RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project report On

DupFree Organizer: Deduplicate and Categorize Images

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Adithyan Darshan Kidav (U2103016)
Aedna Mary Reji (U2103017)
Alan Anu Sam (U2103020)
Allwyn Antony Rodrigues (U2103028)

Under the guidance of
Ms. Sangeetha Jamal
Assistant Professor

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
April 2025

CERTIFICATE

This is to certify that the project report entitled ***"DupFree Organizer: Deduplicate & Categorize Images"*** *is a bonafide record of the work done by* ***Adithyan Darshan Kidav (U2103016)*** *,* ***Aedna Mary Reji (U2103017)*** *,* ***Alan Anu Sam (U2103020)****,* ***Allwyn Antony Rodrigues (U2103028)*** *, submitted to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2024-2025.*

Project Guide
Ms. Sangeetha Jamal
Assistant Professor
Dept. of CSE
RSET

Project Coordinator
Mr. Harikrishnan M
Assistant Professor
Dept. of CSE
RSET

Head of the Department
Dr. Preetha K G
Professor
Dept. of CSE
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Fr Dr. Jaison Paul Mulerikkal CMI**, Principal, RSET, and **Dr. Preetha K G**, HoD ,Computer Science and Engineering for providing us with the opportunity to undertake our main project, **DupFree Organizer: Deduplicate & Categorize Images**.

We are highly indebted to our project coordinator, **Mr. Harikrishnan M**, Assistant Professor, Department of Computer Science and Engineering for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our seminar guide **Ms. Sangeetha Jamal**, Assistant Professor, Department of Computer Science and Engineering, for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, We would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Adithyan Darshan Kidav

Aedna Mary Reji

Alan Anu Sam

Allwyn Antony Rodrigues

Abstract

This project aims to create a complex system that automatically detects, removes and organizes duplicate images in large collections. Using advanced image processing techniques, the system analyzes visual elements such as color, texture, and shape to identify duplicate images. These duplicates are then removed, which is not not only frees up storage space, but also reduces clutter and ensures a more efficient and effective operation manageable collection of images.

After the deduplication process, the system categorizes the remaining images into predefined groups based on their visual similarity. This step improves categorization image collection organization, making it easier for users to find specific images or groups of images based on their content.

The system is designed to handle large data sets efficiently and is capable of processing them large volumes of images with high accuracy and speed. While the cloud deployment aspect primarily supports storage and remote access, allowing users to manage and organize their images from any location, the main strength of the system lies in its ability to automate the entire image management process. This project is particularly valuable for businesses, organizations and individuals who deal with and offer large libraries of digital images a powerful tool for increasing productivity, improving organization and optimizing storage.

Contents

List of Abbreviations

- FPGA - Field-Programmable Gate Array
- CNN - Convolutional Neural Networks
- C-Trans - Convolutional Transformer
- JPEG - Joint Photographic Experts Group
- VGG - Visual Geometry Group
- CLIP - Contrastive Language-Image Pretraining
- AMD - Advanced Micro Devices
- SIFT - Scale-Invariant Feature Transform

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

There is an even more dire need for management of digital images recently. Millions of images are produced every day from fields such as healthcare, media, education, or social media, all baying for attention. Each of these images represents some important bit of information that should be stored and organized as appropriately as possible. Storage and retrieval of an image without proper organization are never really easy; at times, it might be costly and time-consuming. In due course, all these reiterative duplications tend to add up because at different times, the same image has been created over and over, the same image has been saved several times, version errors have occurred, etc. This would waste space, processing costs, and further makes it even more difficult to retrieve particular images on time.

The proposed solution is an automated system to find and erase duplicates from a certain image database related with rest of the organization of images. This would reduce space in storage as well as speed up processing while retrieving photographs. It also allows the users to find images with similar characteristics in viewing, by grouping images automatically together. Traditional methods of collecting and organizing images are labor-intensive and error-prone and therefore is no longer necessary within this digitized age.

The automated system combines hashing algorithms with deep learning models to be an efficient and effective solution. The algorithms will be paired with deep learning models, where hashing algorithms will be by signatures created from each image to find duplicates as compared with ResNet-50 to generate similarity between images based on its content. The application will improve the overall quality and access of image data, making operations in industries that are into digital images facility effective while saving storage costs and human errors.

1.2 Problem Definition

The basis of the work to be done in this project will be constructing an extremely powerful system to manage huge collections of images at the backend to minimize the redundancy of images and organize the rest. With increased dependency on digital photographs, this will serve a customer by providing an entirely seamless automated solution for removing redundant images and sorting the unique images in a clear and organized manner. The project basically automates these tasks in order to improve storage efficiency at the same time quickening the pace at which images can be retrieved and simplifying users' work while providing a good experience.

1.3 Scope and Motivation

This is going to be a project meant specifically for the users or the organizations which need to handle large collections of images to one hence called a single solution that will take care of duplicates and organize images in the respective way. The project will have algorithms developed for automatically finding and eliminating duplicate images as well as grouping the rest of the unique images based on visual features. Advanced techniques for feature extraction and clustering will be applied to datasets of any size and complexity, ranging from personal photo collections to huge institutional picture libraries.

It is a user motivational trustworthy project, especially for the user groups that require handling large volumes of images to make management easier. As there is rapid growth in digital material in large volumes, manual handling of images proves tedious and results in mistakes that lead to scrappy storage, the old cost of space, and wasteful data. Process automation through deduplication and categorization eliminates the need for human efforts. This reduces human errors and provides more accurate data. Eventually, the system will facilitate the easy access and provision of storing images, thereby providing them in a very well-organized and efficient manner to its users.

1.4 Objectives

The most vital objectives of this project are:

1. Automation for detection and elimination of duplicate images in extensive collections.

2. Classification of images based on visual characteristics facilitating navigation.

3. Use of high precision hashing algorithms for duplicate image identification.

4. Storage overhead reduction by preventing redundancy.

5. Improved accessibility of image collections through effective organization.

1.5 Challenges

This system foresees facing some great challenges beyond achieving its overall goals. The first and foremost challenge would be the need for the system to distinguish duplicate images as clearly as possible, even if they may slightly differ because of editing work or different resolutions or rotation. Further, the system should perform effectively and fast on huge datasets without relative limitation or slowdown. At last, the system can act as an effective part of various clustering algorithms that carry categorization. These algorithms need to be tuned for the classification to remain meaningful and allow easy finding of categorized images whenever such images would be needed.

1.6 Assumptions

The following assumptions were made:

1. The users are having large collections of images because small collections won't really benefit much from automated deduplication and categorization.

2. Expected to run in conditions of relatively stable internet connectivity, as with other cloud-based services that perform data processing or storage.

3. Users base understanding of basic principles of data organization would guarantee a better uptake and system usage.

1.7 Societal / Industrial Relevance

This is a project that has social and industrial relevance, particularly in sectors where there is a need to handle very significant volumes of image data. In healthcare, for instance, medical professionals typically have a very huge collection of images such as X-rays or MRIs that need to be well organized for easy retrieval as they help with the preparation of accurate diagnosis and treatment for patients. Similarly, media houses and educational institutions have volumes of visual data that need to be stored efficiently and have easy access as a prerequisite to productivity. There are public sector uses for image deduplication and categorization in historical archives to access critical records more easily.

This project will improve speed through the provision of an automated solution for organizing pictures, thus reducing resource use and bringing in facilities for greater accessibility, which benefits not only an individual user but also any organization relying on a significant count of images.

1.8 Organization of the Report

And so, the report is segmented into chapters, which give it the wider general view. The first chapter is concerned with actually presenting the project, providing necessary background into the project. This chapter defines the problem; specifies the objectives and also the challenges of the undertaking.

This chapter two presents a very elaborate literature review in form of previous studies regarding the problem of integrating language learning into cinema-related activities.

The third chapter deals with the system design that presents the architecture, tools, and workflow of the project. It describes how the project solves the problems of image deduplication and categories through advanced preprocessing, hashing algorithms, and with the use of neural networks like ResNet-50. The structured approach to development is highlighted here through module breakdowns, data, and timelines flow diagrams.

The report is concluded, therefore, in the fourth chapter by looking into the accomplishments of the "DupFree Organizer." It notes that this system efficiently deduplicates a very large dataset of images, provides a saving in the storage space, and improves image organization at the same time. This chapter further speaks of its scalability and relia-

bility, which opens avenues for enhanced systems in streaming deduplication and other features like near-duplicate detection.

The consolidated argument of this chapter explains the background, aims, and relevance of an automated system meant to delete duplicate images as well as organize them. The project focuses on solving the problems of large collections of duplicated images, providing a better avenue for business and individual image usage. This system would, through advanced technology, simplify access to data and save space, as well as minimize errors commonly associated with manual organization. The projects' goals and relevance indicate the increasing need for automation, while assumptions give a good idea on where and how the system will be used. All these points add up to the development of a very solid platform on which to build very efficient tools aimed at improving image management within many other contexts.

Chapter 2

Literature Survey

Digital images, emerging rapidly in much higher volume over the last few years, have started posing challenges in storage, processing, and analysis. Therefore, this chapter reviews significant works on preprocessing, removing duplicates, and classifying images. It examines the methodologies available for faster image processing, repetition reduction, and accuracy improvement. The studies reviewed pertain to making improvements in deduplication imaging techniques, as well as categorizing them along with the merits and demerits of the discussed methods

2.1 A Systematic Investigation of Image Preprocessing on Image Classification (Pegah Dehbozorgi et al., 2024)[1]

Methodology:

The researchers have developed a huge conglomerate dataset of images derived from various subdomains such as nature, medicine, and satellites. Several preprocessing techniques like normalization, resizing, noise reduction, and histogram equalization were explored on these images. To evaluate the effect of all this preprocessing, the researchers trained the image classification models (such as convolutional neural networks) on preprocessed as well as raw unprocessed images and compared the models' performance on both images. The experiments were repeated on different image classification datasets such that the results would be generalized. To statistically analyze the performance difference between preprocessed and raw images, statistical tests were conducted. They analyzed how preprocessing techniques influenced the effect or interaction of factors such as the resolution of images, levels of noise, and imbalance in classes. The aim was to achieve a systematic understanding of how different methods impact the performance of image classification machines.

Results:

It was found through this study that a few data preprocessing methods like normalizing and histogram equalization improved classification accuracies with respect to some data sets. On the contrary, it is found that the effectiveness of any particular method depends on the features of the concerned dataset and on the model used for the classification.

Limitations:

Research mostly results with a data set show that preprocessing methods need adaption to specific data sets. Furthermore, these findings would not hold true for all types of images or domains since different data types have varying needs of preprocessing according to the intended application.

2.2 Novel Hardware Implementation of Deduplicating Visually Identical JPEG Image Chunks (Thang Luong et al., 2024)[2]

Methodology:

The researchers developed a hardware unit using the PXDedup algorithm to eliminate visually identical JPEG image chunks. The unit contained a JPEG Decoder, Chunk Buffer, Pixel Color Average, Subchunk Color Average, Chunk Color Average Buffer, Chunk Hashing, Hash Table, and Store Controller. Such a combination integrated load modules to speedily find and remove duplicate image chunks while leaving the image quality unaffected.

The hardware was developed on Digilent Genesys 2 board comprising Xilinx Kintex-7 FPGA. The researchers have checked the performance of such hardware-based deduplication system with respect to software-based methods. The whole idea was to do speedy operations and compress the size requirement especially in a large-scale storage system.

Results:

The hardware implementation turned out to be quite efficient in eradicating duplicate and visually identical JPEG photographs. These approaches resulted in massive storage savings and made the system respond faster.

Limitations:

The specialized hardware that needs to be developed for deduplication may require huge amounts of initial investment that restricts the smaller bodies or those on limited budgets from adopting it. Thus, the technique may not be accessible or affordable to organizations that cannot outlay the capital to acquire the necessary hardware.

2.3 CE-Dedup: Cost-Effective Convolution Neural Nets Training Based on Image Deduplication (Xuan Li et al., 2021)[3]

Methodology:

By using an image deduplication technique, the researchers have derived a cost-effective technique for training convolutional neural networks (CNNs). CE-Dedup is made up of: Image Chunking, Chunk Hashing, Deduplication, and CNN Training. The first thing to do is chunking the input images, and then use some hash function (e.g., SHA-1) to create unique fingerprints for every created chunk. Record these fingerprints distinctly and delete duplicates in the training dataset. Train the CNN with the deduplicated dataset, thus saving computation and time. The CE-Dedup method was analyzed for its impacts on the training performance and accuracy based on different architectures, namely AlexNet, VGG, and ResNet. The approach that was used to evaluate the efficiency and generalizability of CE-Dedup was standard image classification datasets. There was also a comparison on the cost of training and the model performance of this method against the traditional CNN training methodology.

Results:

The CE-Dedup method has brought out huge savings on the side of cost in computation and time spent during training but only a miniscule effect on the accuracy. Thus, it can be inferred that training these CNNs can be done very efficiently with little regard for performance degradation in the models.

Limitations:

The effectiveness of the CE-Dedup method would be specific to a particular architecture of the neural network and would change with the type of datasets employed. Further

investigation will be needed to comprehend its general applicability since the technique, not all types of neural network models, may demonstrate great efficiency.

2.4 Generation of Novelty Ground Truth Image Using Image Classification and Semantic Segmentation for Copy-Move Forgery Detection (Kang Hyeon Rhee et al., 2021)[4]

Methodology:

The cost-concept development in convolutional neural networks (CNNs) is achieved through image deduplication methodologies. The CE-Dedup entails four main components: Image Chunking, Chunk Hashing, Deduplication, and CNN Training. Here, real input images are chunked into smaller pieces with a hash function (such as SHA-1) which creates individual fingerprints for each chunk, identifies, and removes duplicated chunks from the training dataset. The deduplicated training stock is then passed on to the conversion that feeds the CNN reducing computation and time need. Researchers also examined CE-Dedup's effect on the actual training performance and precision of CNN architectures-including AlexNet, VGG, and ResNet. The experiments were also conducted on the classical image classification datasets to assess the CE-Dedup approach's efficiency and generalizability. The contributors compared CE-Dedup against standard approaches concerning training costs and model performance in CNN training.

Results:

It tightens CE-Dedup, which saves huge computation and training time during minor effects. In fact, it elevates training for CNNs without affecting the performance of the model.

Limitations:

It depends a lot on the accuracy of the classification and segmentation models it uses. Any sort of weakness or error in these models will lead to the mistakes in forgery detection, thus highlighting the necessity of having high precision and quality in the models.

2.5 Summary and Gaps Identified

2.5.1 Summary

Table 2.1: Comparison of Various Image Processing Techniques

Paper Name	**Advantage**	**Disadvantage**
A Systematic Investigation of Image Preprocessing on Image Classification (Pegah Dehbozorgi et al., 2024)	A comprehensive review of different techniques pertaining to performing data preprocessing mechanisms.	Mentioned restrictions will apply to the specific datasets and may hinder generalizability.
Novel Hardware Implementation of Deduplicating Visually Identical JPEG Image Chunks (Thang Luong et al., 2024)	An effective technique for deduplication of visually similar JPEG images.	May involve initial costs.
CE-Dedup: Cost-Effective Convolution Neural Nets Training Based on Image Deduplication (Xuan Li et al., 2021)	Combined image deduplication will lower costs for training.	Does not broadly apply to all kinds of artificial neural networks.
Generation of Novelty Ground Truth Image Using Image Classification and Semantic Segmentation for Copy-Move Forgery Detection (Kang Hyeon Rhee et al., 2021)	C-Trans model has been reported as efficient for different image classification.	Highly weighed on classification and segmentation model accuracies.

2.5.2 Gaps Identified

1. Specific Findings for Datasets: Several works, for example, one related to image preprocessing techniques, show that a particular preprocessing method's successful functioning depend on that certain dataset. Therefore, there is no universal technique of preprocessing used across datasets that do not affect a model's performance.

2. Hardware Implementation Dependence: Very little research on hardware methods for deduplication is as accessible to the smaller institutions; one needs specialized hardware and the costs of such hardware restricts them from being so many organizations rich enough to afford these solutions, rendering many users without any realistic options.

3. Generalizability of Deduplication Techniques: CE-Dedup provides promise for training neural networks for the convolution, but the effectiveness of the method is heavily determined by neural network architectures, and a lot depends on particular setups of data as well.

4. Dependent on Models Accuracy: This technique of forgery detection in copy-move completely depends on how good a classification model and segmentation model can work it in achieving accuracy, because otherwise, these detectors will end up in giving false detection of forgery, and better models are needed to enhance the accuracy of the models.

Up to now, almost everything has touched on image preprocessing, hardware-based duplication, affordable CNN training, and forgery detection. Still, these have the empty gaps waiting upon fresh research. Construction works great are playing forward in the gaps and making the results less dataset-dependent, making the hardware cheaper, making adaptations, and reducing the chances of finding dependence entirely on model accuracy. Further study in preprocessing methodologies might lead to more effective and adaptable outputs.

Chapter 3

System Design

In this chapter, the design and architecture of an automated image filtering system are discussed in addressing image deduplication and categorization issues. Mentioned herein are the system core modules of Preprocessing, Deduplication, Categorization, all with the tools, technologies, and algorithms used. Data flow diagrams, module breakdowns, and project timelines provide a wholesome view of the entire aspect of the structure and process of development of the system.

3.1 System Architecture

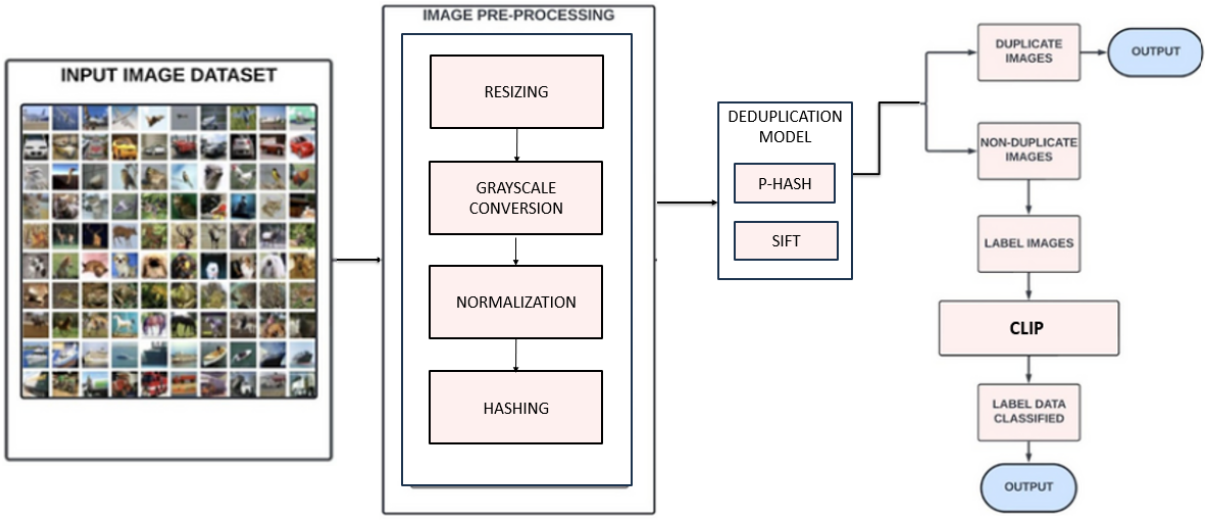


Figure 3.1: Architecture diagram

Image Processing and Classification Pipeline

The image processing and classification pipeline starts off with the image dataset as the input, which contains a composition of unique and duplicate images. Sources of such images are found in any kind of collection, such as private dormant photo libraries or

images pulled from the cloud or from the web by scraping. Whenever raw datasets are said to be dusty, it includes the widespread redundancy of images or images that have little merit in terms of quality; thus, any low-quality images must be pre-processed so that they can be standardized and improved for analysis in the later intervention stages.

3.1.1 Image Preprocessing

Preprocessing is thus all the more important in establishing uniformity, improving quality, and preparing images for the tasks ahead. The first step in this sense is resizing, which puts all the images in standard dimensions that provide efficient processing and allows for compatibility throughout variable models. The next step after resizing is grayscale conversion, which reduces computation load but retains important visual features. The organization may focus on structural features rather than color differences by going down to a single-channel image.

The next stage after gray scaling is the normalization of pixels' intensity values on a fixed scale ((0,1) or (-1,1)). This normalization makes images within the dataset consistent for easier processing by machine learning algorithms. Last of all, hashing is done to produce unique feature representation for every image. Later on, these hash values will be utilized for deduplication; thus, the system can rapidly compare images to detect near-duplicates. The organized way has helped in getting all images to a standard working format, which will certainly facilitate more accurate deduplication and effective classification.

3.1.2 Image Deduplication

After preprocessing, images are deduplicated by eliminating all identical or nearly duplicate creatures. This was performed with the hybrid methods of Perceptual Hashing (P-Hash) and Scale-Invariant Feature Transform (SIFT). In P-Hash, images are transformed to compact hash representations and direct comparison to find out the similarity. The method is comparatively robust toward small variations like change in brightness or some compression artifacts. On the other hand, SIFT detects key visual features in a given set of images and compares respective images based on their features, thus very effective in identifying duplicates whenever images went through other transformations including rotation, cropping, or tiny changes in perspective.

Any duplicate images detected during the deduplication process are discarded from further processing, while those determined to be unique are then passed for classification. The removal of duplicate images saves storage space and improves processing efficiency, only allowing the analysis of images that are meaningful and distinct.

After removing duplicates from the images, the remaining images are classified using CLIP (Contrastive Language-Image Pretraining) - the very large AI model already trained for localized image-text dataset purposes. This is a different approach from traditional classification models, which are label-based, by using text-image matching to generalize and classify images.

In this pipeline, the CLIP first extracts the features of the image and encodes it into a vector representation. These embeddings capture the high-level details of the image, including forms, textures, and its contexts. The system compares these embeddings with the previously defined text labels, which typically include classes like Vehicles, Animals, Food, Natural Objects, and Flowers. Thus, similarity score drives classification, wherein CLIP selects the most appropriate label based on the highest confidence score for that particular image. Each classification result also includes a confidence score which can be used to assess how reliable this categorization is.

This classification process becomes very flexible, scalable, and robust by virtue of the application of CLIP since it can deal pretty effectively with large volumes of image data ranging between highly diverse image datasets without requiring specific training on them. Hence, this makes it an impeccable model for use cases that ask for adaptive image classification.

3.1.3 Output of Labeled Data

The output system after classification provides an organized output with labeled images, metadata, and a well-structured dataset. Each image gets a relevant label, based upon its contents, making the dataset neat and easy to navigate. It may also feature extra metadata, like classification confidence scores, timestamps, and image source details, which are good for further analysis.

The structured dataset produced by this pipeline will have a number of applications. This may be automated photo management, images could be categorized for easy retrieval and organization. This system would also allow for AI-driven semantic search, where

images could be found according to descriptive text queries and would allow for content moderation by ensuring that particular images followed specified guidelines on content, among other applications. Finally, it could lend itself well to recommendation systems, where classified images help suggest relevant content to users.

3.2 Component Design

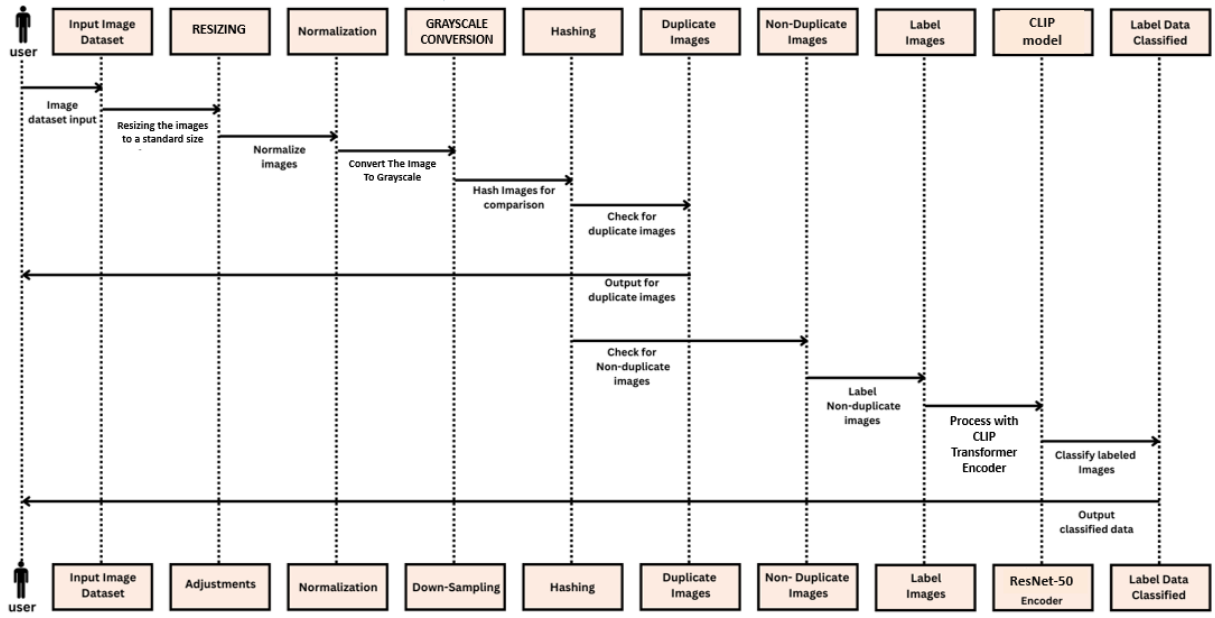


Figure 3.2: Sequence Diagram

This flow chart explains the workflow of an image dataset processing using simple steps to prepare images for classification. This is a succinct description of each stage:

1. Input Image Dataset: Where the process all begins with a batch of visual data to be analyzed.

2. Modifications: The first modification performed on the image to improve the quality of the photo or standardize formats.

3. Normalization: The normalization of images will ensure pixel values are consistent to make them comparable.

4. Down-Sampling: Reduces the resolution of images to save processing time or resources; it still includes critical features in the image.

5. Hashing: A hashing technique is used to give each unique pattern an identifier, enabling easy comparisons between the two images.

6. Duplicate Images Check: The system checks the images for duplicity using their hashes, allowing redundancy to be removed.

7. Processing of Non-Duplicate Images: The images identified not as duplicates undergo further analysis.

8. Labeling Classified Data: The non-duplicate images are classified and their labels defined based on their features.

9. CLIP Encoder: The image encoding by CLIP extracts meaningful features by understanding both visual and textual concepts for semantic similarity.

10. Output: The entire flow culminates in a dataset of labeled images ready for further analysis or machine learning work.

3.3 Data Flow Diagram

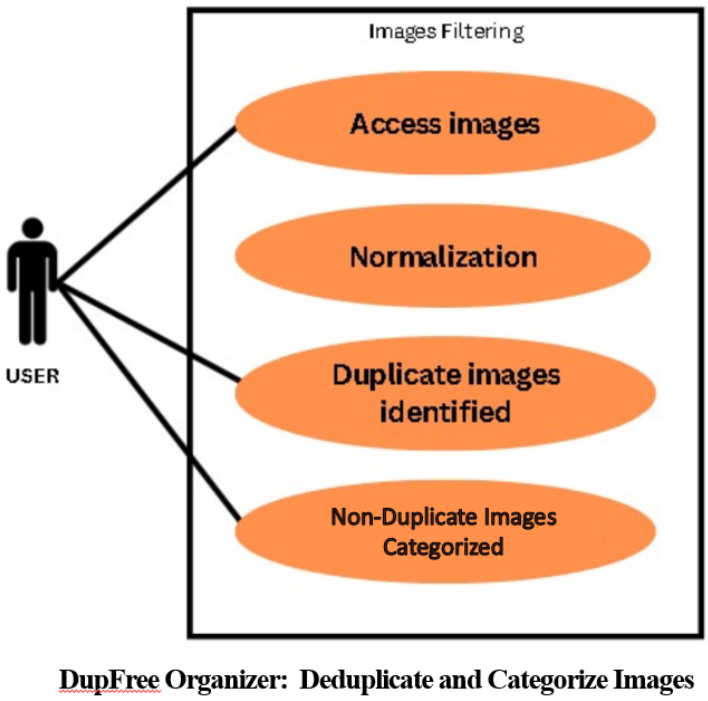


Figure 3.3: Use Case Diagram

3.4 Tools and Technologies

Hardware Requirements:

- RAM: At least 16 GB for running browser smoothly.
- GPU: NVIDIA GTX 1660 (or higher) for reliable performance in machine learning
- Networking: A stable internet connection for google colab and cloud access.

Software Requirements:

- Programming Languages: Python(3.9+)
- Libraries/Frameworks: TensorFlow and Keras for machine learning, Flask/Django for web frameworks, and React for the frontend.
- Cloud Services: Google Drive

3.5 Module Divisions and work break down

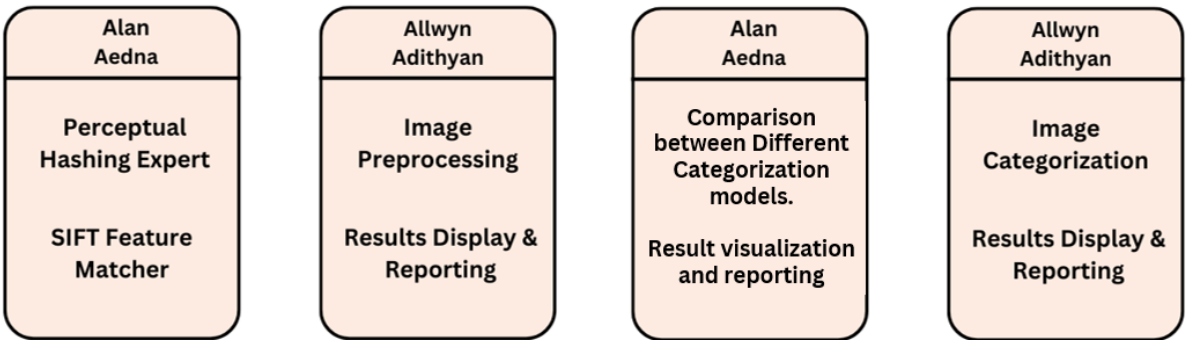


Figure 3.4: Work Division

3.6 Project Timeline

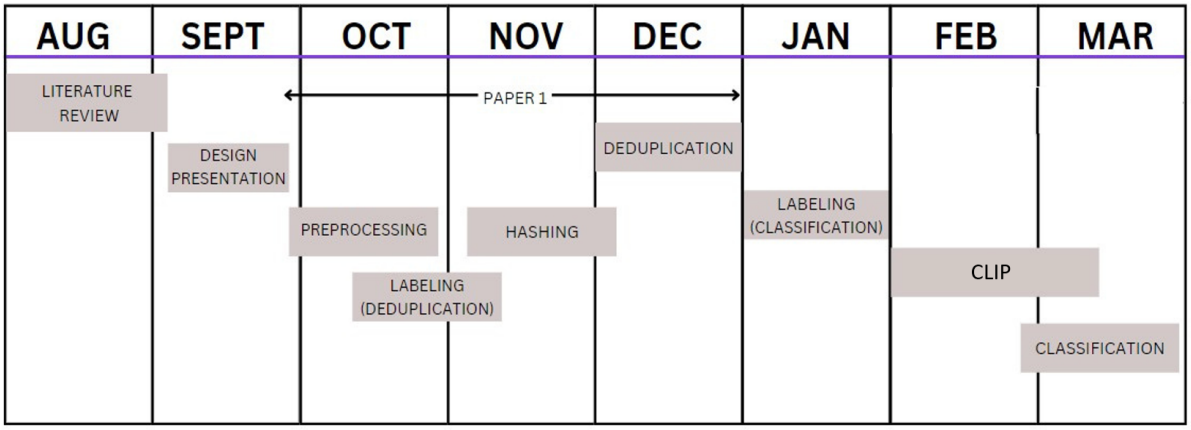


Figure 3.5: Gantt Chart

This chapter introduces an automated image filtration architecture and design which emphasizes modularity in image management while ensuring specifications and performance through advanced preprocessing, deduplication, and categorization processes using p-hash ,SIFT and CLIP model for categorization. Such tools and workflows will act as a base for making a user-centric platform to manage complicated datasets for images, enabling systematic development and on-time delivery of key milestones for execution and qualifying in further developments and tests.

3.7 Comparative analysis of different Object Recognition Models

3.7.1 Swin Transformer

Another major difficulty seen in this Swin Transformer-based model was predicting images accurately for the other seven defined classes. The model's overall performance was quite poor, with an F1 score of 0.43, indicating considerable misclassification. Such striking patterns in prediction were characterized by the model's great tendency to classify those images into "other," suggesting some bias or ambiguity in the definition for that particular category. Further, human and food image classes were often confused through being predicted as other, as the confusion matrix also brought some areas of these confusions to focus. This analysis emphasizes the necessity of further investigations into its architecture,

the training data used, and the characteristics of the other category that can enhance its classification accuracy.

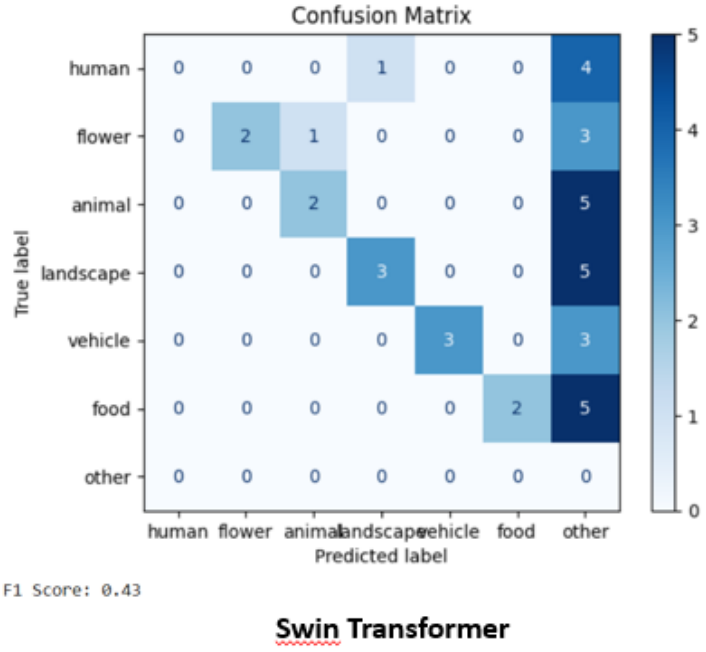


Figure 3.6: Confusion matrix of image categorization using Swin transformer.

3.7.2 Resnet50

The ResNet50 model's outcome in terms of image categorization was higher than that of the Swin Transformer, with an F1 score of 0.72. This means that they were much better in classifying the seven categories correctly in images. In fact, this now clearly showed that ResNet50 had excellent ability in classifying "animal," "vehicle," and "food" images, with very few being misclassified in the procedures. Though that is similar to the Swin Transformer, the model has an area to work on, which involves an "other" category where no images were classified correctly in this group. The model shows further improvement yet still struggles with classifying "human" images. The confusion matrix captured a great reduction in misclassification counts across most compared categories with the Swin Transformer, thus signifying the increased accuracy provided by ResNet50. Moving forward, it would be worthwhile horning in more the "other" category and resolving better human image classification for optimal model performance.

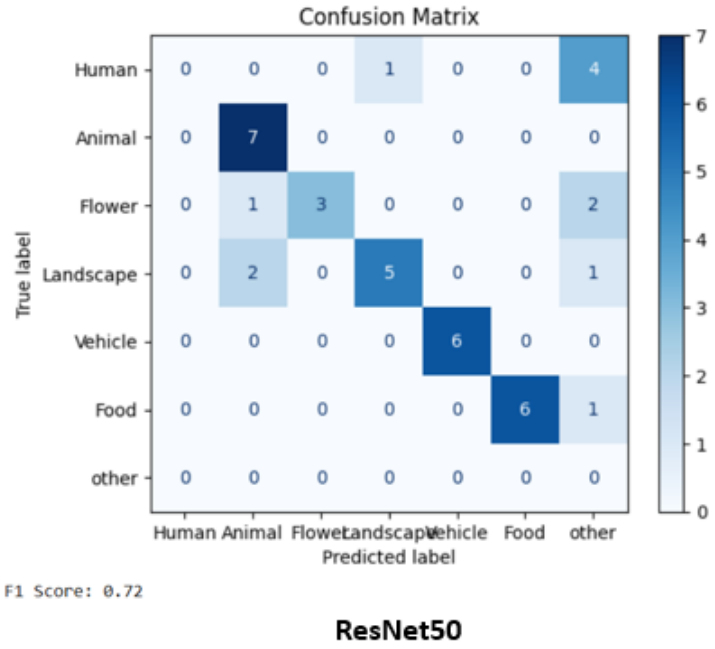


Figure 3.7: Confusion matrix of image categorization using ResNet.

3.7.3 EfficientNetB0

The EfficientNetB0 model has been found to achieve an F1 score of 0.69, reflecting a good performance in the task of image categorization albeit being slightly below that of ResNet50. With regard to classification of images under the categories "animal" and "vehicle," EfficientNetB0 was able to demonstrate the same strong performance as ResNet50. It too struggles, similar to the other models with the "other" category, in classifying images under this particular category. The model also struggled at classifying thscore located at "human" images and misclassification of images "flower" and "food" categories to category "other." On the other hand, EfficientNetB0 clearly outperformed the Swin Transformer, but could not match ResNet50 in performing, primarily in "food" classification. Future work should therefore spend their attention on classifying pets, making the category "other" unambiguous, and misclassified "flowers and food" images as "other" to increase overall accuracy of the model.

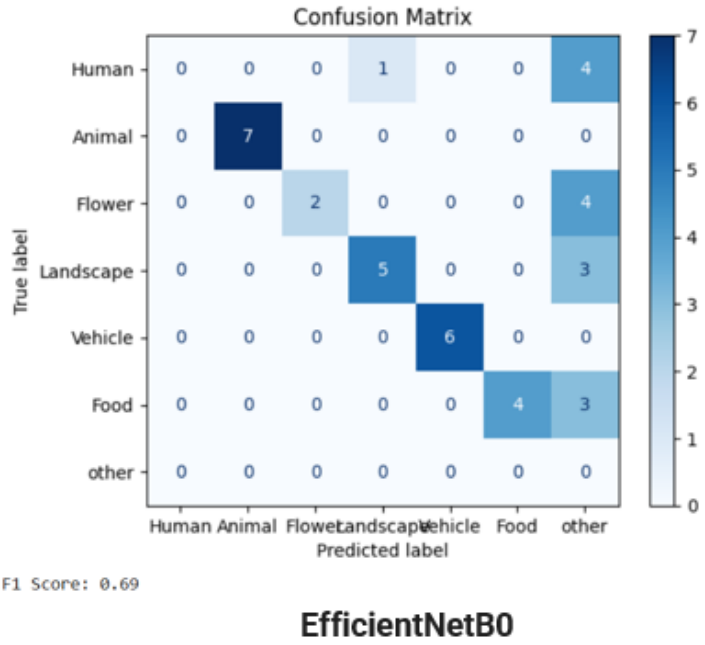


Figure 3.8: Confusion matrix of image categorization using EfficientNetB0.

3.7.4 ConvNextTiny

Moderate image categorization was obtained from ConvNeXtTiny with an obtained score of 0.62 in F1 class. For example, classification was best performed on "animal", while in "human" and "flower" categories, the model misclassified as "other". This is much like what occurred from almost every other model. It performed worse than ResNet50 and EfficientNetB0, indicating a need to work on classification of "human", "flower", and the "other" category to improve overall accuracy.

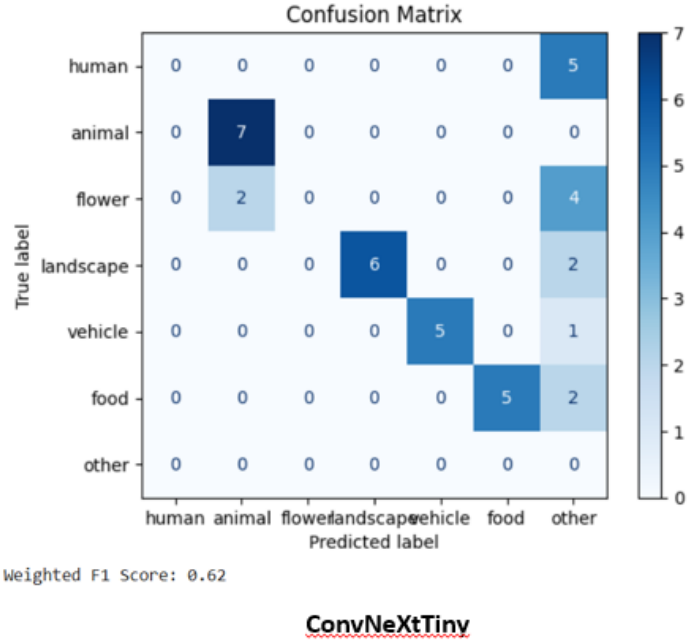


Figure 3.9: Confusion matrix of image categorization using ConvNextTiny.

3.7.5 CLIP

Image categorization is an area wherein the CLIP model performs well, achieving an astonishing weighted F1 score of 0.97. For the most part, the CLIP assignment correctly classified the images from all categories, with perfect or close-to-perfect recall for "animal," "flower," "landscape," and "food" classifications; only one misclassification for "human" proved to be relatively feverish for the model and satisfying in accuracy for "vehicle" classifications. This performance level stands in good contrast to other models tested and indicates the superior ability of the CLIP model to understand and associate images to textual descriptions. This almost neat distinction infers that CLIP makes effective use of its pre-trained knowledge to successfully correlate images and their corresponding categories. Owing to this type of accuracy level, CLIP appears to be an attractive candidate for image categorization tasks as opposed to the well-established convolutional neural networks and transformer-based models.

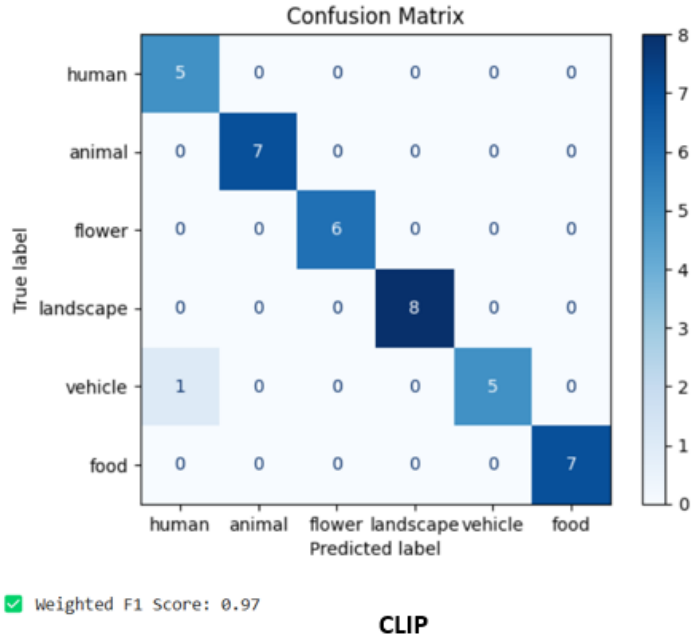


Figure 3.10: Confusion matrix of image categorization using CLIP.

The bar chart presents the error rates achieved by five different image categorization models. It is evident from the chart that there was very inconsistent performance among these models. CLIP model has an outstandingly low rate of error at 0.03, which indicates extremely fine accuracy. Moderate error rates were achieved with ResNet50 and EfficientNetb0, 0.31 and 0.38, respectively, which together indicated a substantial level of performance. Convnext-Tiny has a slightly higher error rate of 0.41, putting it in the middle range. In sharp contrast, the error rates in the Swin Transformer model have marked very high levels of 0.69, making it the least of all five models. This underscores the effectiveness of CLIP and emphasizes the gaping performance differential between CLIP and the rest.

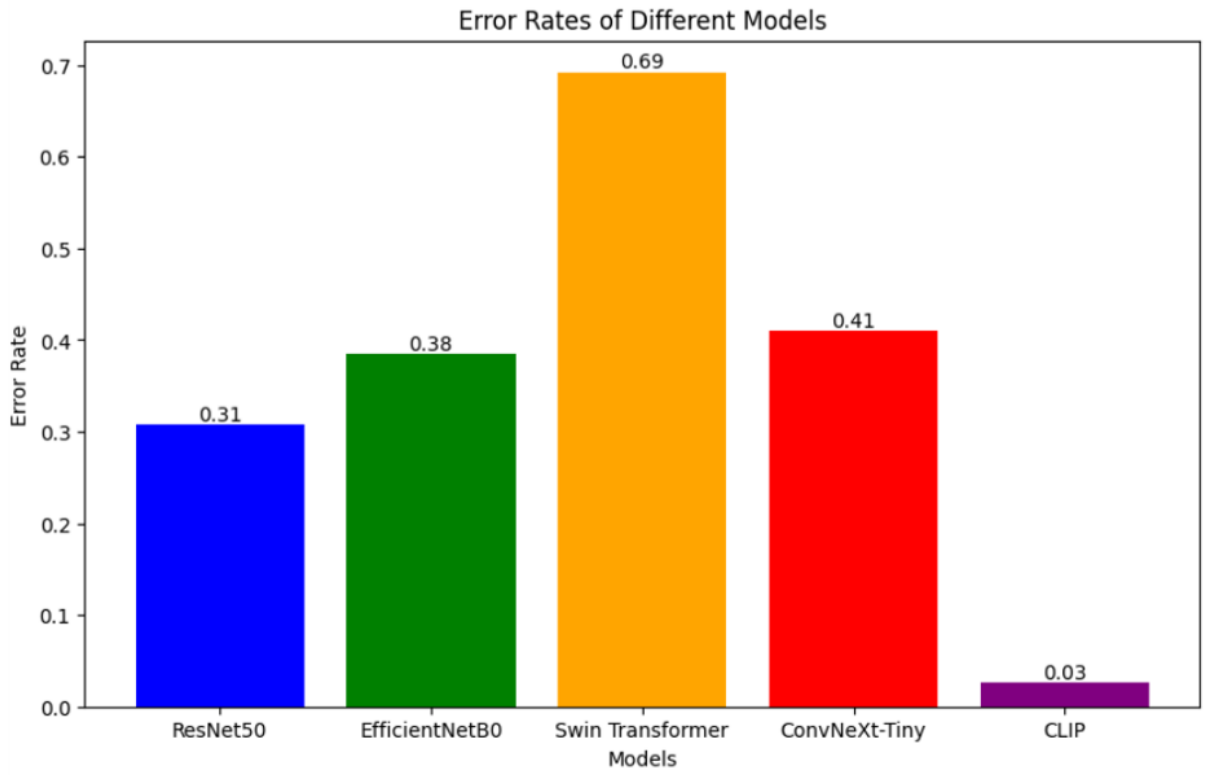


Figure 3.11: Error rates of various categorization methods.

Chapter 4

Results and Discussions

This chapter presents the DupFree Organizer project results and discussions, which consider the act of an image deduplicator and categorizer for large-scale photo libraries. This chapter starts by summarizing the key methodologies that were implemented in the project-PHash (Perceptual Hashing) for fast and effective image deduplication, SIFT (Scale Invariant Feature Transform) for feature matching, and CLIP (Contrastive Language Image Pretraining) model for image categorization. Those techniques were chosen to obtain high accuracy and scalability within the framework of handling a diverse set of image datasets. After that, it goes on to analyze the project's respective outcomes of the methodologies, comparing those with the original objectives. Discussions will include a good review of the effectiveness, insufficiencies, and potential improvements of the outcomes towards judging the success of the said system in tackling concerns on image management.

4.1 Graphical User Interface

The welcome page provides an overview of the project and guides users through the available features.

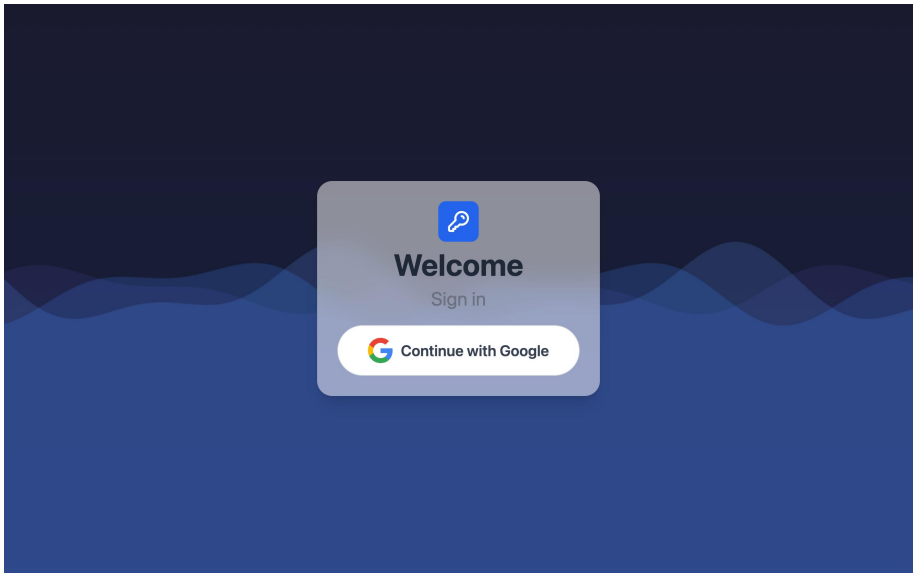


Figure 4.1: Welcome Page.

User authentication is handled through the sign-in page, which connects to Google Drive, allowing secure access to image datasets.

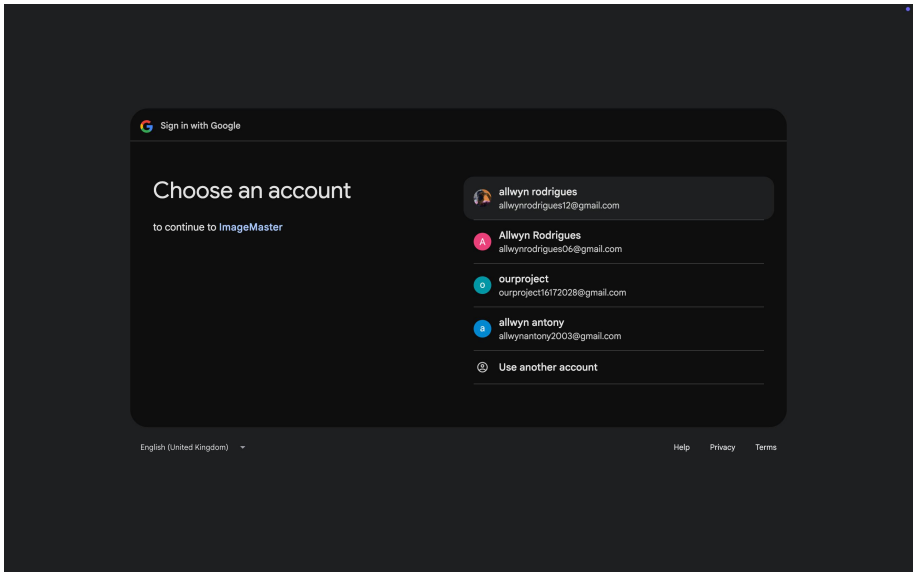


Figure 4.2: Sign-in Page.

The front-end of DupFree Organizer consists of multiple web pages designed to provide a seamless and user-friendly experience. The home page serves as the central hub, introducing users to the system and its functionalities.

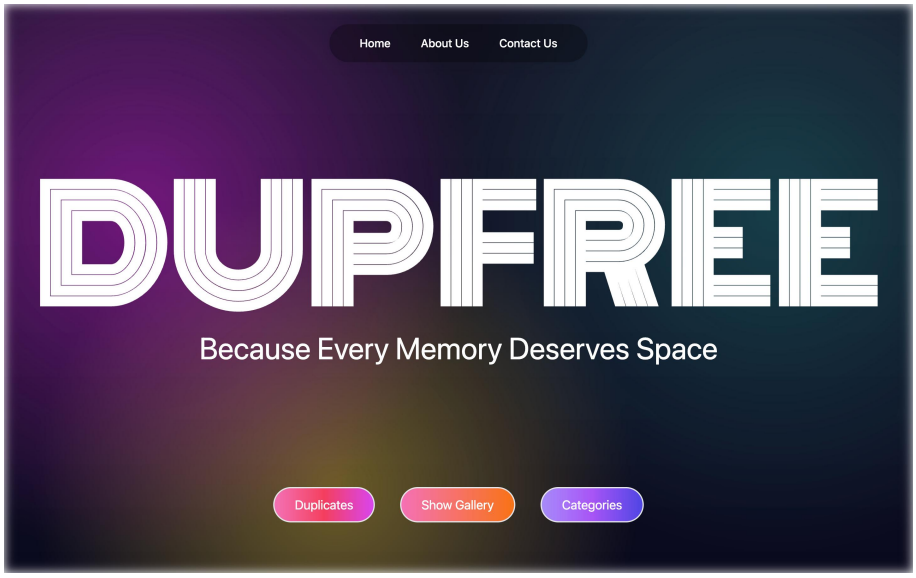


Figure 4.3: Home Page.

The about us page offers insights into the project's objectives, methodology, and team, while the get in touch page provides contact details for support and inquiries.

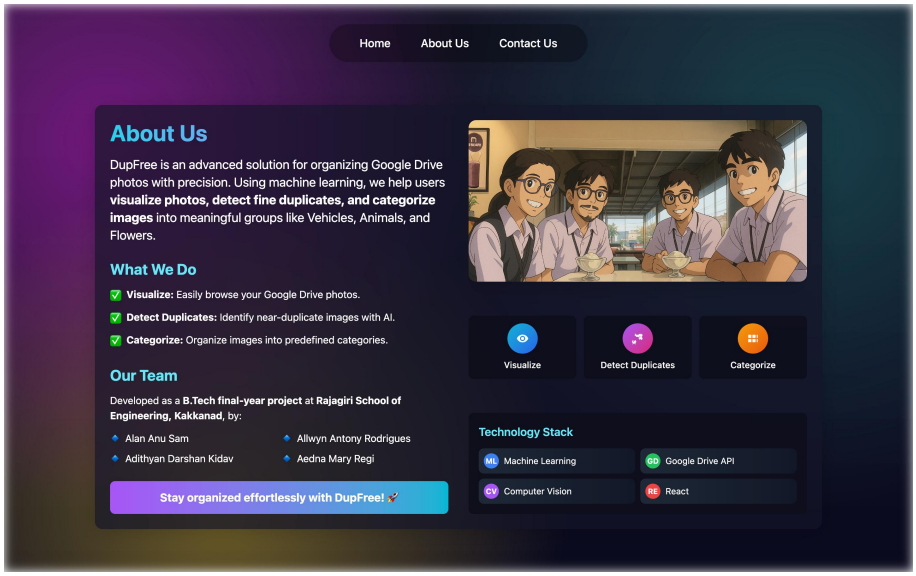


Figure 4.4: About Us Page.

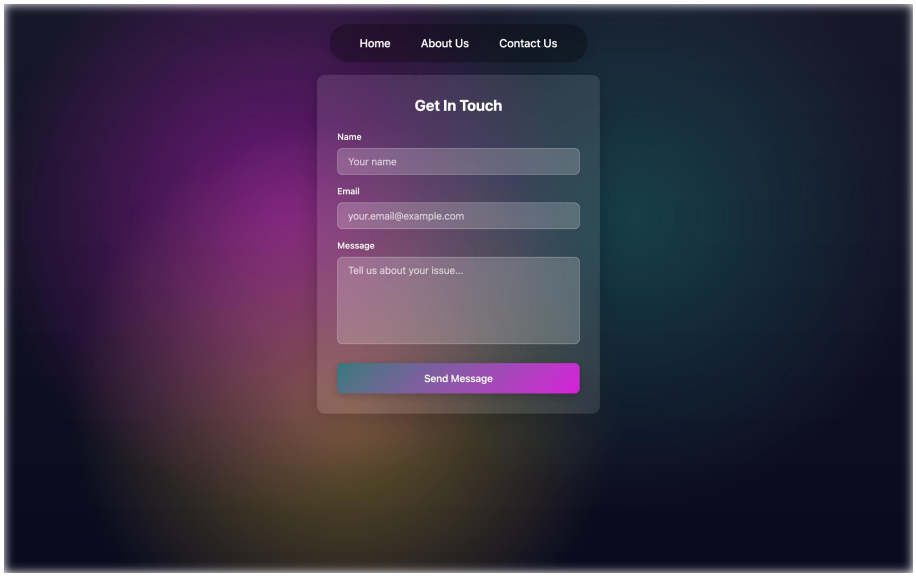


Figure 4.5: Get in touch Page.

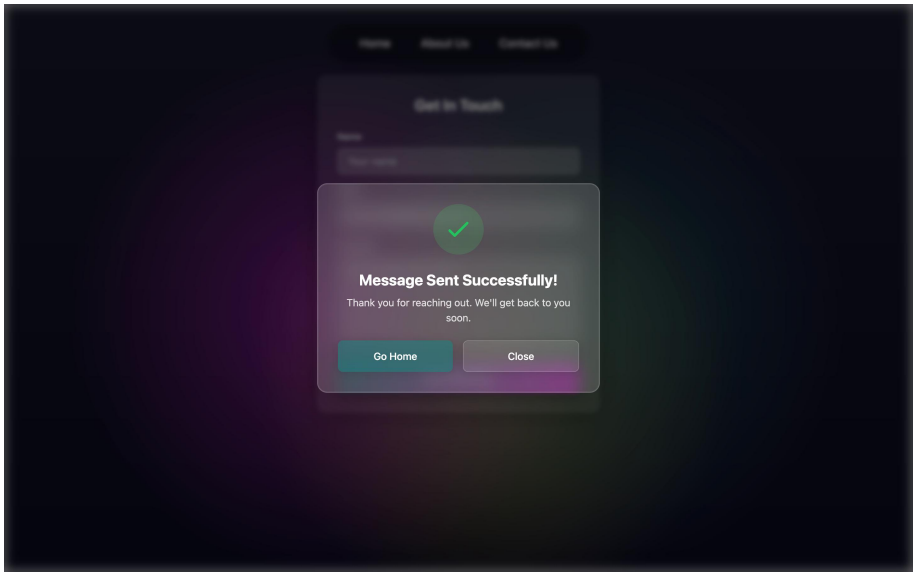


Figure 4.6: Message sent Page.

The following sections include screenshots of these pages, illustrating the layout, design, and core functionalities of the front-end interface.

4.2 Deduplication Performance

The duplication removal procedure, through both P-Hash and SIFT, manages to identify and delete images in duplicate form. P-Hash builds up unique perceptual features Hash-based visual signatures per each image. By using this method, the system could recognize

duplicates even when minor changes are made to the image (for example resizing, cropping, color-change, or contrast variations). Another method to compare images is SIFT (Scale-Invariant Feature Transform), which is based on detecting key features in an image and then using these features for comparison purposes, even if the two images are on a different scale or orientation.

4.2.1 P-Hash

The image compares Image 26, a color photo featuring a red sports car parked on a city street, and its counterpart Image 66, a grayscale rendering of the same scene. A hashing algorithm has returned a 98.83% similarity between two images, thus asserting near-equal content and composition, with Image 66 being devoid of color, as the only difference. Such a high degree of likeness is further corroborated by the accompanying bar chart demonstrating near-perfect similarity scores between Images 26 and 66. All visual and numerical evidence firmly backs up Image 66 as the grayscale conversion of Image 26, which the hashing method aptly detected as a highly similar duplicate.

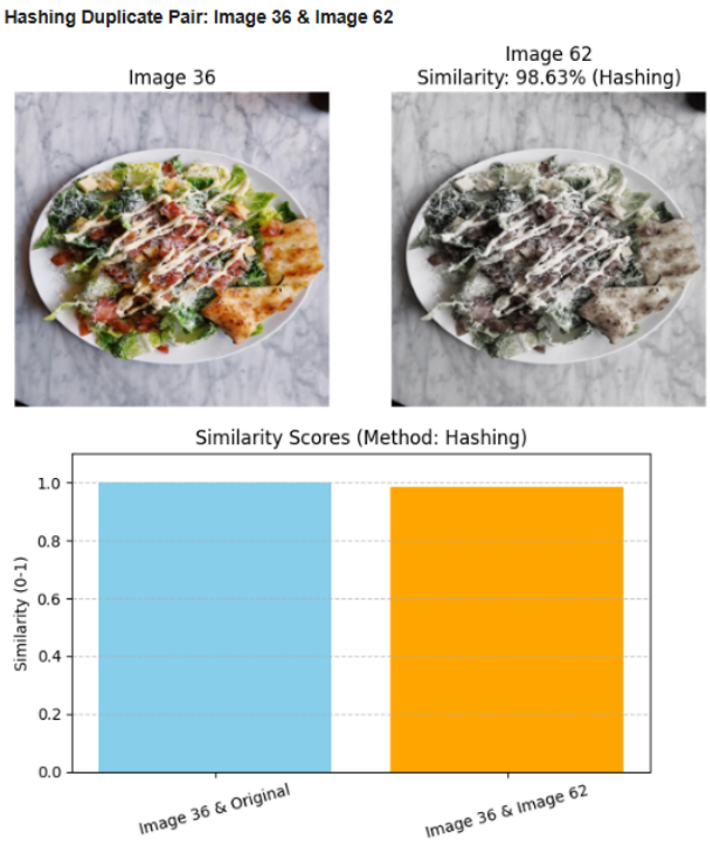


Figure 4.7: Deduplication using hashing along with similarity Scores.

The image compares Image 26, a color photo featuring a red sports car parked on a city street, and its counterpart Image 66, a grayscale rendering of the same scene. A hashing algorithm has returned a 98.83% similarity between two images, thus asserting near-equal content and composition, with Image 66 being devoid of color, as the only difference. Such a high degree of likeness is further corroborated by the accompanying bar chart demonstrating near-perfect similarity scores between Images 26 and 66. All visual and numerical evidence firmly backs up Image 66 as the grayscale conversion of Image 26, which the hashing method aptly detected as a highly similar duplicate.

Figure 4.8: Deduplication using hashing along with similarity Scores.

It also compares the head-on photograph between two images, namely at Image 30, of an airplane against the cloudy sky, with Image 42, absolutely only slightly different from the previous image. The two images are highly similar, based on a hashing-based similarity score of 84.18%, having minor differences mostly due to cropping, resizing, or even little contrast adjustments. A barChart supports this finding of a perfect self-similarity score of 1.0 with Image 30 but left to match it with Image 42 still is significantly

high. The hashing method has efficiently determined that Image 42 is a near-duplicate by taking into account all those small differences between the two images.

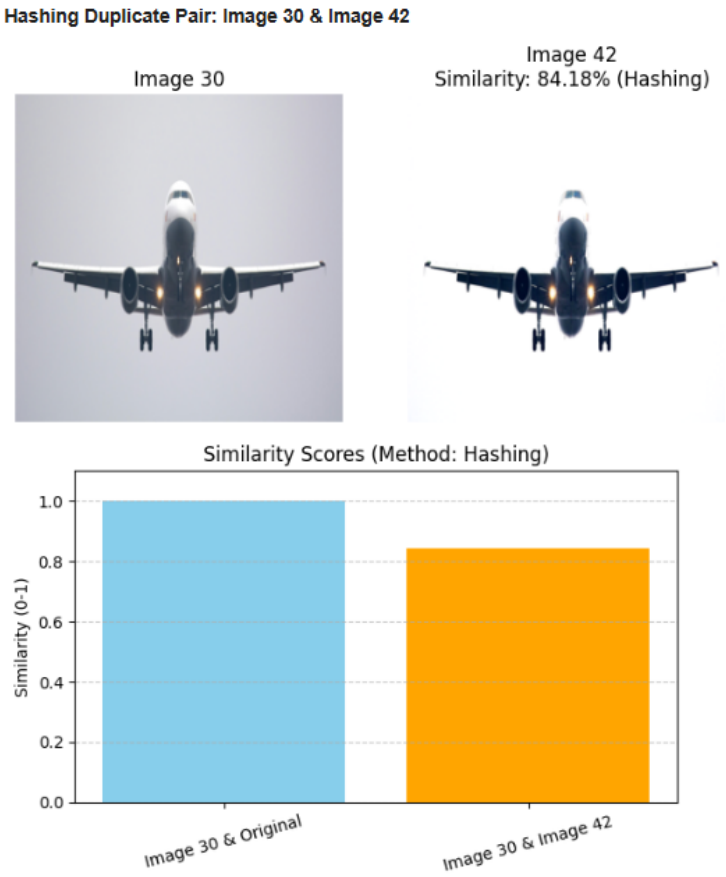


Figure 4.9: Deduplication using hashing along with similarity Scores.

4.2.2 SIFT

Image 69 and Image 70, featuring a laptop, are compared using the SIFT method, yielding a 42.60% similarity. The bar chart confirms this, showing a significantly lower similarity between the two images compared to the perfect match of Image 69 with itself.

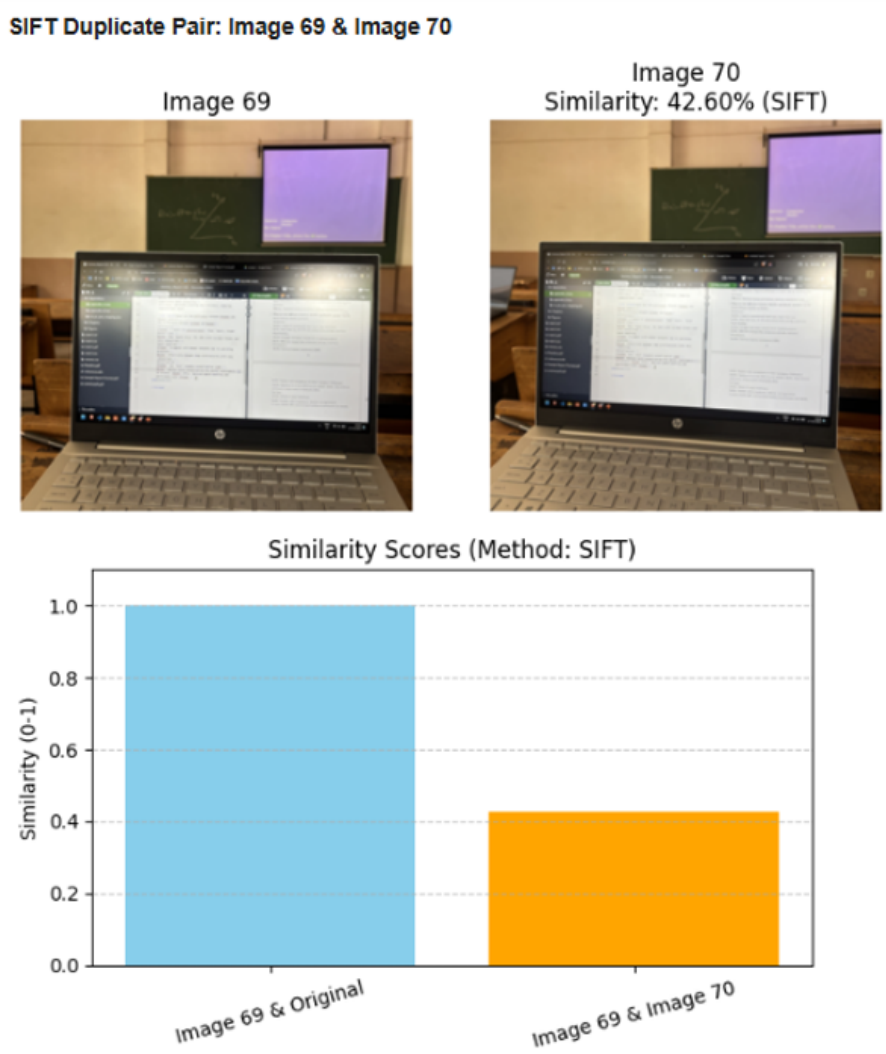


Figure 4.10: Deduplication using SIFT along with similarity Scores.

The differences and similarities between the images were then compared by applying the SIFT method. The cross-comparison of Image 11 against Image 49 gives a percentage similarity of 38.89%. Functionally, such low observational scores-as purposely engineered into the model-derive out the fact that the subject is common amongst both images but never resolved the question of duplicate-hood. This theory finds further support in the Bar Graph where one observed a perfect score of 1.0 between Image 11 against Image 11 and an up-and-down definitive drop in the score for the case of Image 11 against Image 49.

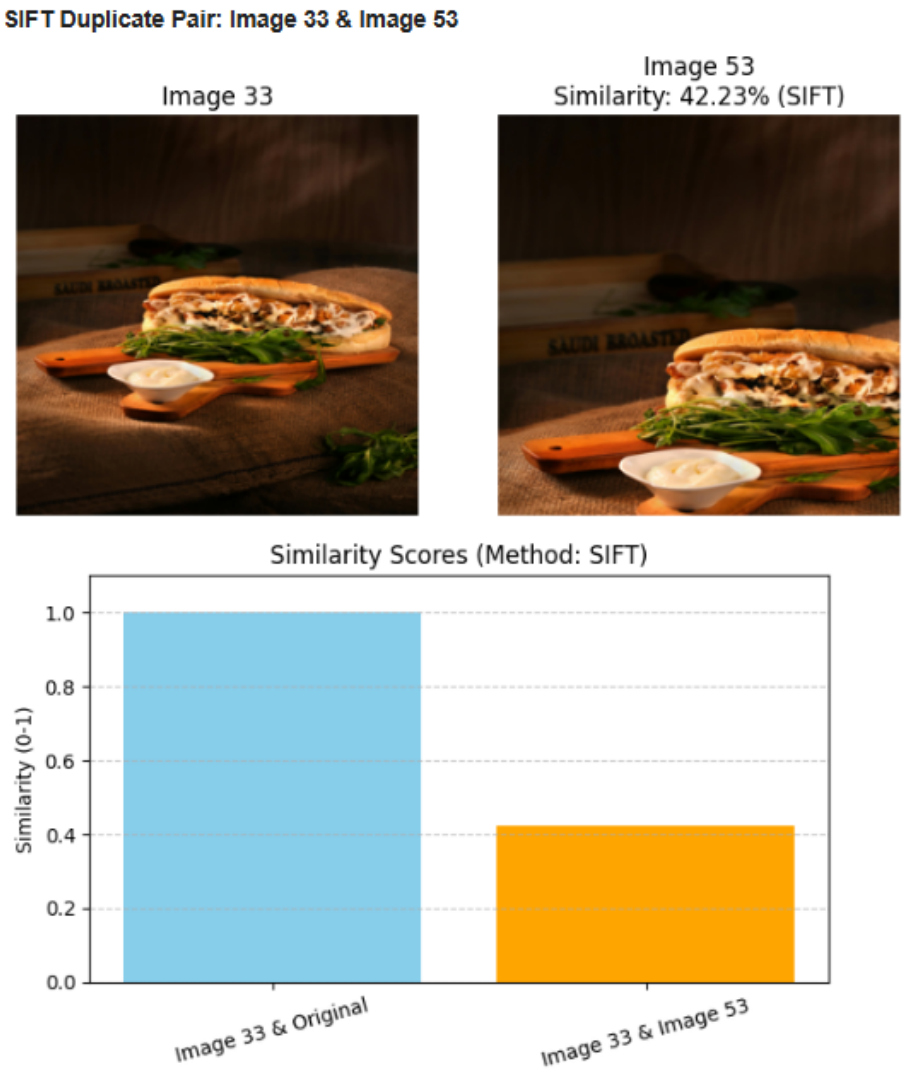


Figure 4.11: Deduplication using SIFT along with similarity Scores.

The Image 33 value depicts a very well-lit photograph of a sandwich luncheon that has been properly arranged on a rustic wooden table, contrasted with Image 53, which has a slightly altered view of the same scene. The SIFT-based similarity score of 42.23% would imply that the visible elements common to these two images do not put the fact of their visible difference out of contention, obviously, for reasons concerning cropping, lighting, or perhaps a slight change of perspective. Indeed, the corresponding bar chart serves to back this claim, as we see perfect self-similarity from Image 33 with a score of 1.0, whereas its score with respect to Image 53 drops substantially. Thus, the SIFT method was effectively used in labeling Image 53 as a partial duplicate, for it points out structural similarity but also takes into ample account the changes that distinguish it from the original.

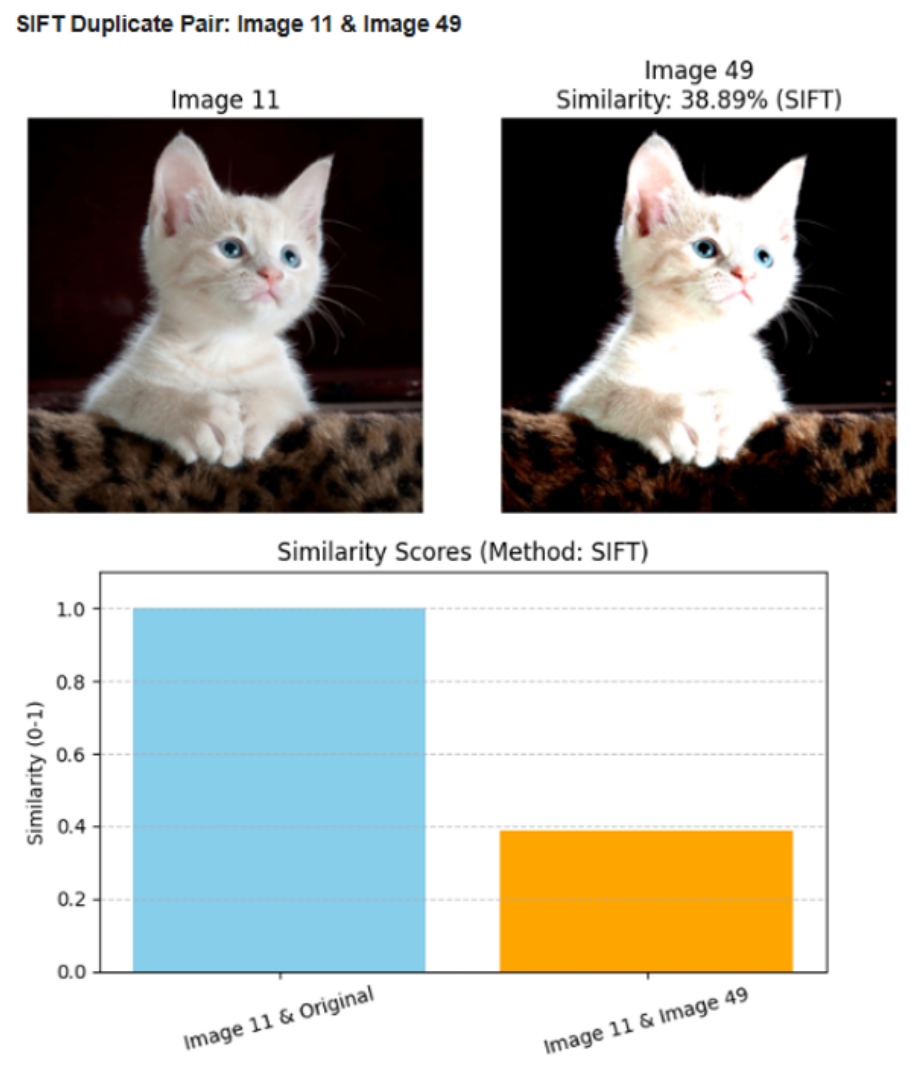


Figure 4.12: Deduplication using SIFT along with similarity Scores.

Classification of all files based on a set of images per user's request into a test dataset, and the method that showed greater similarity was chosen. The method chosen, along with the file name, is provided in the format of a table below, showing each image's similarity score.

Perceptual Hash Similarity Duplicates

Image Names	Similarity (%)
food 5.jpg ↔ dfood 5.png	98.63%
vehicles 5.jpg ↔ dvehicles 5.png	84.18%
vehicles 1.jpg ↔ dvehicles 1.png	98.83%
landscape 4.jpg ↔ dlandscape 4.png	96.09%
animal 2.jpg ↔ danimal 2.png	96.09%
vehicles 3.jpg ↔ dvehicles 3.png	84.38%
random 5.jpg ↔ drandom 5.png	84.18%
human 2.jpg ↔ dhuman 2.png	96.68%
landscape 3.jpg ↔ dlandscape 3.png	77.54%
flower 3.jpg ↔ dflower 3.png	97.46%
animal 4.jpg ↔ danimal 4.png	86.13%
random 3.jpg ↔ drandom 3.png	99.41%
landscape 6.jpg ↔ dlandscape 6.png	72.66%
IMG_1409.jpg ↔ IMG_1408.jpg	58.40%

SIFT Similarity Duplicates

Image Names	Similarity (%)
human 5.jpg ↔ dhuman 5.png	66.02%
flower 5.jpg ↔ dflower 5.png	47.62%
animal 1.jpg ↔ danimal 1.png	38.89%
animal 7.jpg ↔ danimal 7.jpg	44.04%
landscape 2.jpg ↔ dlandscape 2.jpg	33.94%
landscape 8.jpg ↔ dlandscape 8.jpg	43.61%
food 2.jpg ↔ dfood 2.jpg	42.23%
food 4.jpg ↔ dfood 4.png	69.83%
food 7.jpg ↔ dfood 7.png	64.55%
random 6.jpg ↔ drandom 6.png	68.93%
IMG_5403.jpg ↔ lap1.jpg	42.60%

Figure 4.13: Similarity scores of SIFT and Hashing.

4.3 Categorization Performance

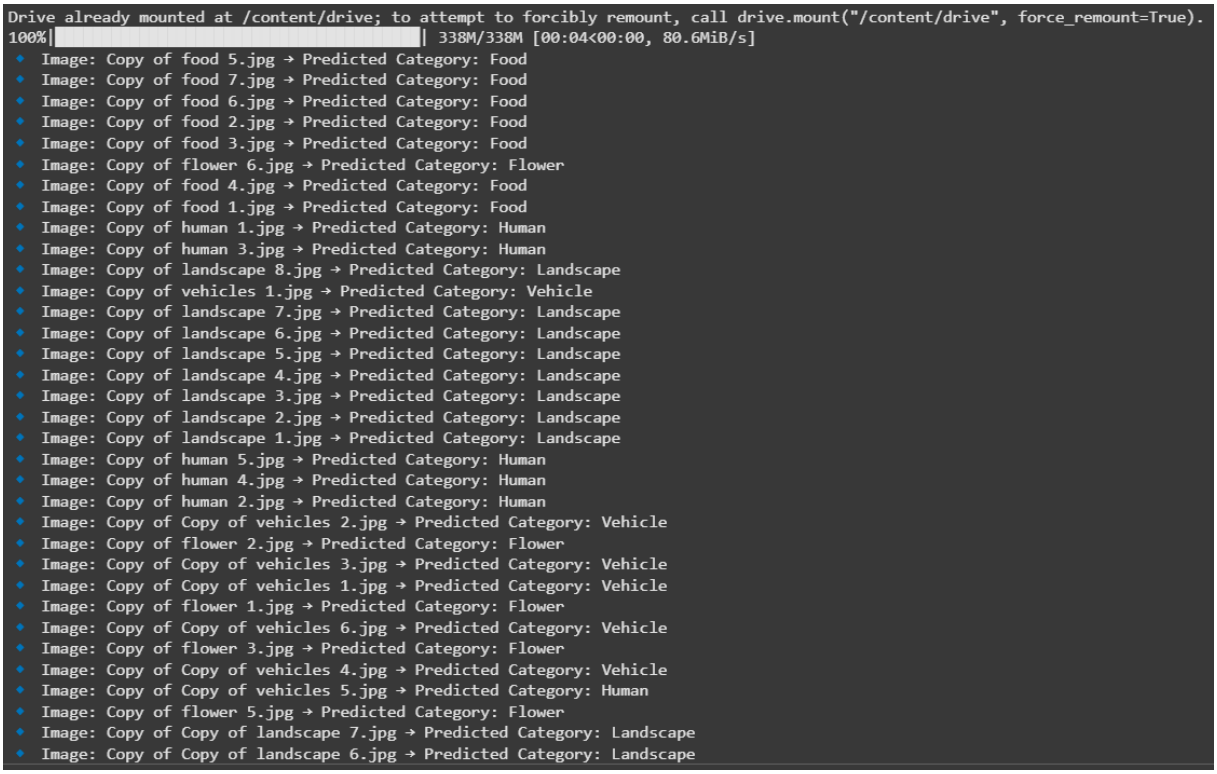


Figure 4.14: Similarity scores of SIFT and Hashing.

The results of this output indicate whether an image belongs to one of six classes: - Flower, Vehicles, Animals, Human, Landscape, or Food. The entire classification was done through CLIP model. In each line, the name of the image file is made with the predicted category of that image depending on the content in it.

Hence, images bearing food contention such as "Copy of food 5.jpg" or "Copy of Copy of food 7.jpg" can invariably be classified in the category of *Food*. Image personalities like "Copy of human 1.jpg" together with "Copy of Copy of human 7.jpg" are classified by the model as *Human*. Likewise, the image of automobiles and other vehicles are appropriately qualified and labeled under *Vehicle*.

Although some of the filenames tend to repeat, while in some cases, they are slightly altered (like "Copy of Copy of vehicles 7.jpg"), the model still assigns the same category correctly. This means that the model treats visually similar images in the same manner with respect to their filenames.

Also, it would denote the clear segmentation between classes. For a flower which

does not end up mixed up with landscapes or animals, each and every such image is correctly grouped into its corresponding category. This output is adequate for the system verifying that a wide range of imagery classifies clearly into various categories-with no visual misclassifications scaling in the examples presented.

Chapter 5

Conclusions & Future Scope

DupFree Organizer is based on an established mainstream system and delivers excellent performance in terms of implementing duplicate image identification and deletion on a large scale with huge image collections. The use of P-Hash and other advanced image hashing techniques creates unique visual signatures that can even enable the detection of duplicates in cases of minor editing and resolution differences. De-duplication algorithms further enhance this by achieving a high level of accuracy and speed while making efficient use of storage space. This is indeed a great step toward automating the management of vast image datasets. The deduplication module supports all kinds of file formats and types of visual content and is therefore scalable and robust, making it suitable for large-scale implementation in different domains. DupFree Organizer optimizes the use of storage space while decluttering the image base and is thus a strong and efficient solution for the effective management of large digital collections.

In terms of the future, there are many possibilities to further expand and improve the capabilities of DupFree Organizer. Some of the advanced ML techniques like deep learning-based models could help enhance the system in classifying images in more elaborate and subtle ways. Other imprints could add real-time processing capability-batch processing for newly uploaded images would streamline image management workflows. Improving the system's ability to handle extremely huge datasets regarding processing time memory management needs some enhancement. Further, future development will aim at integrating a friendly user interface allowing the user to interact with and control his management of image collections. Such steps would open up for even more applications and help enhance the overall utility of the DupFree Organizer for different users and industries.

References

[1] H. R. Kang, "Generation of novelty ground truth image using image classification and semantic segmentation for copy-move forgery detection," *IEEE Access*, vol. 10, pp. 10 123–10 134, December 2021.

[2] X. Li, L. Chang, and X. Liu, "Ce-dedup: Cost-effective convolutional neural nets training based on image deduplication," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, September 2021, pp. 1123–1130.

[3] P. Dehbozorgi, O. Ryabchykov, and T. Bocklitz, "A systematic investigation of image pre-processing on image classification," *IEEE Access*, vol. 12, pp. 567–580, April 2024.

[4] T. Luong, L. Dinh, H. Nguyen, and L. Tran, "Novel hardware implementation of deduplicating visually identical jpeg image chunks," *IEEE Access*, vol. 12, pp. 1223–1235, May 2024.

[5] T. Li, Z. Zhang, L. Pei, and Y. Gan, "Hashformer: Vision transformer based deep hashing for image retrieval," *IEEE Signal Processing Letters*, vol. 29, pp. 827–831, 2022.

[6] Y. Peng, J. Zhang, and Z. Ye, "Deep reinforcement learning for image hashing," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2061–2073, 2020.

[7] S. Li, L. Wang, J. Li, and Y. Yao, "Image classification algorithm based on improved alexnet," in *Journal of Physics: Conference Series*, vol. 1813, no. 1. IOP Publishing, 2021, p. 012051.

[8] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "Resmlp: Feedforward networks for image classification with data-efficient training," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 5314–5321, 2022.

[9] S. Ghazanfari, S. Garg, P. Krishnamurthy, F. Khorrami, and A. Araujo, "R-lpips: An adversarially robust perceptual similarity metric," *arXiv preprint arXiv:2307.15157*, 2023.

Publications

Conference name: 14th Computer Science On-line Conference 2025.

Date: April 23-26, 2025.

Status: Accepted.

To be published in : Springer Series: Lecture Notes in Networks and Systems -ISSN 2367-3370.

Appendix A: Presentation

1

DupFree Organizer: Deduplicate and Categorize Images

GROUP 15 MEMBERS:
ADITHYAN DARSHAN (U2103016)
AEDNA MARY REJI (U2103017)
ALAN ANU SAM (U2103020)
ALLWYN ANTONY RODRIGUES (U2103028), S8 CS alpha

GUIDED BY: Mrs. SANGEETHA JAMAL
Assistant Professor,DCS,RSET

2

CONTENTS

PROBLEM DEFINITION

Manually managing large image collections, particularly when it comes to deduplication, is inefficient and prone to errors. If the collection contains no duplicate images, no deletions will occur. However, without deduplication, unnecessary storage is consumed, and locating specific images becomes challenging.

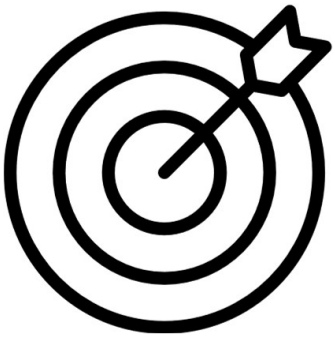
PURPOSE & NEED

- Effective image management is crucial in healthcare, media, and education.
- Challenges arise as image collections grow, requiring automation.
- Automating deduplication and categorization helps by:
  - Improving retrieval efficiency
  - Saving storage space and costs
  - Reducing human errors, ensuring better data quality

PROJECT OBJECTIVE

The main aim of our project is to develop an automated system that can:

- **Detect and eliminate duplicate images from large datasets.**
- **Categorize the remaining images based on visual similarities, making it easier for users to navigate their collections.**
- **Optimize storage and improve the overall management of images, enhancing user experience.**



LITERATURE SURVEY

PAPER NAME	KEY TAKEAWAY
CE-Dedup: Cost-Effective Convolution Neural Nets Training Based on Image Deduplication (Xuan Li et al., 2021)	• VGG's simple yet deep architecture enables efficient feature extraction. • Its small 3x3 filters support pruning and quantization, reducing costs while maintaining accuracy.
Novel Hardware Implementation of Deduplicating Visually Identical JPEG Image Chunks (Thang Luong et al., 2024)	• Perceptual Hashing detects visually identical JPEG chunks by hashing image content, not exact pixels. • It enables efficient deduplication while tolerating minor variations like compression artifacts.
A Systematic Investigation of Image Preprocessing on Image Classification (Pegah Dehbozorgi et al., 2024)	• Resizing standardizes dimensions, reducing cloud computation costs. • Normalization stabilizes models by scaling pixel values. • Grayscale conversion cuts storage and speeds up processing while keeping key details.
Generation of Novelty Ground Truth Image Using Image Classification and Semantic Segmentation for Copy-Move Forgery Detection (Kang Hyeon Rhee et al., 2021)	• C-Tran blends CNNs' feature extraction with Transformers' self-attention for better accuracy and efficiency. • It excels in copy-move forgery detection by capturing spatial relationships and fine details.

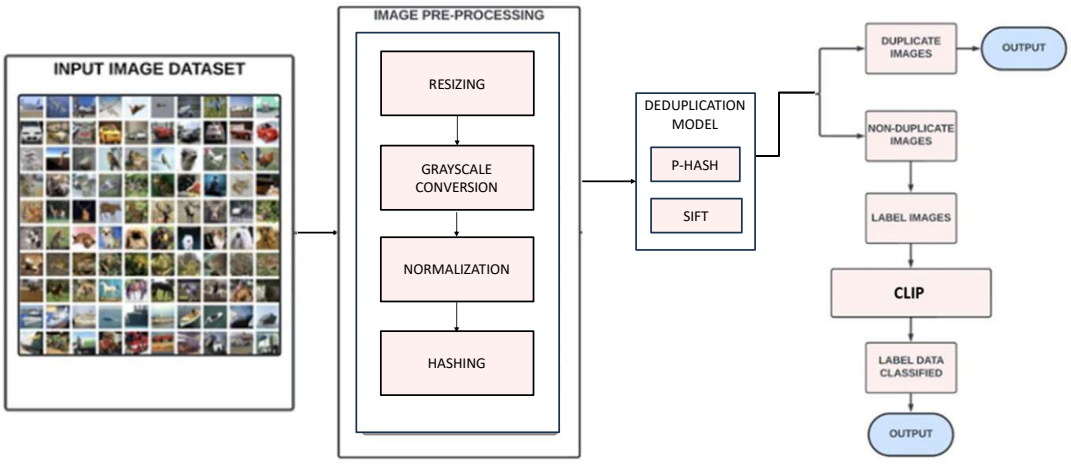
7

METHOD OUTLINE

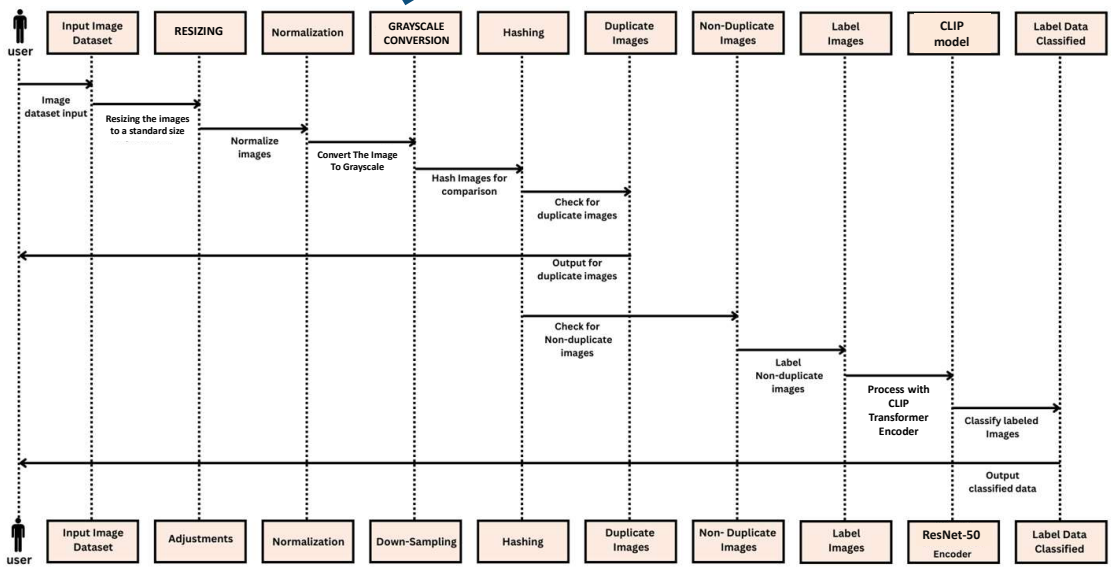
- Clean up the images by pre-processing techniques like resizing and grayscale conversion.
- Apply hashing algorithms to identify and remove duplicates.
- Used CLIP neural network model for grouping all the remaining images.

8

ARCHITECTURE DIAGRAM



SEQUENCE DIAGRAM



USE CASE DIAGRAM

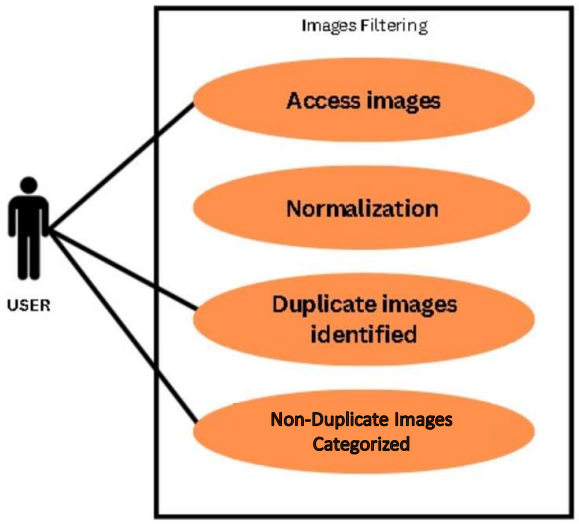


IMAGE PREPROCESSING MODULE

Prepares raw images for further analysis, ensuring consistency across the dataset.

- Resize: Resized & standardization; that is to say all images in the dataset show the same size with respect to the pixels.
- Grayscale Conversion: Original images are converted to gray scale images and give the major part of computational complexity while maintaining almost all the structural details in it.
- Min-Max Normalization: Standardizes pixel intensity from different images so that all images fall in the same scale and improves the performance of features extraction models.

This module ensures images are clean and uniform, providing a stable foundation for deduplication and categorization.

DEDUPLICATION MODULE

Perceptual Hashing (pHash)

- Purpose: Detect near-duplicate images using unique fingerprints.
- How It Works:
  - Convert to grayscale & resize (32x32).
  - Apply DCT to extract key features.
  - Generate binary hash & compare Hamming distance.
- Implementation:
  - Use OpenCV for preprocessing.
  - Compute hash with imagehash.phash().
  - Identify similar images (≥55% similarity).
- Visualization:
  - Display duplicate pairs with similarity scores.
  - Generate a similarity bar chart.

DEDUPLICATION MODULE

Scale-Invariant Feature Transform (SIFT)

- Purpose: Detect and describe local keypoints for image matching.
- How It Works:
  - Extracts keypoints & descriptors.
  - Matches keypoints using Brute Force Matching.
  - Filters matches with Lowe's ratio test.
- Implementation:
  - Convert to grayscale (cv2.cvtColor).
  - Detect keypoints (sift.detectAndCompute()).
  - Match using cv2.BFMatcher().knnMatch().
  - Apply a similarity threshold (e.g., 30%).
- Visualization:
  - Show matched pairs with similarity scores.
  - Generate a similarity bar chart.

CATEGORIZATION MODULE

Image Categorization Methods

Objective: Categorize images into Human, Animal, Vehicle, Landscape, Food, and Flower labels.

- ResNet-50: Deep CNN with residual connections, effective for structured categories but struggles with high variation.
- EfficientNetB0: Lightweight model with compound scaling, balancing accuracy and efficiency for large-scale categorization.
- Swin Transformer: Uses hierarchical self-attention for capturing local & global features, excelling in complex backgrounds.
- ConvNeXtTiny: CNN inspired by transformers, optimizing feature extraction with modern enhancements.
- CLIP: Uses image-text embeddings for semantic categorization, excelling in ambiguous cases.

CATEGORIZATION MODULE

CLIP-Based Image Categorization

- **CLIP (Contrastive Language-Image Pretraining) maps images & text into a shared feature space for similarity-based classification.**
- **Compares each image with category labels (Human, Animal, Vehicle, Landscape, Food, Flower) & assigns the highest similarity score.**
- **Enables accurate classification, even for unseen images.**

Advantages of CLIP

- **No task-specific retraining required, highly adaptable.**
- **Uses text prompts for flexible categorization.**
- **Handles ambiguous images by ranking confidence scores.**
- **Generalizes well across diverse datasets (zero-shot learning).**

ASSUMPTIONS

We're operating under a few assumptions:

- **The system will be used in environments with reliable internet connection for cloud functionalities.**
- **Image Quality & Format Consistency – All input images are assumed to be of sufficient resolution and in supported formats (JPEG, PNG, etc.) to ensure accurate preprocessing, deduplication, and classification.**

WORK BREAKDOWN & RESPONSIBILITES

Alan Aedna	Allwyn Adithyan	Alan Aedna	Allwyn Adithyan
Perceptual Hashing Expert SIFT Feature Matcher	Image Preprocessing Results Display & Reporting	Comparison between Different Categorization models. Result visualization and reporting	Image Categorization Results Display & Reporting

DupFree Organizer

Image Loading & Preprocessing (Adithyan & Allwyn)

- Load & resize images for consistency.
- Apply necessary adjustments before deduplication.

Perceptual Hashing for Near-Duplicate Detection (Alan & Aedna)

- Generate unique digital signatures for images.
- Identify near-duplicates using similarity of signatures.
- Optimize system to prevent redundant comparisons.

Feature Matching with SIFT (Alan & Aedna)

- Detect duplicates despite angle differences.
- Fine-tune SIFT for effective matching.
- Compare SIFT results with perceptual hashing.

DupFree Organizer

Display & Reporting of Results (Adithyan & Allwyn)

•Display duplicate image pairs visually.
•Show similarity scores & highlight detection method.
•Use bar graphs & tables to present similarity metrics.

Categorization & Result Representation

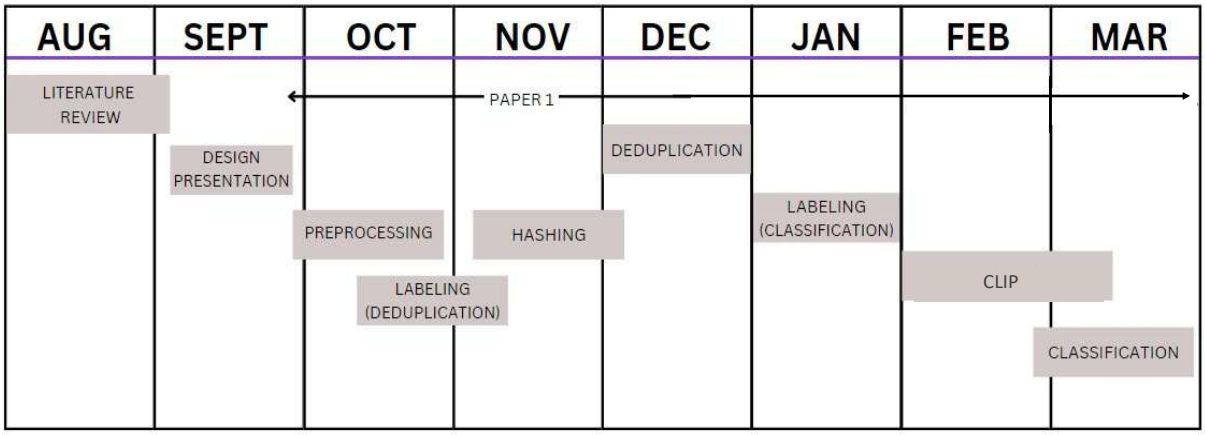
•Alan & Aedna: Implemented ResNet-50 & EfficientNetB0, displayed results via confusion matrices & bar graphs (error rates).
•Adithyan & Allwyn: Implemented Swin-Transformer, ConvNeXtTiny & CLIP, grouped images into category folders.

HARDWARE & SOFTWARE REQUIREMENTS

- Hardware:
  - RAM: At least 16 GB for running browser smoothly.
  - GPU: NVIDIA GTX 1660 or GTX 1080 (or higher) for reliable performance in machine learning
  - Networking: A stable internet connection for cloud access.

- Software:
  - Programming Languages: Python(3.9+)
  - Libraries/Frameworks: TensorFlow and Keras for machine learning, Flask/Django for web frameworks, and React for the frontend.
  - Cloud Services: Google Photos Drive

GANTT CHART



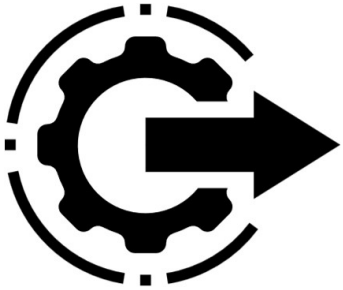
RISKS & CHALLENGES

Like any project, we have some risks to consider:

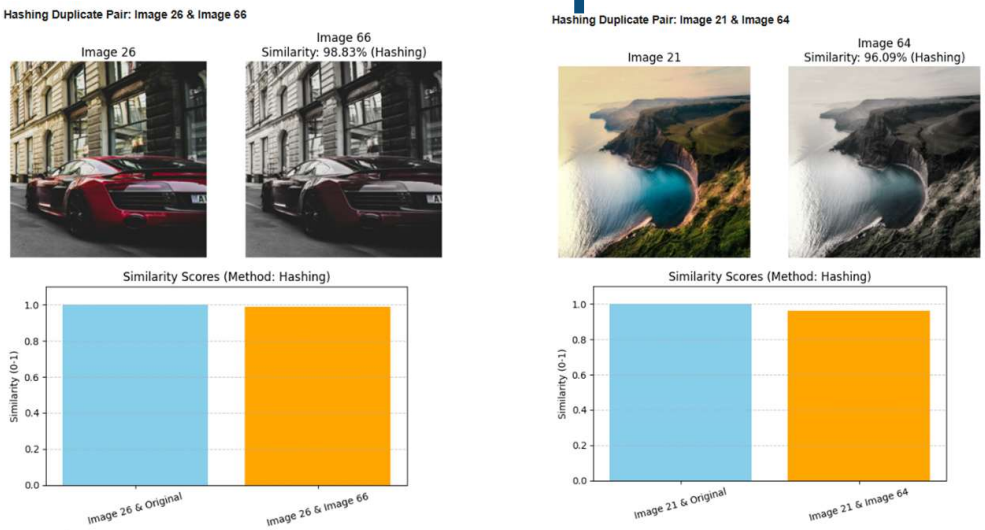
- Technical Risks: A major aspect is ensuring reliable internet connectivity for smooth integration with cloud services, maintaining model accuracy, and supporting fast, high-data transfer rates.
- Operational Risks: Users might need time to adapt to the new system, so training will be crucial.
- Data Security Risks: Since data is diverged among multiple users, ensuring the security of sensitive images during processing and transmission is a top priority.

ACHIEVED OUTPUT

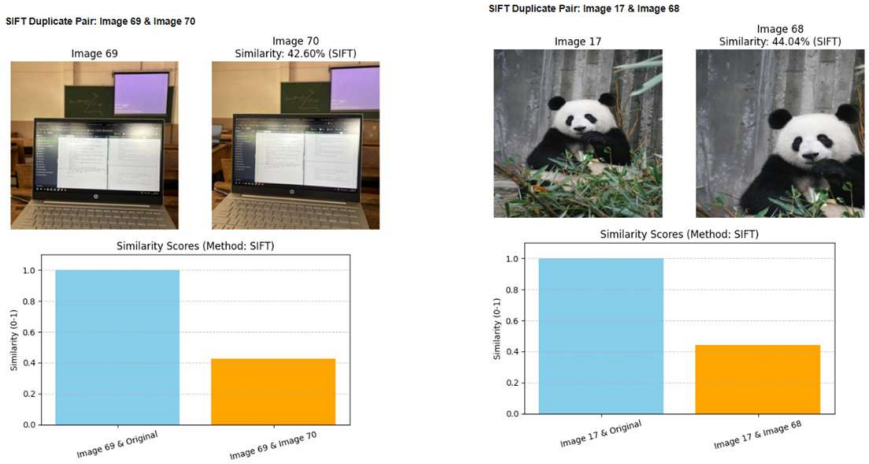
- A robust automated system that efficiently detects and removes duplicate images and then categorizes them.
- A well-organized image database that allows for easy retrieval
- A user-friendly interface that enhances the overall management experience.



OUTPUT- DeDuplication



OUTPUT- DeDuplication



OUTPUT- DeDuplication

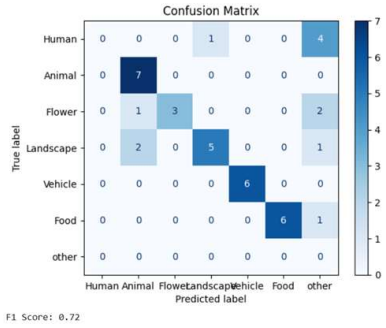
Perceptual Hash Similarity Duplicates

Image Names	Similarity (%)
food 5.jpg ↔ dfood 5.png	98.63%
vehicles 5.jpg ↔ dvehicles 5.png	84.18%
vehicles 1.jpg ↔ dvehicles 1.png	98.83%
landscape 4.jpg ↔ dlandscape 4.png	96.09%
animal 2.jpg ↔ danimal 2.png	96.09%
vehicles 3.jpg ↔ dvehicles 3.png	84.38%
random 5.jpg ↔ drandom 5.png	84.18%
human 2.jpg ↔ dhuman 2.png	96.68%
landscape 3.jpg ↔ dlandscape 3.png	77.54%
flower 3.jpg ↔ dflower 3.png	97.46%
animal 4.jpg ↔ danimal 4.png	86.13%
random 3.jpg ↔ drandom 3.png	99.41%
landscape 6.jpg ↔ dlandscape 6.png	72.66%
IMG_1409.jpg ↔ IMG_1408.jpg	58.40%

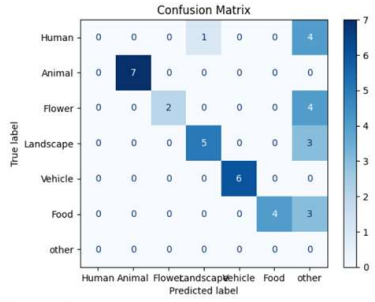
SIFT Similarity Duplicates

Image Names	Similarity (%)
human 5.jpg ↔ dhuman 5.png	66.02%
flower 5.jpg ↔ dflower 5.png	47.62%
animal 1.jpg ↔ danimal 1.png	38.89%
animal 7.jpg ↔ danimal 7.jpg	44.04%
landscape 2.jpg ↔ dlandscape 2.jpg	33.94%
landscape 8.jpg ↔ dlandscape 8.jpg	43.61%
food 2.jpg ↔ dfood 2.jpg	42.23%
food 4.jpg ↔ dfood 4.png	69.83%
food 7.jpg ↔ dfood 7.png	64.55%
random 6.jpg ↔ drandom 6.png	68.93%
IMG_5403.jpg ↔ lap1.jpg	42.60%

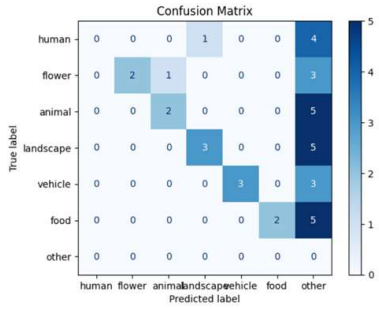
OUTPUT- Categorization



ResNet50

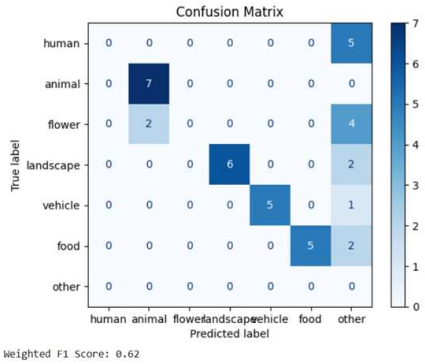


EfficientNetB0

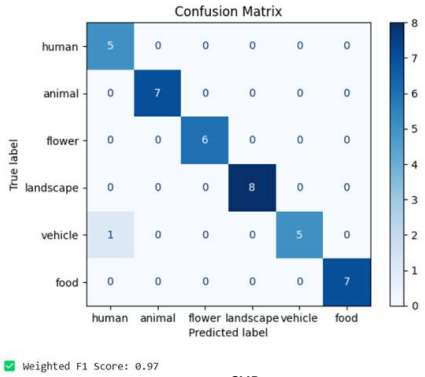


Swin Transformer

OUTPUT- Categorization

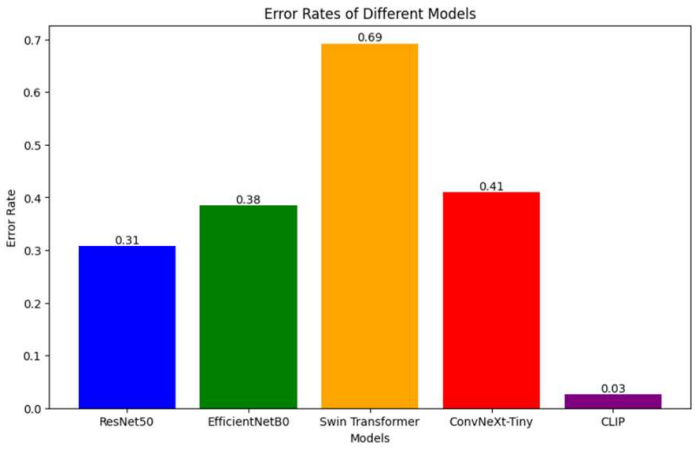


ConvNeXtTiny

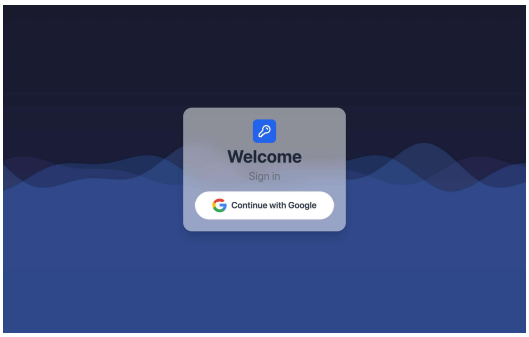


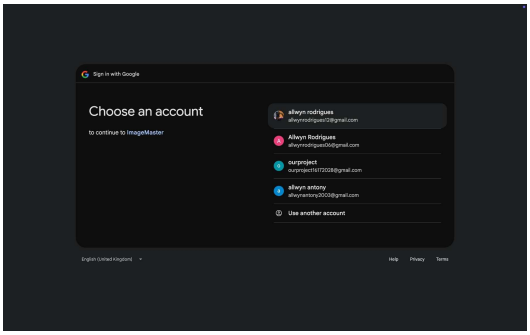
CLIP

OUTPUT- Categorization

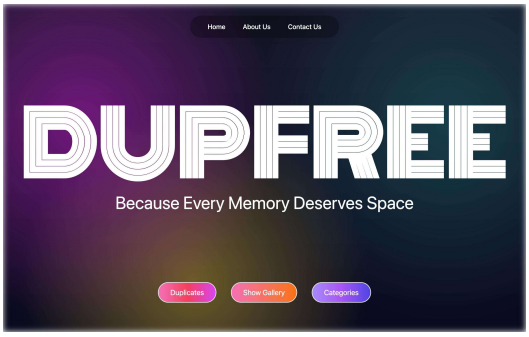


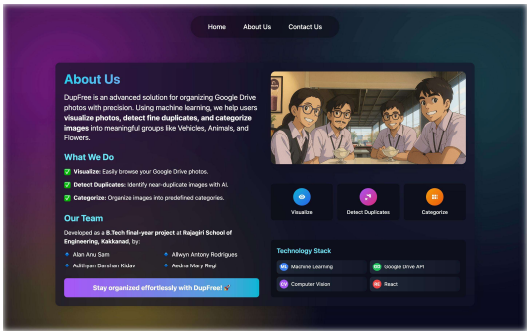
Interface Screenshots



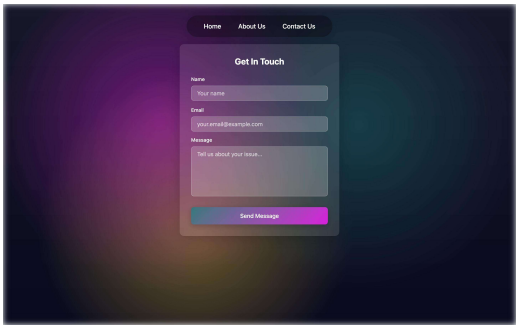


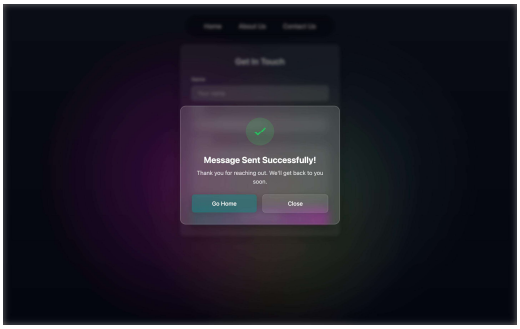
Interface Screenshots



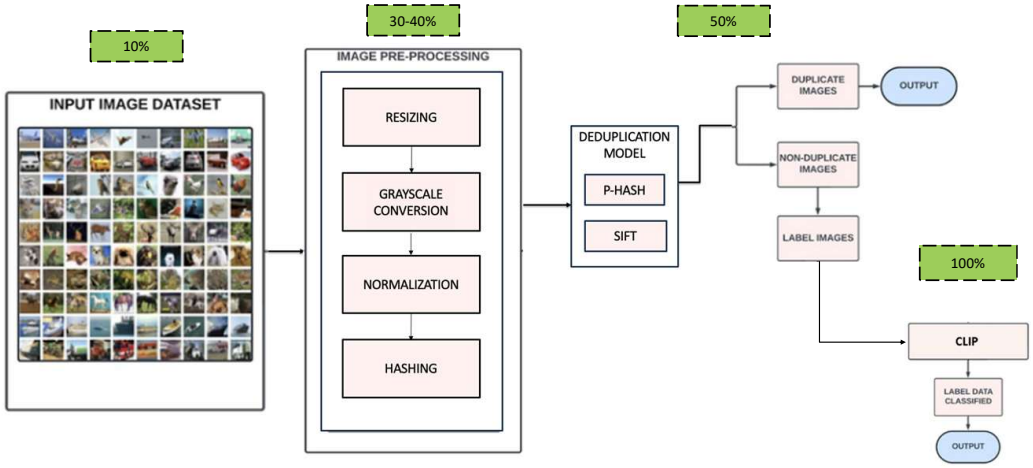


Interface Screenshots





PERCENTAGE COMPLETED



CONCLUSION

The Image Filtering project provides an automated solution for managing large image collections through efficient deduplication and categorization. By utilizing advanced algorithms and cloud integration, we enhance accessibility and streamline workflows. This system aims to improve productivity and reduce errors, making image management easier for users across various sectors.

REFERENCES

1. Kang Hyeon Rhee, "Generation of Novelty Ground Truth Image Using Image Classification and Semantic Segmentation for Copy-Move Forgery Detection", IEEE Access, Vol. 10, pp. 10123-10134, December 2021. Digital Object Identifier 10.1109/ACCESS.2021.3136781.

2. Xuan Li, Liqiong Chang, and Xue Liu, "CE-Dedup: Cost-Effective Convolutional Neural Nets Training based on Image Deduplication", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1123-1130, September 2021.

3. Pegah Dehbozorgi, Oleg Ryabchykov, and Thomas Bocklitz, "A Systematic Investigation of Image Pre-Processing on Image Classification", IEEE Access, Vol. 12, pp. 567-580, April 2024. Digital Object Identifier 10.1109/ACCESS.2024.3395063.

REFERENCES

4. Thang Luong, Luan Dinh, Hung Nguyen, and Linh Tran, "Novel Hardware Implementation of Deduplicating Visually Identical JPEG Image Chunks", IEEE Access, Vol. 12, pp. 1223-1235, May 2024. Digital Object Identifier 10.1109/ACCESS.2024.3401153.

5. Li, T., Zhang, Z., Pei, L. and Gan, Y., 2022. HashFormer: Vision transformer based deep hashing for image retrieval. *IEEE Signal Processing Letters*, *29*, pp.827-831..

6. Chen, Y., Zhang, S., Liu, F., Chang, Z., Ye, M. and Qi, Z., 2022, June. Transhash: Transformer-based hamming hashing for efficient image retrieval. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 127-136).

PUBLICATIONS

Conference name: 14th Computer Science On-line Conference 2025.
Date: April 23-26, 2025.
Status: Accepted.
To be published in : Springer Series: Lecture Notes in Networks and Systems - ISSN 2367-3370.

THANK YOU!

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes (CO)

After the completion of the course the student will be able to:

Course Outcome 1: Identify academic documents from the literature which are related to her/his areas of interest (Cognitive knowledge level: Apply).

Course Outcome 2: Read and apprehend an academic document from the literature which is related to his/her areas of interest (Cognitive knowledge level: Analyze).

Course Outcome 3: Prepare a presentation about an academic document (Cognitive knowledge level: Create).

Course Outcome 4: Give a presentation about an academic document (Cognitive knowledge level: Apply).

Course Outcome 5: Prepare a technical report (Cognitive knowledge level: Create).

Appendix C: CO-PO-PSO Mapping

COURSE OUTCOMES:

SL.NO	DESCRIPTION	Blooms' Taxonomy Level
CO1	Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level:Apply).	Level 3: Apply
CO2	Develop products, processes or technologies for sustainable and socially relevant applications. (Cognitive knowledge level:Apply).	Level 3: Apply
CO3	Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks. (Cognitive knowledge level:Apply).	Level 3: Apply
CO4	Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).	Level 3: Apply
CO5	Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level:Analyze).	Level 4: Analyze
CO6	Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level:Apply).	Level 3: Apply

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
CO 1	2	2	2	1	2	2	2	1	1	1	1	2	3		
CO 2	2	2	2		1	3	3	1	1		1	1		2	
CO 3									3	2	2	1			3
CO 4					2			3	2	2	3	2			3
CO 5	2	3	3	1	2							1	3		
CO 6					2			2	2	3	1	1			3

3/2/1: high/medium/low