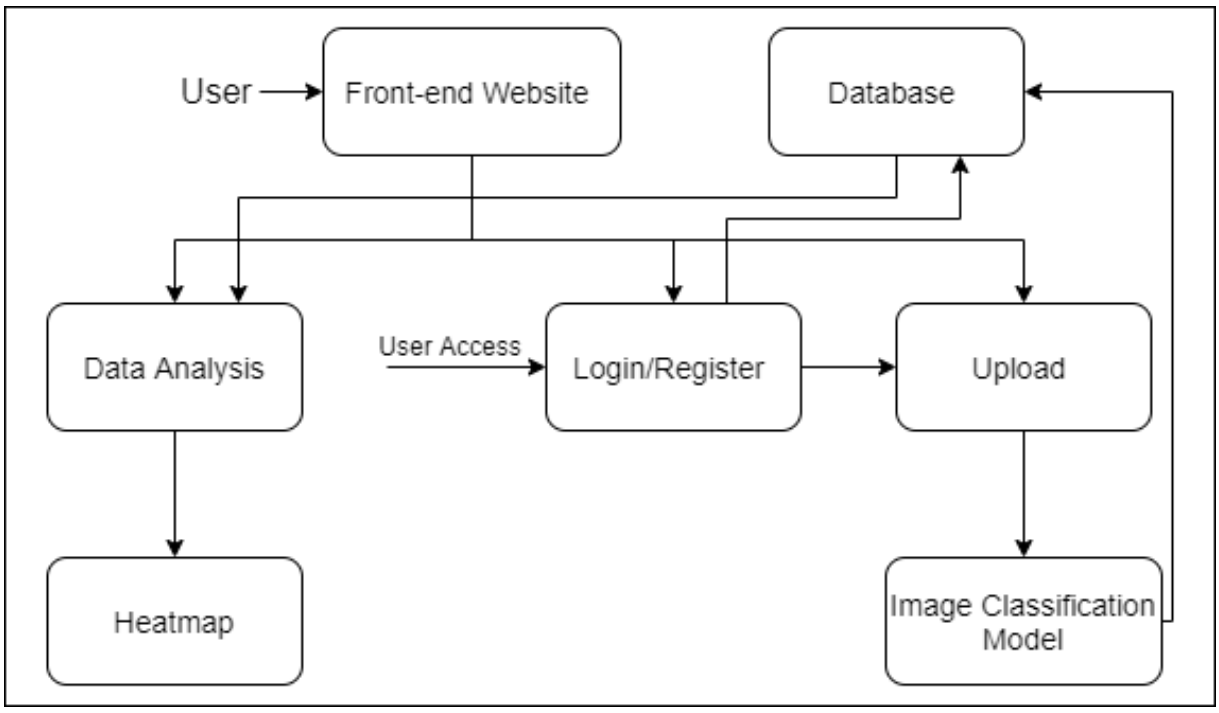


Fig 4.2. Modular diagram

The various modules of the system are as shown in figure 4.1:

1) Front-end: The front end is what users will use to interact with the system. It is made using the flask framework.
2) Database: It is used to store the butterfly data along with images and their metadata.
3) Login/register: This is the authentication method using which valid users are allowed to upload their data to our system.
4) Heatmap: It is the visual representation of the distribution of butterflies in India. It fetches data from the database and displays it as location pins on the map.
5) Upload: It is the feature using which users can upload data to the system. It consists of a form that collects all the necessary metadata regarding the image.
6) Image Classification model: This is used to classify the image that is uploaded by the user and verify if it is a skipper or not. It is passed an image from the upload feature.

4.3 Design of the proposed system with proper explanation of each :

a. Data Flow Diagram:

Level 0:-

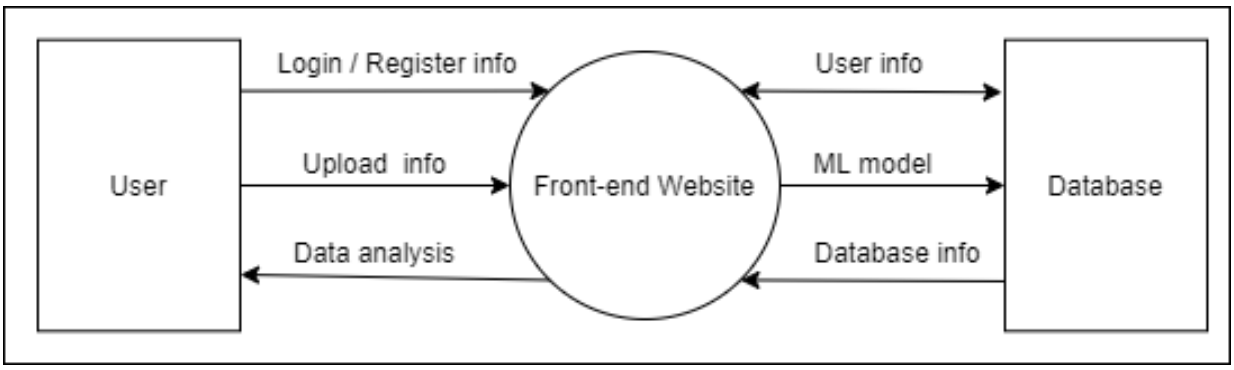


Fig 4.3. Level 0 DFD diagram

Figure 4.3 shows the basic 3 sections of our system in level 0 abstraction. The user is the interacting entity. The front end is the entity that users interact with to access the system. The database is the entity where all the data of the system is stored.

Level 1:-

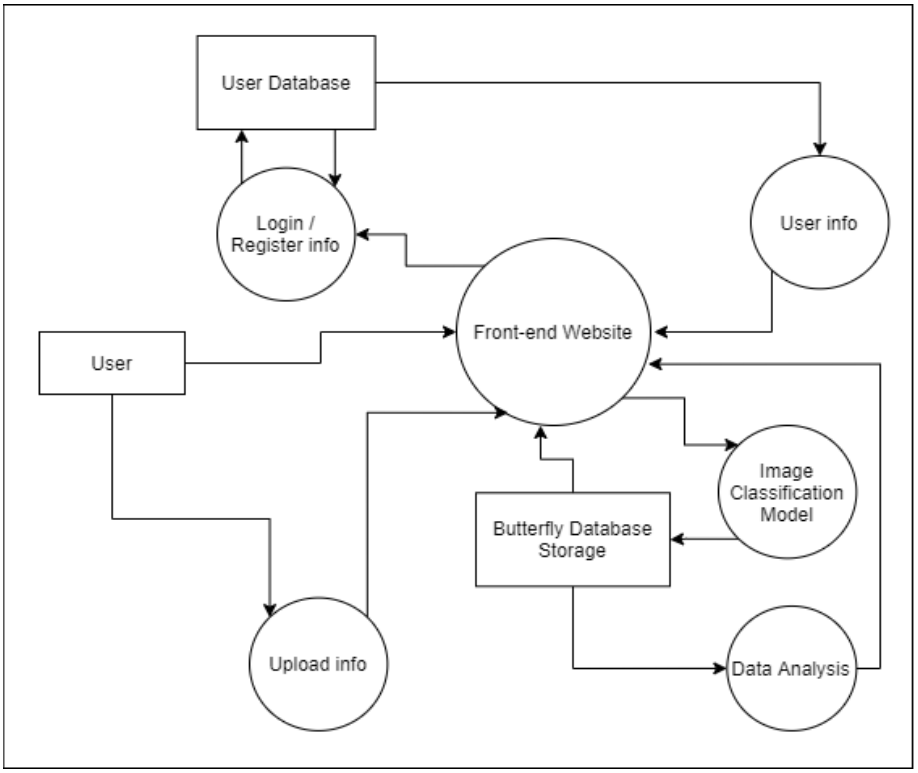


Fig 4.4. Level 1 DFD diagram

Level 1 shows the deconstruction of various modules of level 0 as shown in figure 4.4. The decomposition of the proposed model represented in level 1 DFD is the insertion and processing of data for prediction and verification of the user.

b. Flowchart for the proposed system:

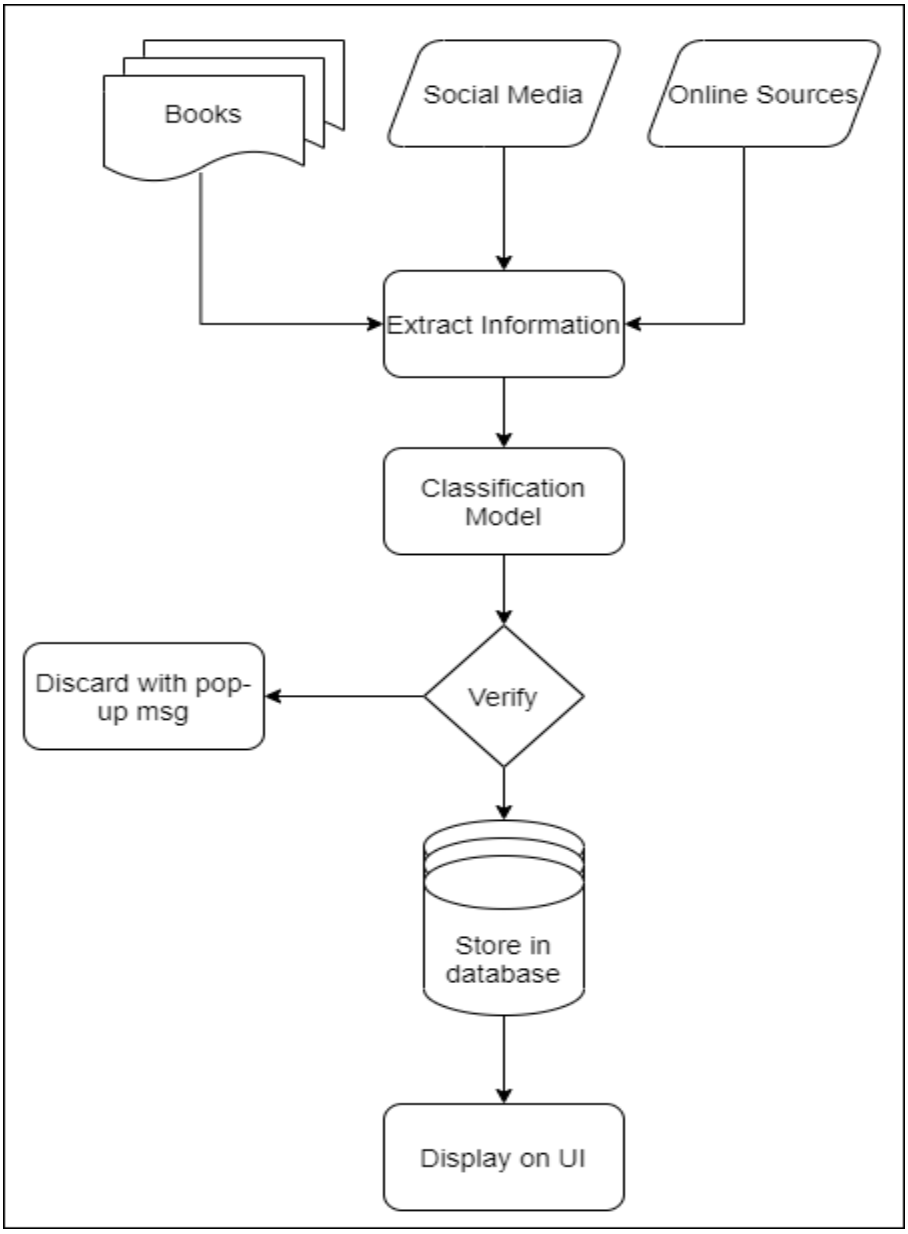


Fig 4.5. Flow diagram of the entire project

4.4. Project Scheduling & Tracking using Timeline / Gantt Chart :

Name	Begin date	End date
Selection of domain	04/06/20	08/06/20
Selection of project i...	09/06/20	11/06/20
Literature Survey	15/06/20	22/06/20
Review of existing sy...	23/06/20	30/06/20
OCR	01/07/20	15/07/20
Web Scraping	16/07/20	16/09/20
Heatmap Creation	17/09/20	28/10/20
Database Generation	29/10/20	09/12/20
Front-end Website	10/12/20	30/12/20
Front-end Website	01/02/21	26/02/21
Model Training	03/03/21	27/04/21
Integration	12/04/21	30/04/21
Testing	03/05/21	13/05/21

Fig 4.6. Timeline

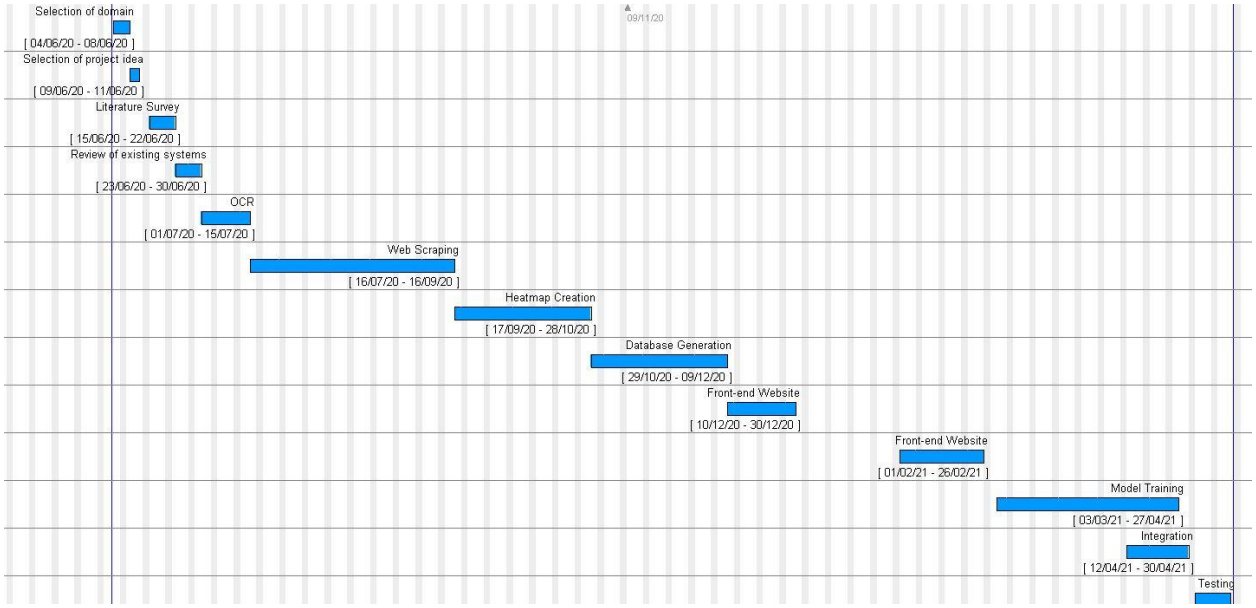


Fig 4.7. Gantt chart

Chapter 5: Implementation of the proposed system

5.1 Methodology applied:

The project was broken down into 2 sections: website and ML model. The sequential steps followed throughout the project were:

1) Data gathering:

Data was gathered from various sources using OCR, graph API and web scraping. All the methods used for web scraping are listed in section 5.3

2) Data cleaning:

The data collected from Facebook required a lot of analysis because it had certain shortcomings in terms of missing fields for example absence of species name, location, etc. and there was no way of verifying the validity of the images being uploaded. Therefore, the missing fields were populated based on the comments on those posts. For validation of images, a butterfly enthusiast was consulted before adding them to our database. The images, along with their corresponding metadata, were stored in a database to be fetched later. The coordinates of the place are needed to plot the images on a heatmap. But the data acquired from the user regarding the image is usually the name of the place where the image is clicked. Hence all locations are converted to the latitude and longitude format. This can be done manually or by using google maps API to auto-generate coordinates nearest to the place named in the Facebook post.

3) Database creation:

All this data needs to be stored in proper tables so that they can be fetched by the front end of the website and displayed. Along with this, other tables were also generated like user details, filtering details, etc.

4) Training machine learning model:

All the images found during web scraping were used in training ML models. All the images were split into a train and test set using python and then used to analyze multiple models until the proper fitting model was found.

5) Generation of heatmap:

Python libraries for plotting various maps such as GeoPandas, Folium, Matplotlib, etc. are well known. For our purposes, we have implemented the interactive library in Python called 'Folium', which then visualizes the data into a Leaflet map provided by the Leaflet.js library. The dataset is directly imported from the database into the map using the Pandas library of Python. The main motive behind choosing folium is that it supports Image, Video, GeoJSON, and TopoJSON overlays which helps in highly eye-appealing map content.

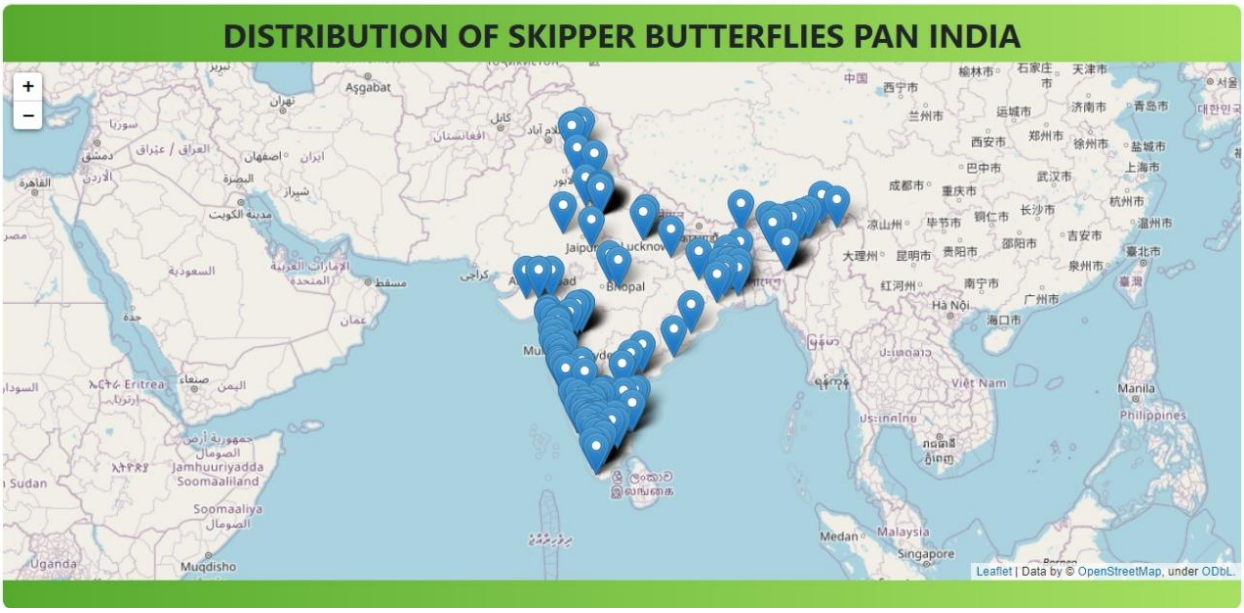


Fig 5.1. Heatmap generated

As Leaflet.js is natively supported on browsers, it is not resource-hungry. The fetched data is passed to Folium, which generates a Leaflet map and overlays all the required

markers onto the map. It then generates a .html file which is saved as an instance on the server and sent to the user when necessary. Due to Flask being a lightweight framework, the webpage is quickly rendered As the data in the map is dynamically loaded, users are given the option to filter the values based on location, species, and date. Every time a new set of filtered values is passed, the map plots the filtered ones as shown in figure 5.1.

6) Creating website:

The website is the front of the project using which users can interact with the data and add new data elements to it if they have any. This is done using the flask frontend framework. The website consists mainly of 2 sections: viewing and uploading.

Viewing: In this section, the users can view the skipper-related data that currently exists in the system. They can do so in tabular format as a gallery or as a map with location markers. The viewing section also has a filter section. This allows the user to filter the results based on the name of the species, location, or even date. All these features can be used in unison to give fine-grained filtering capabilities.

Uploading: The entire system runs on user-generated content. To increase our database and the number of points on the map, we allow users to upload images clicked by them in public. The users are allowed to upload images which are then classified by the ML model. Based on this, the image is added to the database.

7) Working of the entire website:

The flow of the website is as follows: The user can run the website on any device of their choice. On opening the website, they are greeted with a front-page listing the reasons for making the website. They can then go to the data section where they'll find all the user-generated images along with an option to filter. Here, they can also look at the map to better understand the butterfly distribution in India. If a user has any new images to upload, he/she can sign in or register for the website. Once it is done, they can access the upload feature where they can enter the relevant details listed in the form. On submitting, this image is sent to the ML model which processes it to check whether the image is of a skipper or not. If not accepted, it shows the user a popup that says that the butterfly in the image is not a skipper. If accepted, the image is added to the database and this change is

reflected instantly in the map so if a user opens the map and checks the location, they will find their post along with credits to them for uploading.

5.2 Algorithms implemented:

While training the model, multiple algorithms were tried to acquire the best accuracy. The data was split as 70% training, 30% testing. The images were loaded as 255*255*3 arrays. The algorithms implemented were:

1) Convolutional neural network with layers designed by us:

A Convolutional Neural Network, also known as CNN or ConvNet, is a type of neural network that specializes in processing image-based data as shown in figure 5.2. Convolutional neural networks have been applied to classification tasks with enormous picture datasets. A CNN can learn essential basic filters consequently and join them hierarchically to enable the depiction of latent concepts for pattern recognition. A digital image is a binary representation of visual data. An image consists of a series of pixels arranged in a grid-like fashion that contains pixel values to denote how bright and what color each pixel should be. In our project, we implemented CNNs made by us using multiple layers such as Pooling, Convolution, LeakyReLU, and softmax. The images were first passed through a reshaper function which reshaped all the images to be compatible with the input size of the neural network. Then, the image was passed through multiple layers of the network alternating between pooling, convolutions, and leakyRelU. Slowly, the dimensions of the image are reduced using pooling until it reaches the last layer, where we use the softmax function to classify the image as a Skipper or non-skipper. This model was designed by us. However, due to low counts of layers, the model wasn’t able to completely decipher all the features present in the image. Hence, it gave an accuracy of around 70%.

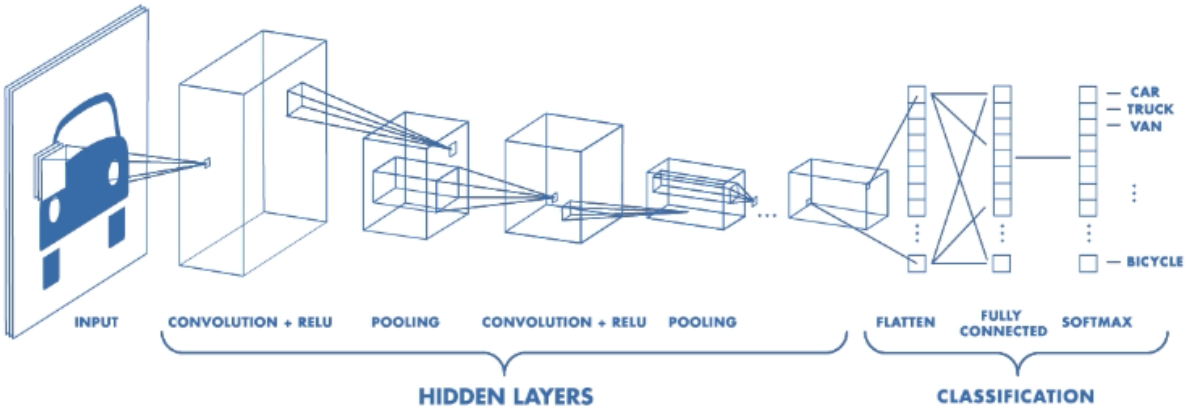


Fig 5.2. Example of CNN layers

2) Transfer learning:

From our preliminary testing, we realized that the model wasn't performing well as the network wasn't deep enough. As the network only had 5-6 layers, it wasn't able to recognize all the features present in the image. Hence, we moved on to using transfer learning with pre-trained networks like VGG-16, Resnet, etc. These provided much better results with accuracies reaching above 90%

a) Resnets:

One of the problems ResNets solve is the famous known vanishing gradient. This is because when the network is too deep, the gradients from where the loss function is calculated easily shrink to zero after several applications of the chain rule. This results in the weights never updating their values and therefore, no learning is being performed.

With ResNets, the gradients can flow directly through the skip connections backward from later layers to initial filters. The architecture is shown in figure 5.3. Resnets have already been trained on vast amounts of data and possess the ability to recognize edges and objects in images. This was useful for us as the old network designed by us wasn't able to do so. We used ResNet-34 as the base model. New layers were appended to the resnet which were trained on our data. The resnet was not included in the training process. Instead, it acted as a filter that recognized valuable information from the images

and passed this data to our layers in order to help them in training. This model was more efficient with accuracies reaching up to 80%.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Fig 5.3. ResNet architecture

b) **VGG-16:**

VGG-16 is a convolution neural net (CNN) architecture that was used to win ILSVR (Imagenet) competition in 2014. It is considered to be one of the most excellent vision model architecture to date. The most unique thing about VGG-16 is that instead of having a large number of hyper-parameters, it focuses on having convolution layers of 3x3 filter with stride 1 and always uses the same padding and max pool layer of 2x2 filter of stride 2 (Figure 5.4). It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end, it has 2 FC (fully connected layers) followed by a softmax for output. The 16 in VGG-16 refers to it having 16 layers that have weights. This network is pretty large and it has about 138 million parameters. VGG-16 is a very expansive network and takes a lot of time to learn. Due to its large number of pretrained parameters, it possesses great recognition capabilities but it takes longer and more memory to train. VGG-16 was used as the base model to which new layers made by us were added. This model had an accuracy of around 94% which was the best among all the algorithms tried.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

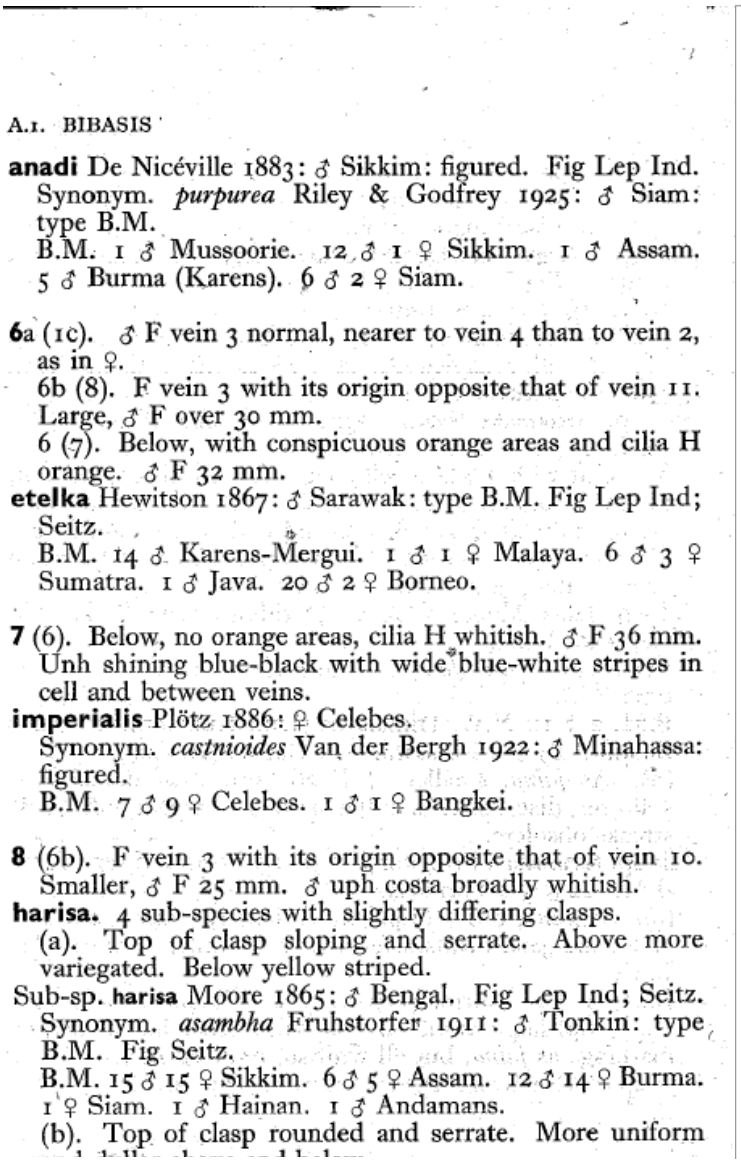
Fig 5.4. VGG-16 layers

5.3 Datasets source and utilization:

The most challenging part of the entire project was to gather data. As skipper butterflies (family Hesperiidae) aren't well documented, gathering data regarding them proved to be difficult. Multiple methods were used to extract relevant information.

1) OCR:

Initially, books were referred to extract the physical features of the butterflies. For this purpose ___ book was initially referred. It is a very old book from the colonial era where ecologists from London had toured India to gather data on various species of butterflies in India. During this, they had caught, studied, and documented various aspects of all the butterflies that they came across. Features such as length, the size of wings, proboscis, etc. were documented along with gender. This data could be useful as using it, we could extract relevant data from butterfly images and then cross-reference them with the data from the book to gain insight on the species. However, as the books were very old, they weren't available in digital format. So to extract data from these books, OCR was applied (optical character recognition). PyTesseract library was used for this purpose. It recognizes and "reads" the text embedded in images. It is a wrapper for Google's Tesseract-OCR Engine. It can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, BMP, tiff, and others. Additionally, if used as a script, Python-tesseract prints the recognized text instead of writing it to a file. However, the images of the books weren't good enough. Even after applying preprocessing to the images, the results obtained weren't legible or useful in any way as seen in figure 5.5 . Then we attempted to physically extract data from the book by reading it and documenting relevant data. However, the book contained data on all species of butterflies and did not document all the butterflies in India. As a result, using this data would have skewed the results, and butterflies not listed in the books would have gone unnoticed. Hence, this data was deemed unusable and dropped.



A.1. BIBASIS

anadi De Nicéville 1883: ♂ Sikkim: figured. Fig Lep Ind. Synonym. *purpurea* Riley & Godfrey 1925: ♂ Siam: type B.M.
B.M. 1 ♂ Mussoorie. 12 ♂ 1 ♀ Sikkim. 1 ♂ Assam. 5 ♂ Burma (Karens). 6 ♂ 2 ♀ Siam.

6a (1c). ♂ F vein 3 normal, nearer to vein 4 than to vein 2, as in ♀.
6b (8). F vein 3 with its origin opposite that of vein 11. Large, ♂ F over 30 mm.
6 (7). Below, with conspicuous orange areas and cilia H orange. ♂ F 32 mm.
etelka Hewitson 1867: ♂ Sarawak: type B.M. Fig Lep Ind; Seitz.
B.M. 14 ♂ Karens-Mergui. 1 ♂ 1 ♀ Malaya. 6 ♂ 3 ♀ Sumatra. 1 ♂ Java. 20 ♂ 2 ♀ Borneo.

7 (6). Below, no orange areas, cilia H whitish. ♂ F 36 mm. Unh shining blue-black with wide blue-white stripes in cell and between veins.
imperialis Plötz 1886: ♀ Celebes.
Synonym. *castnioides* Van der Bergh 1922: ♂ Minahassa: figured.
B.M. 7 ♂ 9 ♀ Celebes. 1 ♂ 1 ♀ Bangkei.

8 (6b). F vein 3 with its origin opposite that of vein 10. Smaller, ♂ F 25 mm. ♂ uph costa broadly whitish.
harisa. 4 sub-species with slightly differing clasps.
(a). Top of clasp sloping and serrate. Above more variegated. Below yellow striped.
Sub-sp. **harisa** Moore 1865: ♂ Bengal. Fig Lep Ind; Seitz.
Synonym. *asambha* Fruhstorfer 1911: ♂ Tonkin: type B.M. Fig Seitz.
B.M. 15 ♂ 15 ♀ Sikkim. 6 ♂ 5 ♀ Assam. 12 ♂ 14 ♀ Burma. 1 ♀ Siam. 1 ♂ Hainan. 1 ♂ Andamans.
(b). Top of clasp rounded and serrate. More uniform

OUTPUT:

Aa. BIBASIS

ava

de Nicéville i883: ¢ Sikkim: ligured, Wig Lep Ind.
Synonym. puspurce Kiley & Godlrey igag: ¢ Siam
type 1M.

BM. 1 Mussoorie. 12 ¢ a 4 Sikkim. 1 gf Assim,
5 4 Burma (Ravens). 4 2 Siam,

61 (1c), vein 3 normal, nearer to vein g than to vein 2,
as in.
Gh (8). F vein 3 with its origin opposite
Large, 1 over 30 mm.
649). Helow, wiih co
ouiipe, #1 32 ant
a Mewitsen 18677 Su

hat of vein 11,

ous orange areas and cilia U

wak: type BLM. Fig Lep Lads

BML 14 of Ravens Mergui,

Fig 5.5. Input page for OCR and results obtained

2) Web scraping:

The next source of information was the internet where we could find images to train our model with. Hence, an attempt was made to scrape data from different websites such as Flickr, Google, found butterflies, Facebook, etc. However, as skippers aren't well-explored, the number of images was limited. Many people click photos of butterflies and upload them on various internet sites from which we could extract them. Images for training purposes were obtained from these websites. However, the legitimacy of these images is brought into question. So we have to rely on only trustworthy sources like other butterfly documentation websites and groups made by naturalists, enthusiasts, or lepidopterists. Such groups were used to extract image data (websites like ifoundbutterflies.org, inaturalist.org, etc. Once proper websites and pages were selected, a web scraping code was written using Selenium web driver. Selenium Webdriver is an open-source collection of APIs which is used for testing web applications. The Selenium Webdriver tool is used for automating web application testing to verify that it works as expected or not. It mainly supports browsers like Firefox, Chrome, Safari, and Internet Explorer. It is used along with chromium driver to scrape data using google chrome browser. Using this, sufficient training data was acquired with both positive and negative sets.

3) Facebook:

Another great source of information is Facebook groups, where naturalists and photographers/enthusiasts upload pictures of butterflies, and people verified those images in the comment section. This data proves useful for Heatmap generation as it not only provided images but also provided metadata to be used along with it which can be shown on the heatmap. For scraping this data, Facebook graph API is used along with manual scraping to verify that the contents were accurate. This data, being undocumented, is quite useful as it meets the requirements of our application and is used as the base for heatmap as well as for training purposes. The data acquired were images, species and subspecies name, name of the photographer, and the date on which the photo was taken.

4) **Field visits:**

In order to gather data about local butterfly species and to better understand their patterns, we, along with our mentor Dr. Sharmila Sengupta visited a butterfly garden situated in Airoli. This visit provided useful insight on butterflies and the caretakers also gave us more data regarding butterflies in the form of images.

Data utilization:

The data was used in 2 ways: for training models and for displaying content on the heatmap.

1) Training models:

The data acquired using scraping was visual. Hence, it was used to train deep learning models like convolutional neural networks as stated above.

2) Website and heatmap:

To show skipper butterflies to users, data from Facebook was used in heatmap generation and website pages. This data contained the name, species, etc

5.4 Screenshots (GUI) of the project:

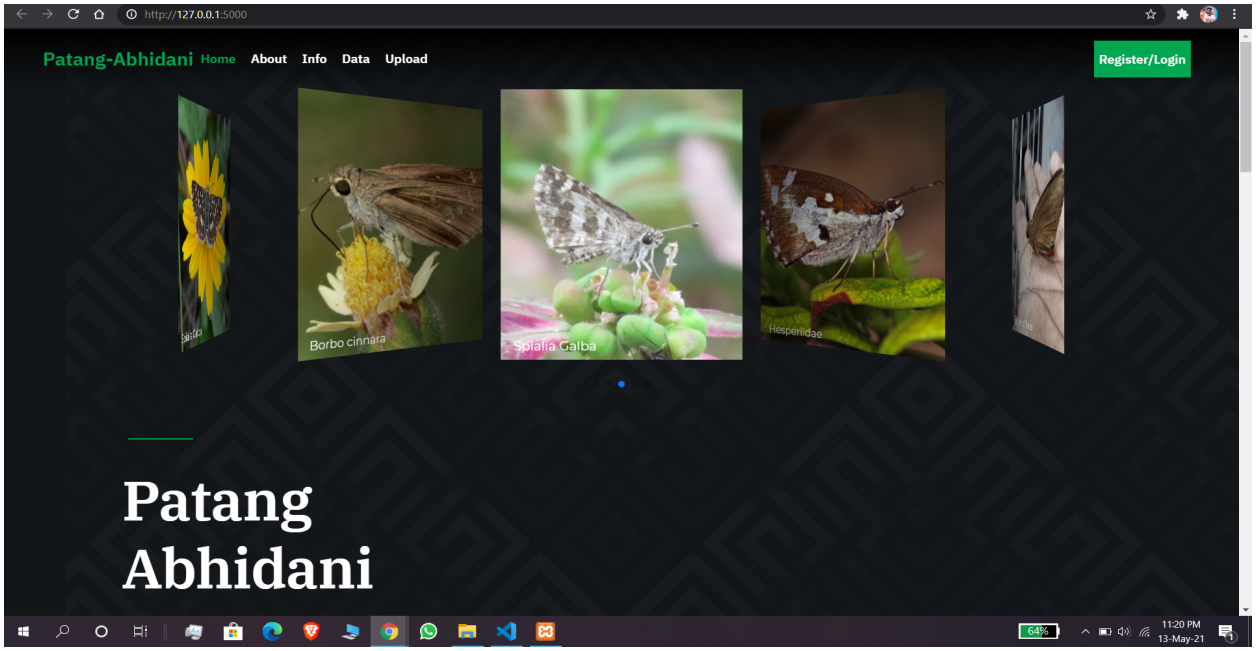


Fig 5.6. Website home page

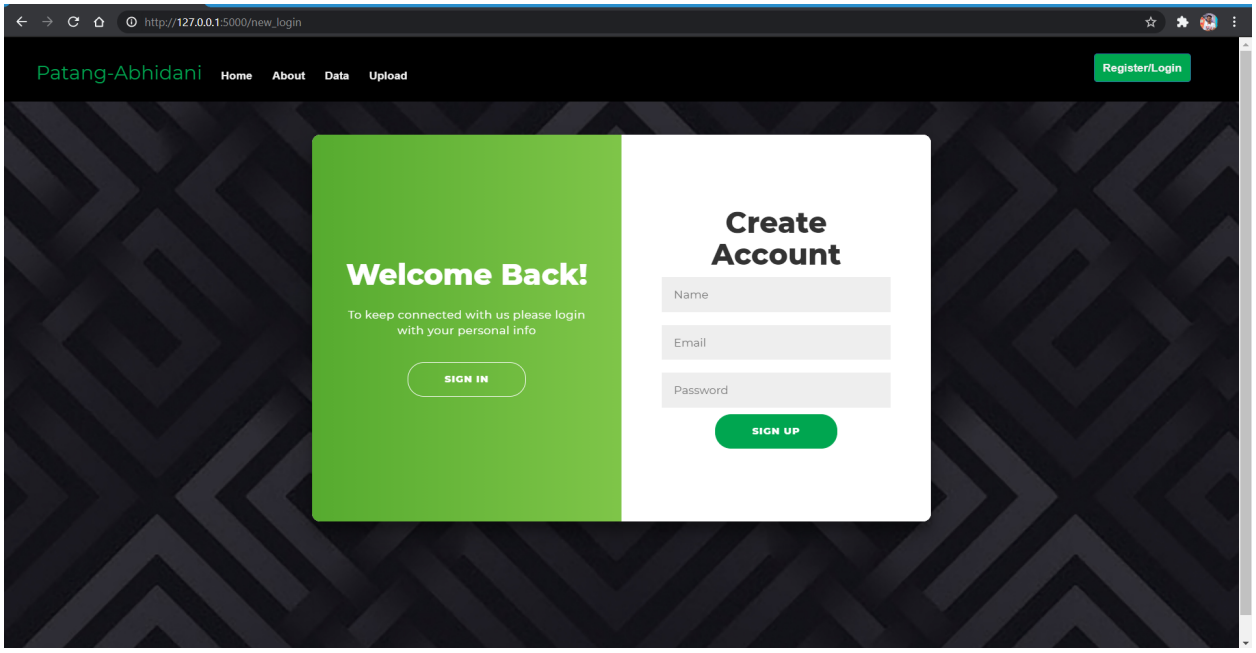


Fig 5.7. Signup page

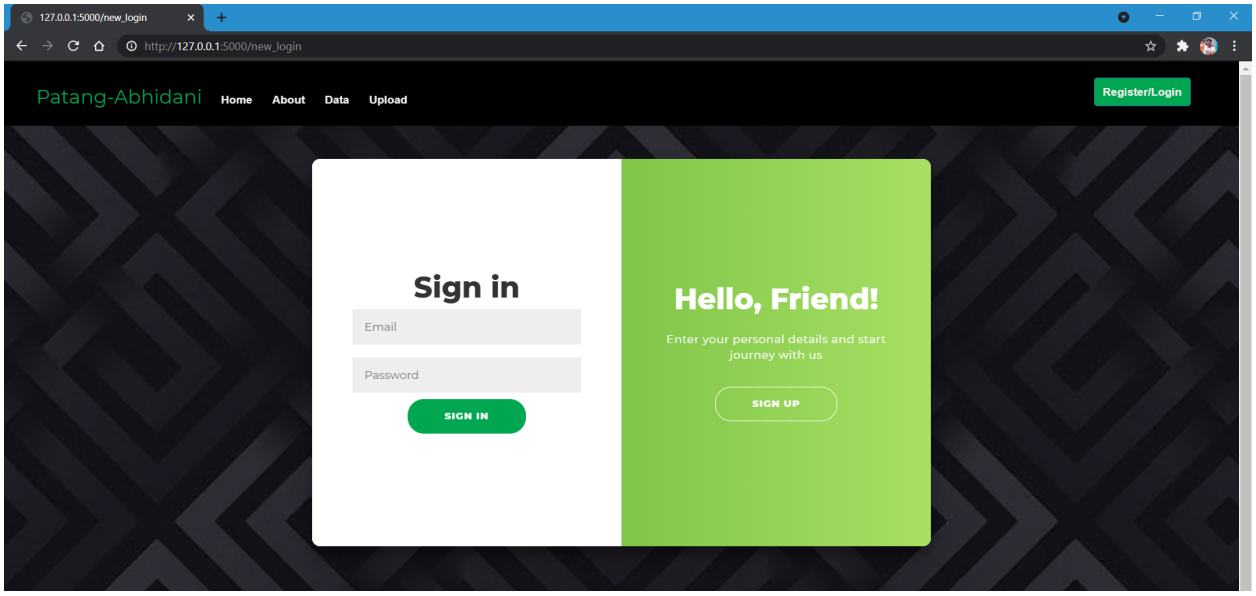


Fig 5.8. Login page

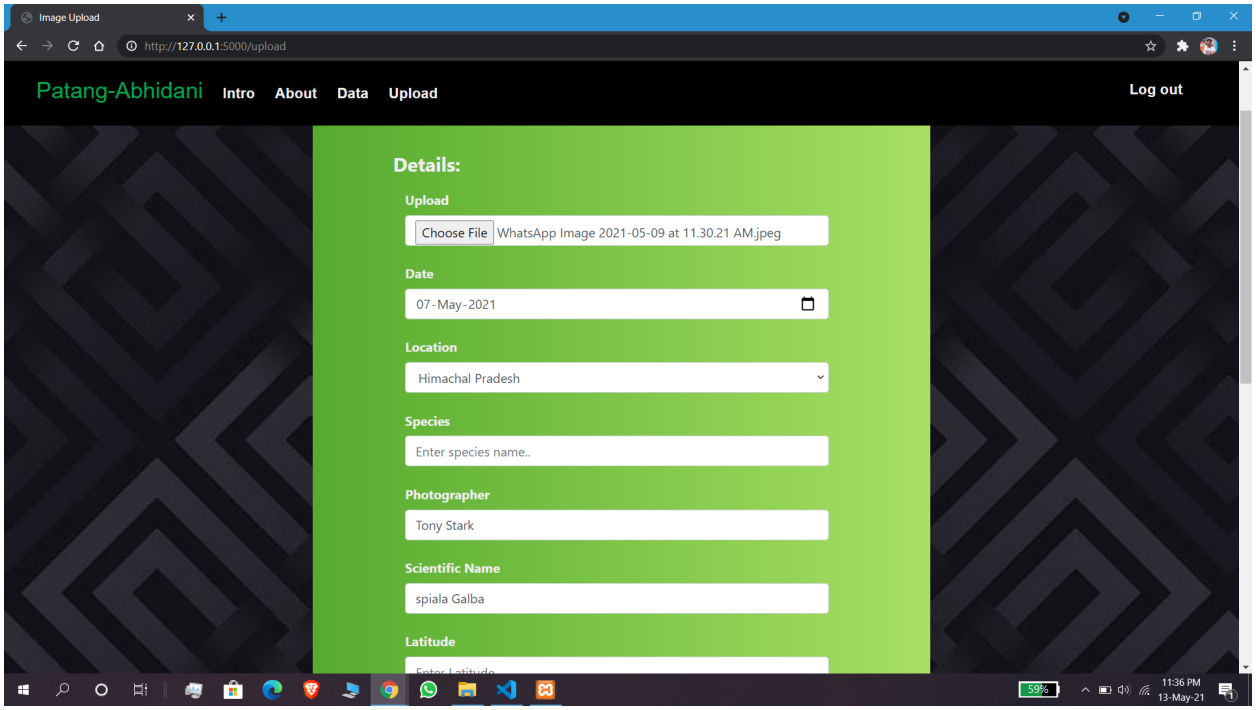


Fig 5.9. Upload Page

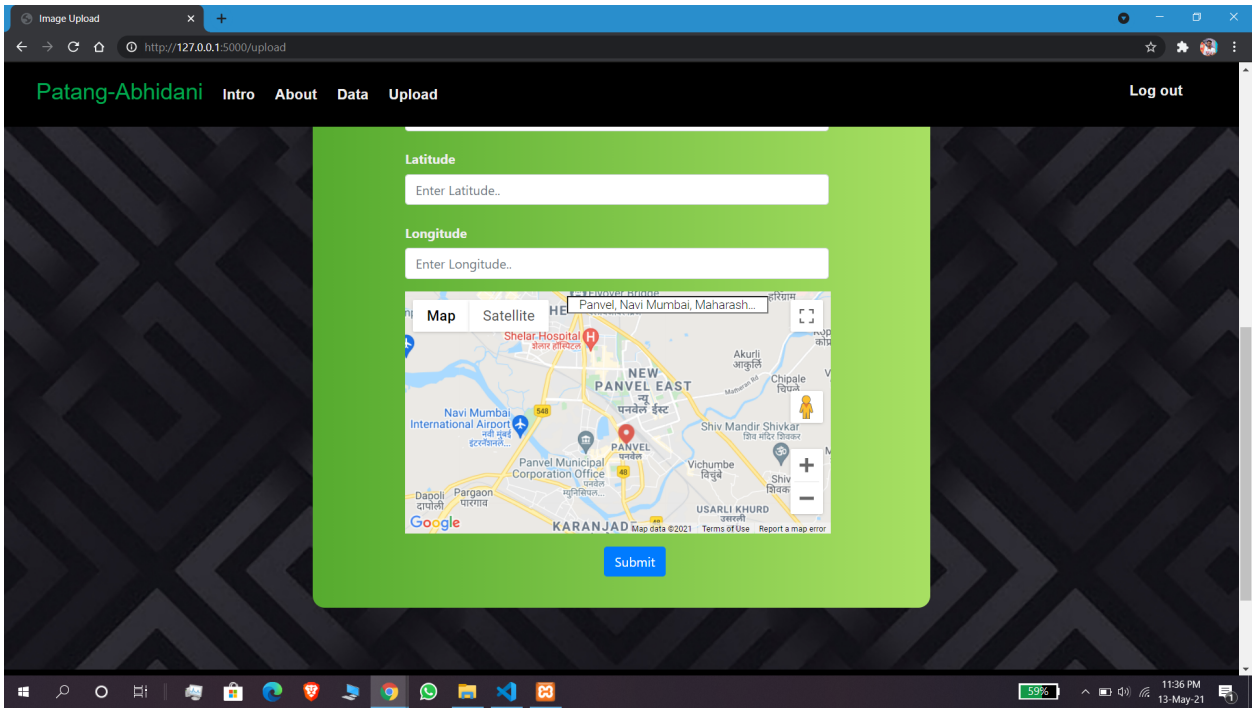


Fig 5.10. Google maps integration for location search

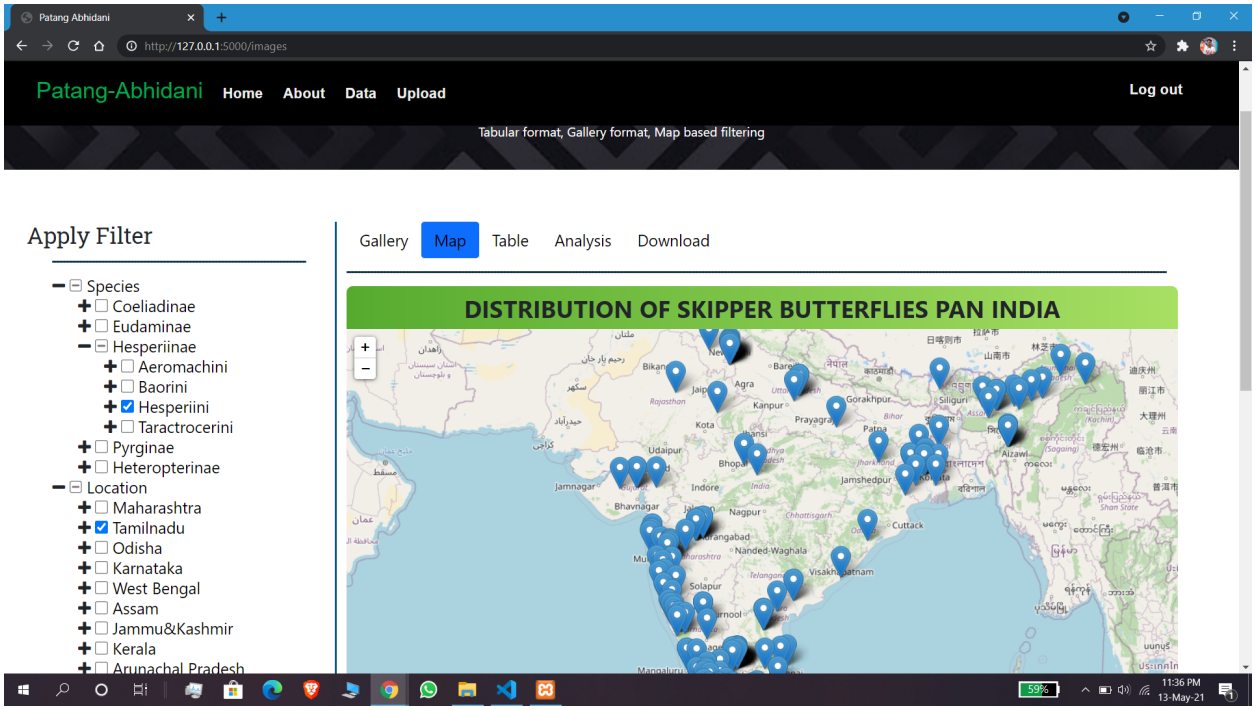


Fig 5.11. Folium map displaying the distribution

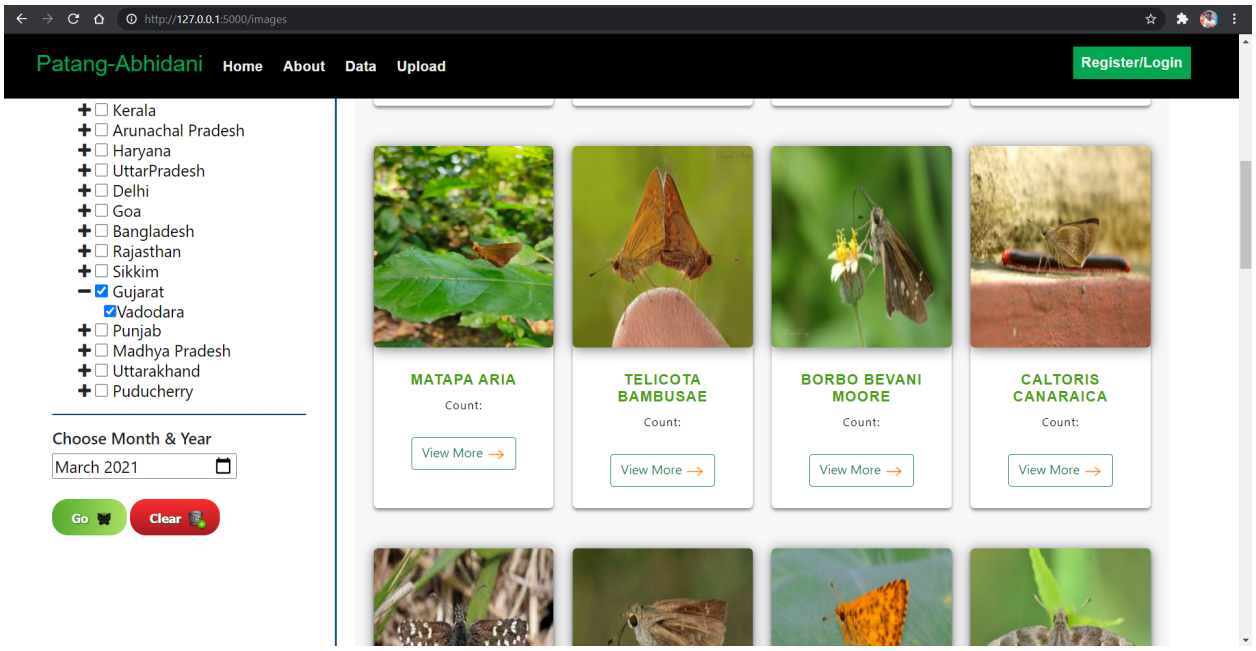


Fig 5.12. Gallery displaying the data

Chapter 6: Testing

6.1. Definition of testing:

In general, testing is finding out how well something works. In terms of human beings, testing tells what level of knowledge or skill has been acquired. In computer hardware and software development, testing is used at key checkpoints in the overall process to determine whether objectives are being met.

6.2. Types of tests:

1. Unit testing:

Exercises specific paths in a component's control structure to ensure complete coverage and maximum error detection. Components are then assembled and integrated. Focuses testing on the function or software module. Concentrates on the internal processing logic and data structures. Is simplified when a module is designed with high cohesion. Reduces the number of test cases.
Allows errors to be more easily predicted and uncovered. Concentrates on critical modules and those with high cyclomatic complexity when testing resources are limited.

2. Integration testing:

Integration testing is a systematic technique for constructing the software architecture while at the same time conducting tests to uncover errors associated with interfacing. The objective is to take unit-tested components and build a program structure that has been dictated by design.

3. Validation Testing:

Validation testing follows integration testing. Focuses on user-visible actions and user-recognizable output from the system. Demonstrates conformity with requirements.
Designed to ensure that all functional requirements are satisfied, all behavioral characteristics are achieved, all performance requirements are attained, documentation is correct, usability and other requirements are met.

4. System testing:

Verifies that all system elements (software, hardware, people, databases) mesh properly and that overall system function and performance are achieved. Software is only one element of a larger computer-based system. Ultimately, the software is incorporated with other system elements (e.g., hardware, people, information), and a series of system integration and validation tests are conducted. These tests fall outside the scope of the software process and are not conducted solely by software engineers. However, steps taken during software design and testing can greatly improve the probability of successful software integration in the larger system.

6.3. Type of Testing considered with justification:

● Statement coverage: In this technique, the aim is to traverse all statements at least once. Hence, each line of code is tested. In case of a flowchart, every node must be traversed at least once. Since all lines of code are covered, helps in pointing out faulty code.

● Branch Coverage: In this technique, test cases are designed so that each branch from all decision points are traversed at least once. In a flowchart, all edges must be traversed at least once.

6.4 Various test case scenarios considered:

Test case 1 - Login form

Test Scenario a - Check the working of Login with valid credentials.

Test steps -

1. Enter valid username
2. Enter correct password
3. Click on the login button

Test Scenario b - Check the working of Login with invalid credentials.

1. Enter invalid username
2. Enter incorrect password
3. Click on the Login button

Test case 2 - Upload

Test Scenario a - Check the working of upload form with Skipper image as input.

Test steps -

1. Upload a skipper image
2. Enter other relevant information
3. Click on the upload button

Test Scenario b - Check the working of upload form with Non-Skipper image as input.

Test steps -

1. Upload a non-skipper image
2. Enter other relevant information
3. Click on the upload button

Test case 3 - Filter Species, location or date

Test Scenario a - Filter Data according to particular species or a set of species.

Test steps -

1. Select a species or a set of species from the dropdown.
2. Click on the filter button

Test Scenario b - Filter Data according to particular location or a set of locations.

Test steps -

1. Select a location or a set of locations from the dropdown.
2. Click on the filter button

Test Scenario c - Filter Data according to particular date.

Test steps -

1. Select a date.
2. Click on the filter button

Test Scenario d - Filter Data according to species, location and date, all together.

Test steps -

3. Select species, location and date, all together from the dropdown.
4. Click on the filter button

6.5. Inference drawn from the test

Test case 1 (Scenario a):

Expected Result - The user should be granted login permission.

Pass/Fail - Pass

Test case 1 (Scenario b):

Expected Result - The user should not be granted login permission.

Pass/Fail - Pass

Test case 2 (Scenario a):

Expected Result - Image is valid and should be uploaded.

Pass/Fail - Pass

Test case 2 (Scenario b):

Expected Result - Image is invalid and hence shouldn't be uploaded.

Pass/Fail - Pass

Test case 3 (Scenario a):

Expected Result - All the data should be filtered out according to selected species.

Pass/Fail - Pass

Test case 3 (Scenario b):

Expected Result - All the data should be filtered out according to the selected location.

Pass/Fail - Pass

Test case 3 (Scenario c):

Expected Result - All the data should be filtered out according to the selected date.

Pass/Fail - Pass

Test case 3 (Scenario d):

Expected Result - All the data should be filtered out according to the selected location,species and date.

Pass/Fail - Pass

Chapter 7: Result Analysis

7.1. Module(s) under consideration:

The major part of the project to be analyzed is the image classification system. We tried implementing various CNN models with and without transfer learning with varying degrees of success. In the end, we discovered that Deep neural networks with VGG-16 as the base along with our layers performed the best, giving an accuracy of around 90%.

7.2. Parameters considered:

For our project, the metrics considered by us are:

1. Classification Accuracy.
2. Log Loss.
3. Area Under ROC Curve.
4. Confusion Matrix.
5. Classification Report.

1. Classification Accuracy:

It is the ratio of the number of correct predictions to the total number of predictions made by the model. It is the most commonly used metric as a simple ratio is used. However, it isn't the most suitable metric for testing all models. It is suitable only when there are equal numbers of observations in each class. It also considers that all predictions and prediction errors are equally important.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig 7.1. Classification accuracy formula

2. Log Loss:

It is a performance metric that is used to calculate the probability of membership to a given class. It lies between 0 and 1 and can be seen as a measure of confidence for a prediction by a model. Predictions that are correct or incorrect are rewarded or punished proportionally to the confidence of the prediction. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log-loss value means better predictions. Log Loss is a slight twist on something called the Likelihood Function. Log Loss is -1 * the log of the likelihood function. So, we will start by understanding the likelihood function. The likelihood function answers the question "How likely did the model think the observed set of outcomes was".

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Fig 7.2. Log loss formula

3. Area Under ROC Curve:

Area Under ROC Curve (or ROC AUC for short) is a performance metric for binary classification problems. The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random. A ROC curve is a plot of the true positive rate and the false-positive rate for a given set of probability predictions at different thresholds used to map the probabilities to class labels. The area under the curve is then the approximate integral under the ROC Curve.

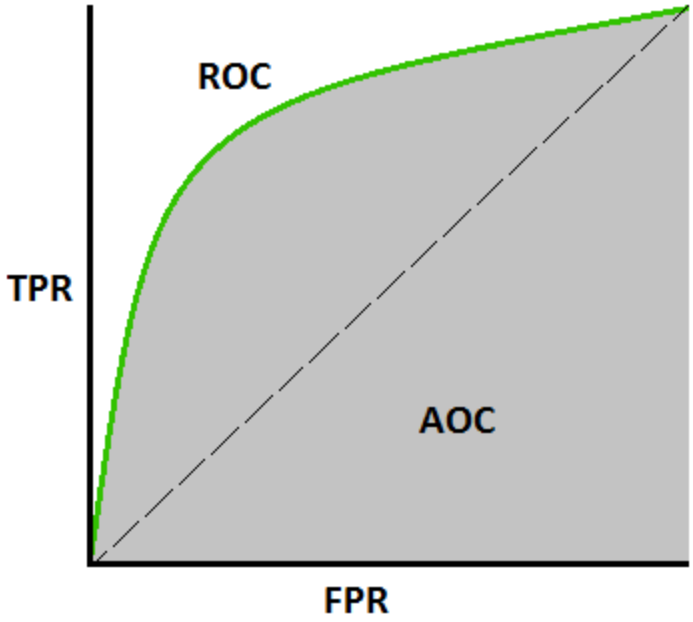


Fig 7.3. AUC-ROC graph

4. Confusion Matrix:

The confusion matrix is useful for representing the accuracy of a model with 2 or more classes. It generates a table that shows predictions on the x-axis and accuracy outcomes on the y-axis. For example, a machine learning algorithm can predict 0 or 1 and each prediction may have been a 0 or 1. Predictions for 0 that were 0 appear in the cell for prediction=0 and actual=0, whereas predictions for 0 that were 1 appear in the cell for prediction = 0 and actual=1. And so on.

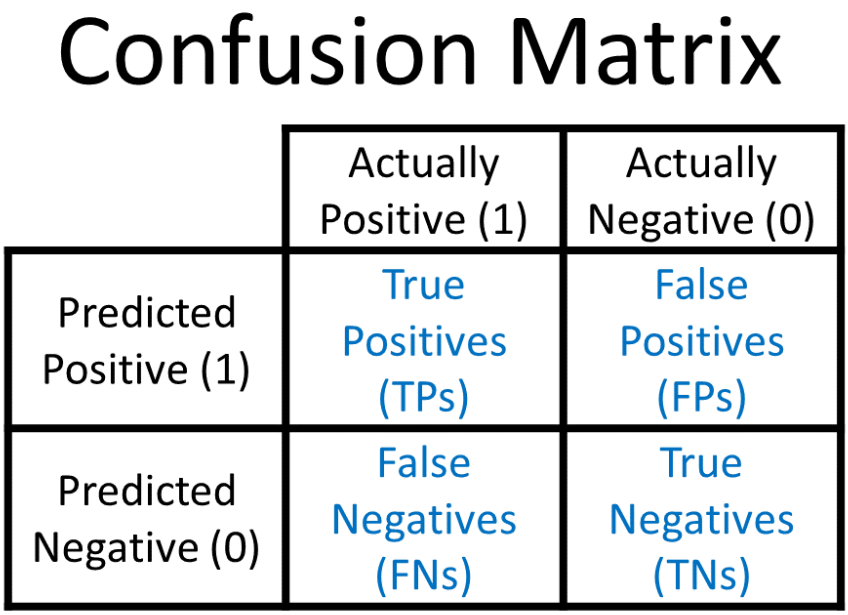


Fig 7.4. Confusion matrix fields

5. Classification Report:

A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives, and False Negatives are used to predict the metrics of a classification report.

TN / True Negative: when a case was negative and predicted negative

TP / True Positive: when a case was positive and predicted positive

FN / False Negative: when a case was positive but predicted negative

FP / False Positive: when a case was negative but predicted positive

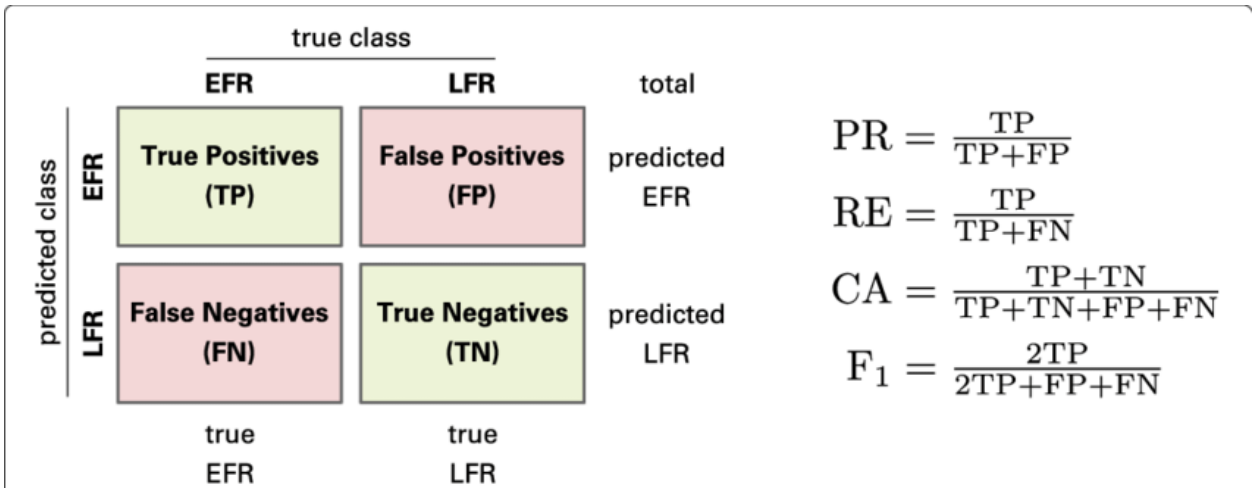


Fig 7.5. Formulae for all the analysis parameters used in the classification report

7.3. Screenshots of User Interface (UI) for the respective module:

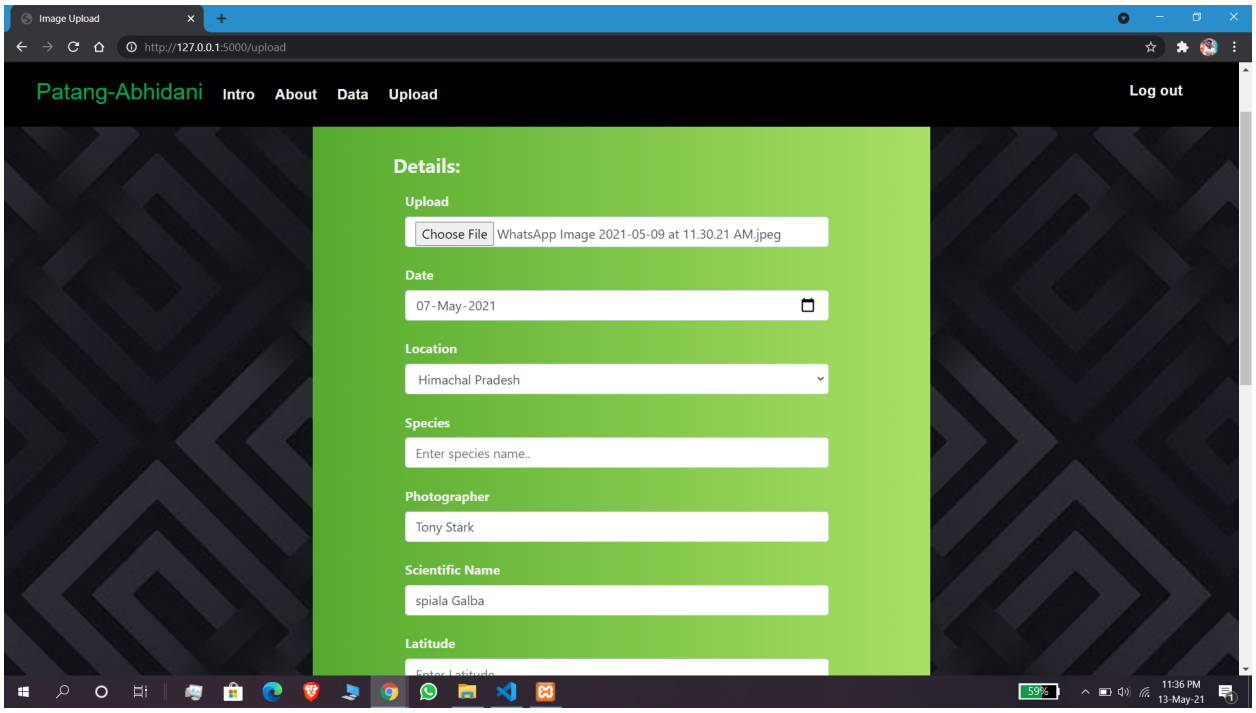


Fig 7.6. Front end for image recognition module

7.4. Evaluation of the developed system (Accuracy, Effectiveness, Efficiency):

Using the classification accuracy formula, we achieved an accuracy of around 90%. However, this measure is not enough to properly assess the system as it considers all errors to be equally important

```
from sklearn.metrics import accuracy_score
print("Accuracy: ",accuracy_score(y_val.argmax(axis=1), predIdxs)*100,"%")
```

```
Accuracy:  89.8 %
```

Fig 7.7. Classification accuracy

The log loss of the model was 3.523. It is the average when it comes to classification. The accuracy can be considered as the distance from 0. The closer the value is to 0, the better is the accuracy.

```
from sklearn.metrics import log_loss
print("Log loss: ",log_loss(y_val.argmax(axis=1), predIdxs))
```

```
Log loss:  3.5229935729575397
```

Fig 7.8. Log loss

Score: ROC AUC=0.897

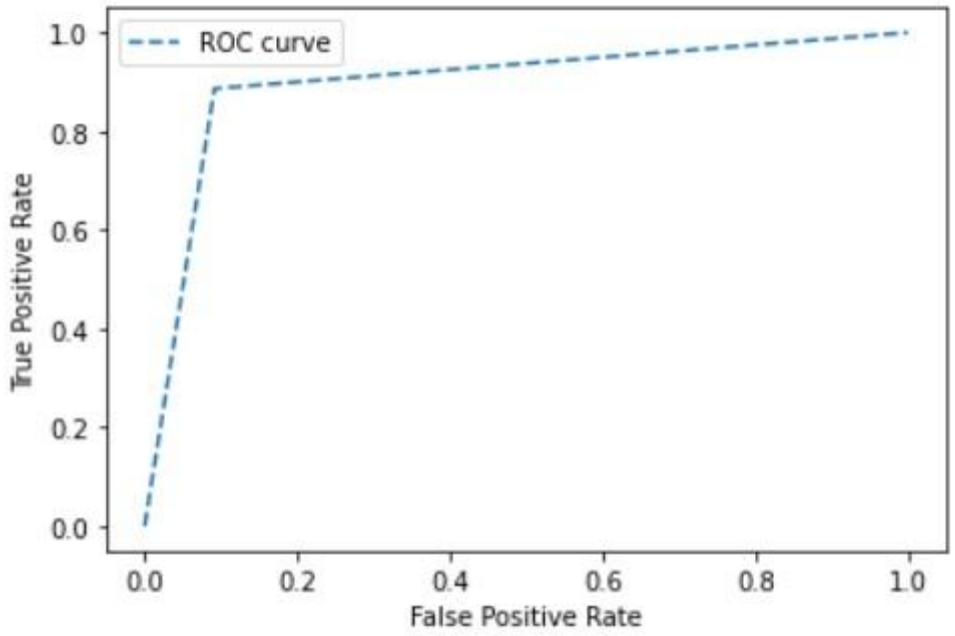


Fig 7.9. ROC curve

```
from sklearn.metrics import confusion_matrix
rounded_labels= y_val.argmax(axis=1)
print("Confusion matrix: ")
cmtx = pd.DataFrame(
    confusion_matrix(rounded_labels, predIdxs),
    index=['true:yes', 'true:no'],
    columns=['pred:yes', 'pred:no']
)
print(cmtx)
```

```
Confusion matrix:
          pred:yes  pred:no
true:yes       477       48
true:no         54      421
```

Fig 7.10. Confusion matrix

7.4. Graphical outputs of the various scenarios considered:

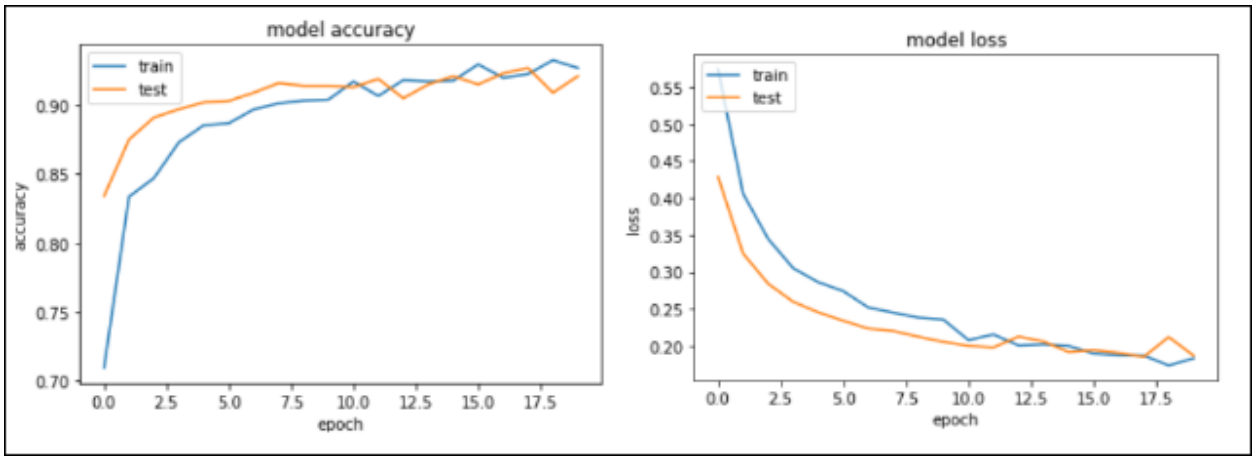


Fig 7.11. Graphs depicting model accuracy and loss of train and test datasets

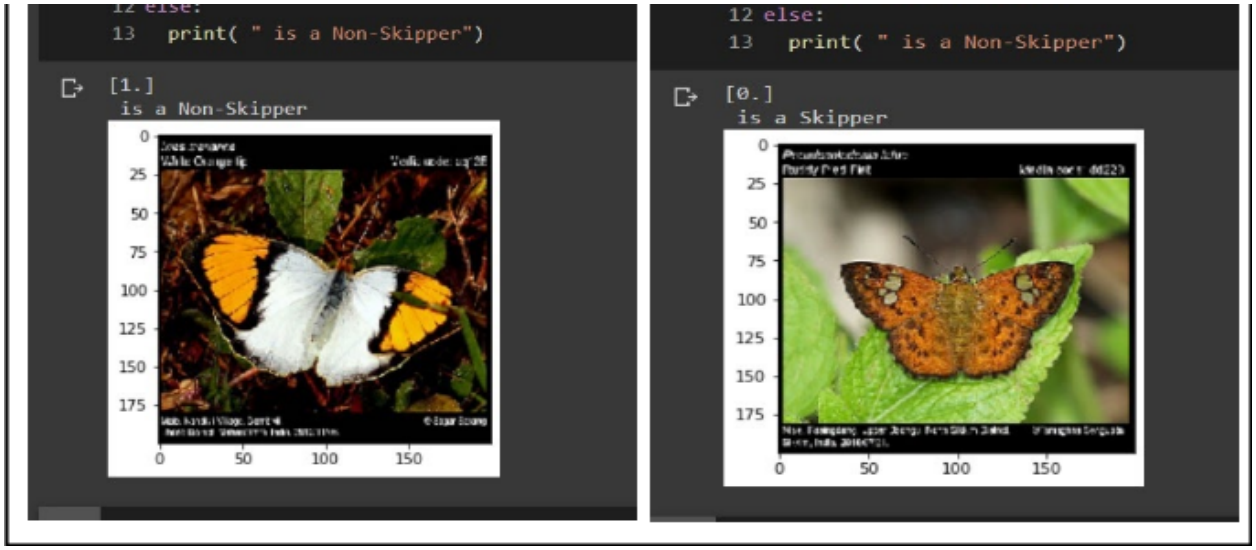


Fig 7.12. Model classifying the image inputted as skipper or not.

7.5. Reports generated / Tables obtained:

```
print(classification_report(y_val.argmax(axis=1), predIdxs))
```

	precision	recall	f1-score	support
0	0.90	0.91	0.90	525
1	0.90	0.89	0.89	475
accuracy			0.90	1000
macro avg	0.90	0.90	0.90	1000
weighted avg	0.90	0.90	0.90	1000

Fig 7.13. Classification report

7.6 Comparison with the existing systems (wrt results):

As compared to other systems, ours gives respectable accuracy with F1-score of 90%. The system was limited by training data and storage space and hence, it can be improved upon. Other existing systems exhibit similar accuracy. Some specialized models trained on only selective butterfly species exhibit higher accuracy of about 94% but any system that tries to classify a large number of butterflies has its accuracy drop down to similar ranges. The reason behind this is that most systems make use of transfer learning techniques which yield similar results with various species of butterflies.

Chapter 8: Conclusion

8.1 Limitations:

Although the system works as an all-encompassing guide for skipper butterflies, it has a few fallacies like:

1) The current system does not have 100% accuracy and hence it can sometimes recognize the species incorrectly.
2) The model is capable of recognizing only skipper butterflies and hence it becomes very limited.
3) There is no way to verify if the user uploading the image is the actual person who clicked the image.

8.2 Conclusion:

The diversity of flora and fauna is essential for the well-being of our planet. Many plants and animals rely on them for their sustenance and reproduction. Skipper butterflies are important pollinators, and hence, it is necessary to make sure that their population counts are maintained to save them from extinction. This can be done by tracking their movements around the nation and documenting them to establish the sustainability of the selected species of butterflies. The proposed system enables users to do so by providing them with ample information about the species along with the ability to study their migration patterns. Such a platform thrives off of user-generated content to increase the pool of information which is made available to everyone.

8.3 Future Scope:

1. The project can be further improved by using newly acquired data and training the model based on that data to increase the accuracy of it.
2. We can branch out and add new species to the model so that it can start considering more species while classifying images.
3. We can also add the intraspecies classification to figure out the subspecies of butterflies.
4. The entire training process can be automated.

References

[1] Hari Theivaprakasham, 2021. Identification of Indian butterflies using Deep Convolutional Neural Network, Journal of Asia-Pacific Entomology, Volume 24, Issue 1, ISSN 1226-8615,https://doi.org/10.1016/j.aspen.2020.11.015.

[2] Xie, J., Lu, Y., Wu, Z. et al. Investigations of butterfly species identification from images in natural environments. Int. J. Mach. Learn. & Cyber. (2021). https://doi.org/10.1007/s13042-021-01322-8

[3] Dey, Pritam & Payra, Arajush & Mondal, Krishnendu. (2017). A study on butterfly diversity in Singur, West Bengal, India. e-planet. 15. 73-77

[4] Sahoo RK, Warren AD, Wahlberg N, Brower AVZ, Lukhtanov VA, Kodandaramaiah U. 2016. Ten genes and two topologies: an exploration of higher relationships in skipper butterflies (Hesperiidae) PeerJ 4:e2653 https://doi.org/10.7717/peerj.2653

[5] Xin D, Chen Y-W, Li J (2020) Fine-Grained Butterfly Classification in Ecological Images Using Squeeze-And-Excitation and Spatial Attention Modules. Applied Sciences 10:1681. DOI, 10.3390/app10051681

[6] Alhady S.S.N., Kai X.Y. (2018) Butterfly Species Recognition Using Artificial Neural Network. In: Hassan M. (eds) Intelligent Manufacturing & Mechatronics. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-10-8788-2_40

[7] Liu Aozhi [CN]; Wang Jianzong [CN]; Xia Zimin [CN]; Xiao Jing [CN], "Butterfly identification network construction method and apparatus and computer device and storage medium," CN201810735895A·June 17, 2018.

[8] Wang Jianzong; Wang Yiwen,Zhang Shuang. "A butterfly identification method based on a neural network and a related device", CN201910025247A·November 1, 2019.

[9] Fan Zhun; Huang Longtao; Lu Jiewei; Mo Jiajie; Wu Yuming; Zhu Guijie, "A butterfly automatic classification method based on depth learning," CN201811070920A· September 14, 2018.

Appendix:

Paper details:

1. Paper acceptance:

Thanks and Regards,
Dr. Sharmila Sengupta,
Associate Professor,
Computer Engineering Department,
Vivekanand Education Society's Institute of Technology,
Chembur-74.

On Thu, May 6, 2021 at 9:35 PM ICDIS-2021 <icdis2021@easychair.org> wrote:

> Dear Authors,
>
> Congratulations! On behalf of the ICDIS-2021 program committee and technical committee, we are very pleased to inform you that your paper
> A comprehensive survey on skipper butterflies for Lepidopterists and butterfly enthusiasts
>
> has been accepted as a REGULAR paper for presentation at the ICDIS-2021 and for inclusion in the conference proceedings to be published in Springer Book Series on Lecture Notes in Networks and Systems" ** Indexing: The books of this series are submitted to ISI Proceedings, SCOPUS, Google Scholar and Springerlink **
> This email provides you with all the information you require to revise and complete your paper as per the reviews given below in this mail and submit it for inclusion in the proceedings.
> Here are the steps you have to follow in re-submitting the final version of your paper:
> 1. Please see the Springer's FORMATTING REQUIREMENTS at download section of http://icdis.ssedconferences.org/images/quickoads/Manuscript+guidelines+for+English+books.pdf
> STRICTLY follow the requirements and REVIEWERS' COMMENTS for your paper, which are intended to help you to improve your paper before the final publication. The listed comments should be addressed carefully in your revision, as this acceptance is conditional on your appropriate responses to the formatting requirements and reviewers' comments in your revision. It is mandatory for all authors to incorporate all the changes suggested by reviewers else their paper will be excluded from conference proceedings.
> 2. Each paper must not exceed 10 pages including figures and references. Papers beyond TEN (10) pages are subject to page surcharges) and must not less than 6 pages. All papers must be re-submitted even if the reviewers indicated that no change is required.
> 3. In order for your paper to be published in the ICDIS-2021 conference proceedings, a signed Springer Copyright-Form (Consent to Publish) must be simultaneously submitted for each your accepted paper. (Refer Quick Links section at http://icdis.ssedconferences.org/)
> 4. Register each your accepted paper to the conference by following instructions of registration:

2. Paper publication:

We have presented the paper at the conference on 14th May 2020 and are currently waiting for certification.

3.Plagiarism report:

PLAGIARISM SCAN REPORT

Report Generation Date: May 13,2021

Words: 833

Characters: 5551

Excluded URL :

0%	100%
Plagiarism	Unique
0	**46**
Plagiarized Sentences	Unique Sentences

Content Checked for Plagiarism

Hesperiidae family of Skipper butterfly are one of the major families with over 4000 species of butterflies under it belonging to over 567 different genera. The need to keep track of butterfly hotspots has become increasingly important due to climate change accounting for the decline of their population. The objective of this paper is to provide one authentic source of knowledge about skipper butterflies in India for aspiring researchers. This has been done through a fully integrated comprehensive web application, "Patang Abhidhani". There is a lack of an integrated and regularly maintained data dictionary with skipper butterfly-related information. The data gathering process is quite exhaustive as information is not readily available. The data is cleaned and analyzed before applying an image classification model (VGG-16 architecture with custom layers). This system will allow the Lepidopterists and butterfly enthusiasts to upload relevant images to the website and automatically populate the map with their information.

Keywords: Skipper butterflies · Hesperiidae · Lepidoptera · Butterfly classification · Transfer Learning · VGG-16 architecture · Coliadinae · Eudaminae · Hesperiinae · Pyrginae · Heteropterinae.

The biodiversity available in India is quite vivid. It is home to a plethora of animals, plants, insects, and birds. Due to its tropical location, organisms find it easy to flourish in India. However, due to deforestation and excessive farming, and urbanization, the green cover of the nation has been steadily decreasing. As a result, most species are losing their habitats.[1] including butterflies. Butterflies are some of the most biodiverse insects on the planet. They act as indicators of environmental change and are also an integral part of our ecosystem. They are needed to pollinate plants to acquire fruits and vegetables. Due to their sensitivity to minor changes to the environment and decreasing availability of flowers for pollen, their populations have been seen migrating to different places.

However, in recent years, pollution has gripped our planet leading to a decrease in variety and a count of various fauna. Butterflies are one such affected avian species. Hence, the preservation and categorization of butterflies have become even more important. Among all the butterfly species available in India, study on Hesperiidae (Skippers) is hardly available.

Skippers are a family of Lepidoptera (moths and butterflies) named Hesperiidae. Since they are diurnal, they are often referred to as butterflies. They were historically classified as a separate super family, Hesperioidea; however, the most recent taxonomy positions them in the superfamily Papilionoidea, the Butterfly super family, confirming their status as Butterflies. Their fast, darting flight

habits have earned them the name. The antenna tips of the majority have been modified into narrow, hook-like system ions. Furthermore, most skippers lack the wing-coupling structure found in most moths. They can be found worldwide, but the Neotropical regions of Central and South America have the most variety.

All subspecies of skippers do not look like moths. Looking at all the members of its family, we also notice some of its sub-species to show beautiful colors, which also make them hard to differentiate. Due to the wide range of features that these butterflies possess, many people find it difficult to categorize them. [2]

The website, "Patanga Abhidani" would exhibit detailed information including gallery, heatmap, etc as well as have an option to upload images on Skipper butterflies. Therefore, enthusiasts and common people can further research this family. The user would upload images and details of the butterfly, which would then be sent to the machine learning model to classify it. If verified, the data is then sent to a database, and the results are reflected in the heatmap, which also shows the locations of the butterflies.

Using this data, naturalists, photographers, and enthusiasts can ascertain where exactly skipper butterflies are prevalent to figure out changes in their living patterns and gather more data regarding particular sub-species. This also allows us to record various instances of a butterfly species existing at different places based on which we can better document and understand the species.

2.1 Non-existent automated models :

The major shortcoming of today's systems is that there isn't a system that can automatically detect a butterfly species. The current systems rely on manual verification of the butterfly species by an expert or trust-based systems where what enthusiasts and photographers say is considered valid.

2.2 Image data unavailability :

The second issue is the unavailability of species-specific data. For the process of detection to be automated, the model would need many images. Although butterfly data is available in plenty, images of skipper butterflies found in India are quite a few and far apart. As a result, the database regarding this is quite small.

2.3 No platform for normal people to get their photos verified :

The websites related to butterflies that exist today do not allow any common person to upload images to their database. You need to contact the administrators and show credentials for your images to be featured on their website. The other alternative is Facebook groups which are a lot more chaotic and disarrayed.

PLAGIARISM SCAN REPORT

Report Generation Date: May 13,2021

Words: 1329

Characters: 8492

Excluded URL :

0%	100%
Plagiarism	Unique
0	65
Plagiarized Sentences	Unique Sentences

Content Checked for Plagiarism

In the paper by Hari Thelvaprakasham, they construct a new butterfly dataset with 34,024 butterfly images belonging to 315 species from India. In order to identify butterflies based on images, 11 pretrained networks were tested to get varying degrees of accuracy. [3]
Identification of butterfly species from images taken in their natural environment is a complex and challenging task. This task can be simplified by using a pretrained network called retinaNet which gives mean precision of upto 79% which is far better than recurrent neural networks which is much better than the state of the art systems that exist today. [4]
In terms of classification models, there has been a lot of research on classifying images native to various nations and on a global scale. Still, due to the unavailability of data, the research in India has been stunted. The models are primarily trained to classify butterflies based on stock photos rather than natural photos of butterflies in their habitat [5].
In the paper by R Zhao, C Li, S Ye, and X Fang, the authors used an R-CNN network to classify butterflies based on textures of different butterflies. The images were converted to grayscale to extract relevant features from them, which could then create a recurrent neural network. This data was used to create a co-occurrence matrix to classify 19 different species of butterflies. The accuracy rating lay at around 98%. [6]
Researchers have used pre-trained CNN models called GoogleNets for species-based identification. It works by interpolating data based on four species of butterflies. [7].
Detection of butterflies in images can be a daunting task. This process can be simplified by removing background and then applying a local binary pattern (LBP) to extract relevant information and then use ANNs to process the acquired patterns [8].
Probably the most popular and well-known group of insects is butterflies. India is home to over 1400 species of butterflies, out of which around 218 are classified as skipper butterflies. India also has various hotspots where we can find butterflies. Western ghats are home to over 330 species of butterflies, and they are also found in large numbers in eastern ghats near Bengal and the seven sister states [9].
RK Sahoo and his team explored the biological difficulties in distinguishing between skippers and explored the topologies of Skippers based on markers. [10]
Copepods is another insect species normally found in marine conditions. Although they are very different from butterflies, they are equally rare to find and train on. As a result, a model that is good enough to uniquely identify each species under copepods can be made using ANNs which gives an accuracy of up to 93% on multiclass classification. This study can be used to draw similarities with our

studies on butterflies [11].
Skippers do possess some features that make them distinguishable when compared to other butterfly species. They have triangular wings with acute forward apexed and rounded hindwings. However as the family has over 4000 species of butterflies, each species under it does not exhibit these same characteristics and hence using these specific features for recognition is insufficient [12]

4 Motivation And Uniqueness

The primary reason for research on this topic was that Skipper butterflies aren't well documented in India. Due to rapid urbanization, butterfly counts have been receding and it is important to track the migration patterns of these butterflies to prevent them from going extinct. The proposed system would accomplish this by gathering user submitted data and verifying its validity. The proposed system is unique as currently no websites exist that automatically verify this specific butterfly species and document them. Most current systems rely on manual verification and so the target of this system is to automate the entire process. It also focuses on one specific species rather than trying to categorize all species of butterflies which makes it more fine tuned and adept for accurate classification.

5 System Design And Working

The end product is a web application that allows users to upload their images and also see previously uploaded images by the community. The modules of the website are:
To upload images of butterflies.
Training and implementation of machine learning model.
Generation of a heatmap.

All these modules are interconnected to produce a streamlined process that makes it easier for people to gain information about skipper butterfly species quickly. The user would be able to upload images to the web application which passes them through the ML model. If the model detects a skipper butterfly in the image, it will store it in the database for storage. To prevent inaccurate data from being added to the database, any negative results would be sent to an expert for verification. Once the image and its metadata are stored in the database, it is sent to the heatmap to be added as a marker. The user is asked to provide location details based on which the map is plotted

The Skipper butterflies are not well documented, and so the availability of data required for training is quite low. As a result, we had to fall back on unconventional methods to acquire the required data.

6.1 OCR :

Fig. 1 Sample page of the book

Naturalists during the 1900s had documented the existence of various butterfly species in India [16]. The books written by them are available as open-source documents and need manual extraction as shown in figure 1. On implementing optical character recognition on these books, data on various species was acquired. However, due to the age of the books and varying syntax of documentation, pinpointing correct data proved difficult which can be seen in figure 1, so methods were switched to manual extraction. Other books were also referred for population counts and density information, but no useful information could be extracted from those books due to a severe lack of documentation. An attempt was made to extract species-specific features such as wingspan, size, and so on but to no avail.
6.2 Web scraping :
Many people click photos of butterflies and upload them on various internet sites from which we could extract them. Images for training purposes were obtained from these websites. However, the legitimacy of these images is brought into question. So we have to rely on only trustworthy sources like

other butterfly documentation websites and groups made by naturalists, enthusiasts, or lepidopterists. Such groups were used to extract image data (websites like ifoundbutterflies.org, inaturalist.org, etc.

6.3 Facebook :

Another great source of information is Facebook groups, where naturalists and photographers/enthusiasts upload pictures of butterflies, and people verified those images in the comment section. This data proves useful for Heatmap generation. For scraping this data, Facebook graph API is used along with manual scraping to verify that the contents were accurate. This data, being undocumented, is quite useful as it meets the requirements of our application and is used as the base for heatmap as well as for training purposes. The data acquired were images, species and subspecies name, name of the photographer, and the date on which the photo was taken.

7 Cleaning And Organizing Data

The data collected from Facebook required a lot of analysis because it had certain shortcomings in terms of missing fields for example absence of species name, location, etc. as well as validation of the images being uploaded. Therefore, the missing fields were populated based on the comments on those posts. For validation of images, a butterfly enthusiast was consulted before adding them to our database. The images, along with their corresponding metadata, were stored in a database to be fetched later.

The coordinates of the place are needed to plot the images on a heatmap. But the data acquired from the user regarding the image is usually the name of the place where the image is clicked. Hence all locations are converted to the latitude and longitude format. This can be done manually or by using google maps API to auto-generate coordinates nearest to the place named in the Facebook post.

PLAGIARISM SCAN REPORT

Report Generation Date: May 13,2021

Words: 1061

Characters: 6657

Excluded URL :

0%	100%
Plagiarism	Unique
0	52
Plagiarized Sentences	Unique Sentences

Content Checked for Plagiarism

Python libraries for plotting various maps such as GeoPandas, Folium, Matplotlib, etc. are well known. For our purposes, we have implemented the interactive library in Python called 'Folium', which then visualizes the data into a Leaflet map provided by the Leaflet.js library. The dataset is directly imported from the database into the map using the Pandas library of Python. The main motive behind choosing folium is that it supports Image, Video, GeoJSON, and TopoJSON overlays which helps in highly eye-appealing map content.
The data stored in tables are constantly updated based on the upload feature, and the website is listening for changes in the database; if any changes are made, then the heatmap is regenerated by Flask with new points included as markers on the map. As Leaflet.js is natively supported on browsers, it is not resource-hungry. The fetched data is passed to Folium, which generates a Leaflet map and overlays all the required markers onto the map. It then generates a .html file which is saved as an instance on the server and sent to the user when necessary. Due to Flask being a lightweight framework, the webpage is quickly rendered As the data in the map(Figure 2) is dynamically loaded, users are given the option to filter the values based on location, species, and date. Every time a new set of filtered values is passed, the map plots the filtered ones.

Fig. 2 Heatmap
9 Training model

A deep learning method involving neural networks is used for the system. The first task was feature selection analysis. As many butterfly species look quite similar to each other, categorizing them based on their physical features such as wingspan, length, etc. were unfeasible as there was no accurate source for that data. Instead, a model relying solely on deep learning and image visualization was selected. Neural networks are efficient when it comes to image classification tasks. The end goal of the model was to create a binary classifier that classified images as skipper and non-skipper.
To gather negative training data, more images of butterflies that aren't skipper and are native to India were web scraped and added to the negative dataset. This led to having two types of data (skipper and non-skipper) which were further subdivided into training, testing, and validation sets.
For image-based classification, Convolutional neural networks were selected. CNNs are known to be quite accurate and, at the same time, being efficient in training which allowed the model to be trained for more epochs. The preliminary network trained by us consisted of 3 layers of the convolution along

with two layers of flattening and a final layer of softmax used for classification. The data was split as 70% training, 20% testing, and 10% validation. The images were loaded as 255*255*3.
The model was trained multiple times by changing the learning rate, changing the optimizers by switching between Adam, stochastic gradient descent, RMSprop, and more. The loss function variables were also tweaked to get the best possible and optimized result from the selected network. Even after changing parameters to the best of our capabilities, the model accuracy reached 65%. This was partly due to the unavailability of images for training and the huge scope over which the model needs to be trained.
Based on these results, the next model was selected which was based on transfer learning. Transfer learning with batch normalization was used to add to the complexity of the network to gain more data and recognize more features from the existing data. State-of-the-art pre-trained CNN models were selected for transfer learning like GoogleNets, ResNets, and VGG. Transfer learning was implemented along with the addition of custom layers to the pre-trained model to make it suit our task. The hyperparameters were also fine-tuned to increase the accuracy of the model. The results obtained were much better, with accuracy reaching as high as 89%. These models were then evaluated based on their F-scores to reach acceptable levels of around 0.7 as shown in figure 3.

Fig. 3 Transfer learning model(VGG-16) metrics

10 Website

A user can view the Distribution map and visualize the skipper distribution in India. They can also utilize a filter option based on the species, location, and date-wise filtering. The website has a Login/Register option to store the user details for the upload feature.
However, the upload feature is accessible only to logged-in users. Once a user logs in and uploads the image along with other relevant fields present in the form, the image is passed to the ML model for verification purposes. If the image is of a skipper, that is, if the prediction score is more than 0.5, then it is added to our database and displayed onto the heatmap. This heatmap is constantly updated, and any user that opens the website would be able to see the upload made by any other user.
In this way, the website works as a community for the people interested in studying Skipper butterflies. Using this, various lepidopterists, conservationists, photographers, and butterfly enthusiasts would be able to share the information amongst each other, leading to an increase in the count of images available on the internet for skipper butterflies in particular.

The application has been developed for a nonprofit that will use it in tandem with their website. In the future, the classification model can be further improved by training on the newly acquired data. The model can also be improved by making the process more fine-tuned to recognize skippers' existence and identify which subspecies of skippers it is. The scope can be increased from just skippers to other species of butterflies that exist in India.

12 Conclusion

The diversity of flora and fauna is essential for the well-being of our planet. Many plants and animals rely on them for their sustenance and reproduction. Skipper butterflies are important pollinators, and hence, it is necessary to make sure that their population counts are maintained to save them from extinction. This can be done by tracking their movements around the nation and documenting them to establish the sustainability of the selected species of butterflies. The proposed system enables users to do so by providing them with ample information about the species along with the ability to study their migration patterns. Such a platform thrives off of user generated content to increase the pool of information which is made available to everyone.

4. Project review sheet:

1) Review sheet 1:

Group no 11 INDUSTRY Dr. Sharmila Sengupta

Title of Project: **Patang Abhidhani - A comprehensive guide for Lepidopterists and butterfly enthusiast**

Group Members: Abhijit Pradeep Thikekar, Saurav Sunil Telge,Aniket Ashok Pawar_____

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Research & Innovative Approach (5)	**Total Marks (50)**
4	3	4	3	4	2	2	2	2	2	3	3	4	**5**	**43**

Comments:ML model to be integrated with website

S.Sengupta

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Research & Innovative Approach (5)	**Total Marks (50)**
4	3	4	3	4	2	2	2	2	2	3	3	4	**5**	**43**

Comments:

Date: 17th March,2021

Name & Signature Reviewer2

2) Review sheet 2:

Group no 11 INDUSTRY Dr. Sharmila Sengupta

Title of Project: **Patang Abhidhani - A comprehensive guide for Lepidopterists and butterfly enthusiast**

Group Members: Abhijit Pradeep Thikekar, Saurav Sunil Telge,Aniket Ashok Pawar

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Research & Innovative Approach (5)	**Total Marks (50)**
5	5	5	3	5	2	2	2	2	3	3	3	4	**4**	**48**

Comments: Look into the scalability issues

S.Sengupta

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Research & Innovative Approach (5)	**Total Marks (50)**
5	5	5	3	5	2	2	2	2	3	3	3	4	**4**	**48**

Comments: try to make the database scalable for images

Mamak