KANTIPUR ENGINEERING COLLEGE

(Affiliated to Tribhuvan University)

Dhapakhel, Lalitpur

[Subject Code: CT654]

A MAJOR PROJECT PROPOSAL ON

BIRD SPECIES IDENTIFICATION FROM AUDIO AND IMAGE USING DCNN

Submitted by:

Gaurav Giri	**[Kan077bct034]**
Iza K.C.	**[Kan077bct039]**
Prajwal Khatiwada	**[Kan077bct56]**
Samrat Kumar Adhikari	**[Kan077bct74]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

5 June, 2024

BIRD SPECIES IDENTIFICATION FROM AUDIO AND IMAGE USING DCNN

Submitted by:

Gaurav Giri	**[Kan077bct034]**
Iza K.C.	**[Kan077bct039]**
Prajwal Khatiwada	**[Kan077bct56]**
Samrat Kumar Adhikari	**[Kan077bct74]**

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

Kantipur Engineering College

Dhapakhel, Lalitpur

5 June, 2024

ABSTRACT

SnapTag introduces an approach for efficiently analyzing handwritten documents by leveraging image processing techniques and Named Entity Recognition (NER). The primary objective is to develop a system capable of extracting meaningful information from handwritten content provided by users and subsequently generating relevant tags for improved document organization and categorization.

SnapTag employs image processing methods such as image binarization, thresholding, denoising to enhance the quality of scanned handwritten documents. Through these techniques, the system effectively preprocesses images, mitigating noise and improving the clarity of handwritten text. Then, the methodology involves the integration of Canny edge detection and Hough line transformation, coupled with K-means clustering, to accurately detect document boundaries. Subsequent stages of the process incorporate image segmentation to isolate words and characters, followed by a classification model that identifies each character within the document. The character recognition phase utilizes a trained classification CNN model, to accurately classify individual characters into predefined classes. This step is crucial for deciphering the handwritten content and preparing it for further analysis.In the final stage, NER is employed to extract meaningful tags from the processed document providing valuable metadata that enhances the document's categorization and searchability.

Keywords−Optical Character Recognition, Binarization, Thresholding, Denoising, Boundary Detection, Hough Line Transformation, K-Means Clustering, Convolutional Neural Network, Named Entity Recognition

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Network
EMNIST	Extended Modified National Institute of Standards and Technology
NER	Named Entity Recognition
OCR	Optical Character Recognition
PDF	Portable File Document
ReLU	Rectified Linear Unit

CHAPTER 1
INTRODUCTION

1.1 Background

Globally, the avian kingdom is vast, with over 11,000 species, a testament to nature's complexity and evolutionary prowess. This figure, sourced from the International Ornithological Committee as of April 2023, merely scratches the surface of avian diversity, each species a unique entity with its own ecological role and evolutionary story. (https://www.worldbirdnames.org/new/updates/)

Turning our gaze to Nepal, a country of remarkable biodiversity and varied ecosystems, from the lowland Terai to the towering Himalayas, it is home to more than 887 bird species, as reported by the Himalayan Nature organization. This represents more than 8% of the world's known bird species, a significant figure given Nepal's relatively small geographical footprint. (https://www.himalayannature.org/works/projects/national-red-list-of-nepals-birds/)

Among these, a number are endangered, their existence threatened by habitat loss, climate change, and human activities. The National Red List of Nepal's Birds, a comprehensive assessment of the country's avian biodiversity, identifies several species at risk. Specifically, Nepal harbors 168 nationally threatened bird species, including 68 Critically Endangered, 38 Endangered, and 62 Vulnerable species, as detailed in a publication by the Journal of Threatened Taxa. (https://threatenedtaxa.org/index.php/JoTT/article/view/2855/3

The plight of these endangered species underscores the urgency of conservation efforts. Technologies such as audio recognition and image classification offer innovative tools for identifying and monitoring bird populations. By analyzing the unique sounds and visual characteristics of birds, researchers can enhance our understanding of species distribution, behavior, and threats. Such technologies not only aid in the conservation of endangered species but also contribute to the broader field of biodiversity research.

1.2 Problem Statement

Given the complexity and significance of avian biodiversity, particularly in regions rich in species diversity such as Nepal, the challenge of accurately identifying and classifying bird species, especially those that are endangered, emerges as a critical concern. The necessity for precise identification and classification is not merely academic but is deeply intertwined with conservation efforts aimed at preserving the delicate balance of ecosystems. In Nepal, where the avian population includes a range of endangered species, the task of monitoring and protecting these birds is compounded by the limitations of traditional observation methods.

The problem, therefore, lies in developing a method that can overcome the constraints of human observation and the vast geographical and ecological diversity of Nepal. This method must be capable of identifying bird species through the sounds they make, a task that requires distinguishing between the myriad of calls and songs in the natural environment. Additionally, it must utilize visual identification through image classification to account for instances where audio data may be insufficient or inconclusive. The integration of audio recognition and image classification technologies presents a promising solution to this problem, offering a way to automate the identification process, enhance accuracy, and significantly improve the efficiency of monitoring endangered bird species.

However, the implementation of such technologies raises several questions: How can audio recognition and image classification systems be effectively trained to recognize the specific calls and visual markers of Nepal's endangered bird species? What are the challenges in collecting and curating the necessary datasets for this training, given the elusive nature of many of these species and the complex acoustics of their natural habitats? And, importantly, how can these technologies be integrated into conservation strategies to not only identify but also protect these species from the myriad threats they face?

Addressing these questions requires a multidisciplinary approach that combines expertise in ornithology, conservation biology, machine learning, and environmental science.

The ultimate goal is to develop a robust system that can contribute to the conservation of Nepal's endangered birds by providing reliable data on species distribution, population trends, and habitat use, thereby informing targeted conservation actions and policies.

1.3 Objectives

i To develop and implement an integrated technological solution that utilizes advanced audio recognition and image classification techniques for the accurate identification and monitoring of bird species in Nepal, with an emphasis on endangered species.

1.4 Application Scope

1. This software can be utilized by researchers and academics to organize their handwritten research notes,allowing for efficient retrieval of information during the writing process. The OCR functionality enables quick keyword searching within the notes, helping researchers locate relevant information, references, and supporting evidence for their papers.
2. Students can leverage the software to manage and organize their handwritten study notes, lecture summaries, and practice questions. The ability to search for keywords within the notes makes it easier for students to locate specific topics or concepts while preparing for exams.
3. Language learners can utilize the software to organize their handwritten vocabulary lists, grammar rules, and language exercises. This enables them to access them quickly, helping in retention by facilitating vocabulary acquisition and language practice.

1.5 Features

1. **Handwritten Note Digitization:**

   The platform allows users to easily convert their handwritten notes into digital format through scanning or capturing images, ensuring easy access and preserva-

tion.

2. **Keyword-Based Searching:**

   Leveraging OCR technology, the platform enables users to search for specific keywords within their handwritten notes, providing quick and accurate retrieval of relevant information.

3. **Cross Platform:**

   The platform can be accessed from various devices, such as computers, tablets, and smartphones, allowing users to retrieve their notes at any time.

4. **Cost-Free and User-Friendly:**

   The platform is free to use and offers a user-friendly interface, ensuring that students and researchers can efficiently organize and search their handwritten notes without any financial barriers.

5. **Efficient Content Indexing:**

   The platform automatically indexes the handwritten text within the notes, generating a searchable database of keywords. The indexing feature significantly improves the speed and accuracy of searching for specific information.

1.6 Feasibility Study

Before implementation of project design, the feasibility analysis of the project must be done to move any further. The feasibility analysis of the project gives an idea on how the project will perform and its impact in the real world scenario. So, it is of utmost importance.

1.6.1 Economic Feasibility

Our system is economically achievable as a result of the development of several tools, libraries, and frameworks. Since all the software required to construct it is free and readily available online, this project is incredibly cost-effective. Only time and effort are needed to create a worthwhile, genuinely passive system. The project doesn't come at a substantial cost. From an economic standpoint, the project appears successful in this sense.

1.6.2 Technical Feasibility

The software needed to implement a project can be downloaded from a wide variety of online resources. Technically speaking, the project is feasible as the necessary software is easily available. We were able to learn the information we required for the project through a variety of online sources, including classes. All the libraries and data are accessible online for free because this project does not require any licensing costs. It is technically possible if one has the necessary information and resources.

1.6.3 Operational Feasibility

Our project intends to improve hand-written notes taking experiences with a primary focus on searching keywords in notes/documents to make it more accessible. The project is intended for a sizable group of learners and students who desire to maintain their handwritten notes. Consequently, our project is operationally feasible.

1.6.4 Schedule Feasibility

The workload of the project is divided amongst the project members. The scheduling is done according to an incremental model where different modules are planned to be assigned to the group members.So, the project fulfills the schedule feasibility requirements.

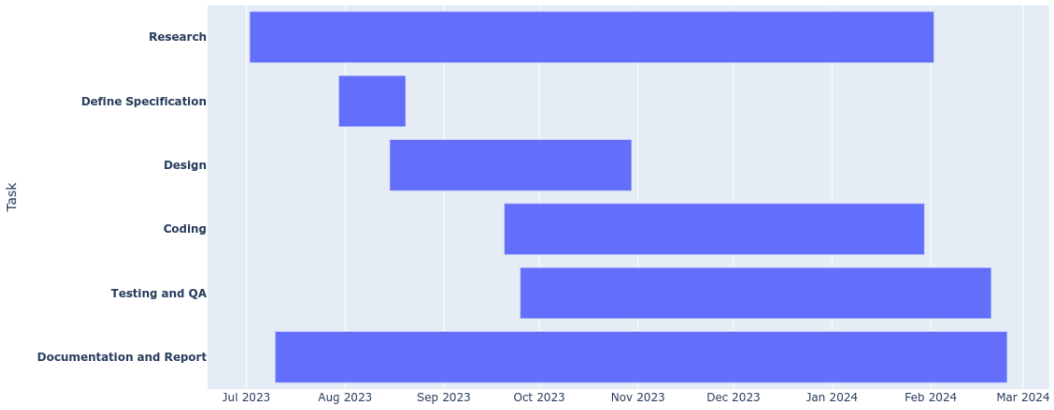

Figure 1.1: Gantt Chart

1.7 System Requirements

1.7.1 Development Requirements

Table 1.1: Development Requirements

Software Requirements	**Hardware Requirements**
Programming Language: Python,Dart, Java	Camera: >= 12 Megapixels
Design Tools: Figma	RAM: >= 8 GB
OpenCV Library, spaCy	CPU: i5 10th (Recommended)
Frameworks: Flutter, Tensorflow	GPU: P100
Testing and Debugging Tools	Storage: >= 10 GB

1.7.2 Deployment Requirements

Table 1.2: Deployment Requirements

Software Requirements	**Hardware Requirements**
Android: >= 10	Camera: >= 12 Megapixels
Read/Write FileSystem	RAM: > 4 GB
Internet Accessibility	Storage: >= 5 GB

CHAPTER 2
LITERATURE REVIEW

Regardless of whether they are documented or not, every project has helped to shape the world as it is today. Other researchers can benefit from documented projects by learning specifics about problems and how to solve them. Additionally, they boostproject efficiency by removing the need to start the project from the beginning and specifying the starting point.

2.1 Related Works

BirdNET is a cutting-edge research platform developed through collaboration between the K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology and the Chair of Media Informatics at Chemnitz University of Technology. Its primary aim is to detect and classify bird sounds using machine learning technologies, serving both experts and citizen scientists in their efforts to monitor and protect bird populations.

BirdNET can identify around 3,000 of the world's most common bird species, with plans to expand this number. Features such as a live submissions map and a Twitter bot are included to engage the community and share real-time data. The project is supported by donations and collaborations, offering opportunities for researchers and developers to contribute to its growth. BirdNET serves as an invaluable tool for bird enthusiasts, conservationists, and biologists alike, providing innovative solutions for large-scale acoustic monitoring and contributing to the conservation and understanding of avian biodiversity.

The BirdCLEF 2023 competition on Kaggle is a significant data science challenge that falls under the broader LifeCLEF initiative, aimed at pushing the boundaries of species identification and biodiversity monitoring through technological innovation. This particular competition focuses on the development of machine learning models that can identify bird species based on audio recordings. It presents a complex and realistic

challenge due to the diversity of the audio recordings, which are collected from various environments and feature a wide range of bird species.

2.2 Related Research

The research paper 'Audio Classifier for Automatic Identification of Endangered Bird Species of Nepal' focuses on developing an audio classifier to identify endangered bird species in Nepal using deep learning techniques. The dataset, collected from xeno-canto.org, comprises 2215 audio recordings of 41 bird species, 38 of which are endangered. This dataset was expanded to 6733 recordings through 10-second audio splitting and Gaussian noise augmentation, with 5407 recordings used for training, 639 for validation, and 687 for testing. The methodology involved handling imbalanced class distribution through data augmentation, employing Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction, and developing a custom Convolutional Neural Network (CNN) model and an EfficientNet model. The hyperparameters of these models were optimized using a genetic algorithm. The Mel spectrograms were created using Short-Time Fourier Transform, converting amplitudes to decibel scale, and applying Mel filter banks to the spectrograms. Similarly, MFCCs were derived by framing the audio signals, applying Discrete Fourier Transform, logarithmic scaling, Mel scaling, and Discrete Cosine Transform. The EfficientNet architecture utilized compound scaling for network depth, width, and resolution. The findings indicated that the proposed approach achieved satisfactory results in classifying the bird species. However, limitations include the relatively small dataset size and the need for further enhancement in model robustness and accuracy. Future enhancements could involve expanding the dataset, exploring additional feature extraction techniques, and incorporating more advanced deep learning models to improve classification performance. This research contributes significantly to the conservation efforts by providing a reliable method for automatic bird species identification, aiding in monitoring and protecting endangered species.[**?**]

The paper 'Audio Bird Classification with Inception-v4 extended with Time and Time-

Frequency Attention Mechanisms' presents an innovative adaptation of the Inception-v4 deep convolutional network for bioacoustic classification, focusing specifically on bird sound recognition. The datasets employed include various bird sounds, prominently from the BirdClef2017 challenge, consisting of 1500 bird species recordings. The methodology revolves around treating bird sound classification as an image classification problem through transfer learning. The Inception-v4 model, initially pre-trained on ImageNet, was adapted to process time-frequency representations of bird sounds by converting these sounds into RGB images using three log-spectrograms generated via fast Fourier transform at different scales (128, 512, 2048 bins). Data augmentation techniques, common in computer vision, were applied to these spectrograms to enhance the robustness of the model. The findings demonstrate that the model, termed 'Soundception', integrates time and time-frequency attention mechanisms effectively, significantly improving classification accuracy. The results highlight Soundception's outstanding performance, achieving a mean average precision (MAP) of 0.714 in classifying 1500 bird species, 0.616 MAP for background species, and 0.288 MAP for soundscapes with time-codes, making it the top model in the BirdClef2017 challenge across multiple tasks. However, limitations include the incomplete convergence of the model due to computational constraints and the extensive GPU resources required for training, which restricted the full potential realization within the challenge's timeframe. The paper concludes with a discussion on future improvements, such as exploring different scalable optimizations and incorporating stacked GRU layers for better audio-to-image representation learning, underscoring the potential of transfer learning from advanced image classification models to acoustic domains

The research paper 'Bird Species Identification using Deep Learning' presents a methodology using Deep Convolutional Neural Networks (DCNNs) to classify bird species, leveraging the Caltech-UCSD Birds 200 (CUB-200-2011) dataset, which contains 11,788 annotated images of 200 bird species. The methodology involves converting images to grey scale to reduce computational complexity, followed by the application of DCNNs using TensorFlow to extract hierarchical features from images such as edges, textures, and complex patterns. The neural network architecture includes convolutional

layers for feature extraction, pooling layers for dimensionality reduction, activation layers for non-linearity, and fully connected layers for classification. Key findings show the DCNN model achieved a testing accuracy of 80% and training accuracy of 93%, with a validation accuracy of around 75%, indicating robust performance across different data splits. The study highlights the effectiveness of combining multiple features (head, body, color, beak) over single-feature classification, with generated autographs and score sheets facilitating identification. The system's usability is enhanced through a web interface for image uploads, and future directions propose mobile app development and cloud integration to improve accessibility and scalability. However, the research also identifies limitations such as the dependency on the quality and diversity of the dataset, potential overfitting due to the complexity of the model, and the computational resources required for training deep learning models. Addressing these limitations could further enhance the model's accuracy and applicability in real-world scenarios.

The research paper 'Bird species classification from an Image using VGG-16 Network' utilizes machine learning and deep learning techniques has shown significant advancements in both methodology and accuracy. The primary focus has been on the creation and utilization of diverse datasets, with a notable example being the dataset of 1600 images across 27 bird species, as well as the Caltech-UCSD Birds-200-2011 Dataset. Methodologically, the use of Convolutional Neural Networks (CNNs), particularly the VGG-16 model, has been prevalent due to its effectiveness in feature extraction from images. This approach has been complemented by transfer learning techniques, which have notably improved classification accuracies, as seen in the use of pre-trained networks like AlexNet and VGG-16, achieving accuracies up to 85.4% and 92.13% respectively. The application of machine learning algorithms such as SVM with a linear kernel and KNN has also been explored, with SVM achieving an accuracy of 89% after parameter optimization. Additionally, the integration of computer vision methods to extract Histogram Oriented Gradients and RGB histogram values has further enhanced classification performance. Despite these advancements, the studies have encountered limitations, including the challenge of accurately classifying bird species from vari-

ous angles and positions, and the need for extensive preprocessing to remove noise from datasets. Overall, the body of work demonstrates a trend towards higher accuracy in bird species classification through the innovative use of deep learning techniques and sophisticated feature extraction methods, though challenges remain in dealing with the variability of natural images and the computational demands of processing large datasets.

This research paper presents significant advancements in the field of 'Automatic bird species identification through the integration of audio signal processing and neural networks'. The study, conducted by Chandu B, Akash Munikoti, Karthik S Murthy, Ganesh Murthy V, and Chaitra Nagaraj from the BNM Institute of Technology, outlines a robust methodology for identifying bird species from audio recordings, leveraging a combination of meticulously curated datasets and sophisticated machine learning techniques. The dataset, a critical component of the study, was manually compiled from both local recordings and online resources such as xeno-canto.com, featuring audio clips from species like cuckoo, sparrow, crow, and laughing dove, as well as ambient noise and human voices to simulate real-world conditions. Pre-processing techniques including pre-emphasis, framing, silence removal, and reconstruction were applied to the audio clips to enhance the relevant frequency components and eliminate unnecessary noise, ensuring the purity of the dataset. Spectrograms of these processed clips were generated and used as input for training a convolutional neural network (CNN), specifically AlexNet, chosen for its high accuracy in image classification tasks. Through transfer learning, AlexNet was adapted to recognize bird species from the spectrograms, achieving a classification accuracy of 97% in controlled environments. However, recognizing the variability of real-world conditions, the researchers retrained the model with datasets containing ambient noise, achieving a real-time classification accuracy of 91%. Despite these promising results, the study acknowledges limitations such as the relatively small size of the dataset and the need for further tuning of performance parameters to improve robustness. The potential for future applications is vast, including the development of mobile applications and hardware implementations for ecological monitoring, highlighting both the scientific and practical significance of automated bird

species identification systems.

The paper 'An Ensemble of Convolutional Neural Networks for Audio Classification' delves into a comprehensive study on CNN classification using different architectures, data augmentation techniques, and audio signal representations, aimed at enhancing audio classification tasks across various datasets. The study employs three datasets: BIRDZ, CAT, and ESC-50, each offering unique challenges in audio classification. The methodology involves training five convolutional neural networks (CNNs) with four audio representations combined with six different data augmentation methods, resulting in thirty-five subtypes of ensembles. The audio representations include techniques such as the Discrete Gabor Transform (DGT), Waveform Similarity OverLap Add (WSOLA), and Phase Vocoder. The data augmentation methods encompass procedures like short spectrogram augmentation, random time shift, and frequency masking. The CNN architectures are pre-trained models fine-tuned with these augmented datasets to boost classification accuracy. The findings reveal that the ensemble method outperforms standalone networks, achieving 97% accuracy on the BIRDZ dataset, 90.51% on the CAT dataset, and 88.65% on the ESC-50 dataset. The study also highlights that the best-performing CNNs are VGG16 and VGG19, with DGT as the most effective signal representation. However, the study acknowledges limitations, such as the computational cost of training ensembles and the variability in performance across different augmentation techniques. Notably, no single augmentation protocol consistently outperforms others across all datasets. The results underscore the potential of combining different CNNs and augmentation policies to enhance performance, although this approach demands significant computational resources. Additionally, the study provides the MATLAB source code for reproducibility, contributing to the research community's efforts in advancing audio classification technologies.

The literature on 'Bird species identification using deep learning techniques' provides a comprehensive exploration of the datasets, methodologies, findings, results, and limitations associated with this field. The study utilized the largest publicly available dataset, consisting of over 33,000 recordings of 999 different bird species, highlighting the sub-

stantial challenge due to the extensive diversity and volume of data. The methodology employed a convolutional neural network (CNN) architecture with five convolutional layers and one dense layer, utilizing rectified activation functions and max-pooling layers for feature extraction. Preprocessing involved separating sound files into signal and noise parts, computing their spectrograms, and dividing these into equally sized chunks to serve as training and testing samples. Data augmentation techniques, such as time shifts and pitch shifts, were crucial in enhancing the model's generalization capability. The results were notable, with the model achieving a mean average precision (MAP) score of 0.69 and an accuracy score of 0.58, marking the highest recorded performance in the BirdCLEF 2016 Recognition Challenge. Despite these successes, the study acknowledged several limitations, including the difficulty of distinguishing between multiple birds singing simultaneously and the challenges posed by variable recording lengths and background noise. Additionally, attempts to improve the model using methods like bi-directional LSTM recurrent neural networks and non-square filters did not yield better performance. Future improvements could involve using an ensemble of neural networks, incorporating meta-data such as date, time, and location, and addressing the disproportionate influence of longer files on the model's predictions.

The literature review focused on the 'analysis of bird call datasets sourced from Xeno-Canto', comprising 72,172 samples from 264 bird species in 16-bit wav format with a 16 kHz sampling rate. The methodology involved preprocessing the audio data to filter out low-frequency noise and normalize signal amplitude, followed by generating Mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) as inputs for deep learning models. The Mel-spectrograms were produced using discrete Fourier transform (DFT), and the MFCCs were derived by applying discrete cosine transform (DCT) to the Mel-spectrogram. The study employed various metrics to evaluate the performance of these methods, including ROC analysis to visualize model effectiveness. Findings indicated that the proposed models showed significant promise in identifying bird species from their calls, with improvements in classification accuracy compared to previous approaches. However, limitations were noted, including potential biases in the dataset due to uneven sample distribution across species and the challenge of back-

ground noise affecting signal quality. Future work suggested enhancing noise reduction techniques and exploring more sophisticated neural network architectures to further improve model robustness and accuracy.

The study conducted an in-depth analysis of 'bird species recognition through acoustic monitoring', utilizing a robust dataset of bird sound samples, meticulously annotated and validated for accuracy. The dataset, referred to as SD, comprises multispecies bird sound recordings, each labeled with species name and sample ID, along with corresponding metadata, providing a comprehensive foundation for model training and evaluation. Methodologically, the research employed a spectrogram-based feature extraction approach, leveraging Short-Time Fourier Transform (STFT) to capture the intricate temporal and spectral characteristics of bird sounds. This was followed by the application of a Multilayer Perceptron (MLP) classifier to distinguish between different bird species. The findings reveal that the proposed model achieved high recognition accuracy, with some species being identified with perfect precision, recall, and accuracy (100%), though the performance varied across species, with a few showing lower recognition rates (86.9%) and precision/recall values ranging between 50-75%. The results demonstrated an overall classification accuracy of 96%, with cross-validation accuracy standing at 81.4%, highlighting the model's robustness yet indicating room for improvement in generalizability across diverse datasets. Despite the promising results, the study acknowledges several limitations, including the variability in recognition accuracy among different species and the potential influence of environmental noise on model performance. Future work is suggested to explore feature and model fusion techniques, integrate the model with cloud-based systems for real-time recognition, and expand the dataset to include a broader range of bird species to enhance the model's applicability and accuracy in practical scenarios.

CHAPTER 3
METHODOLOGY

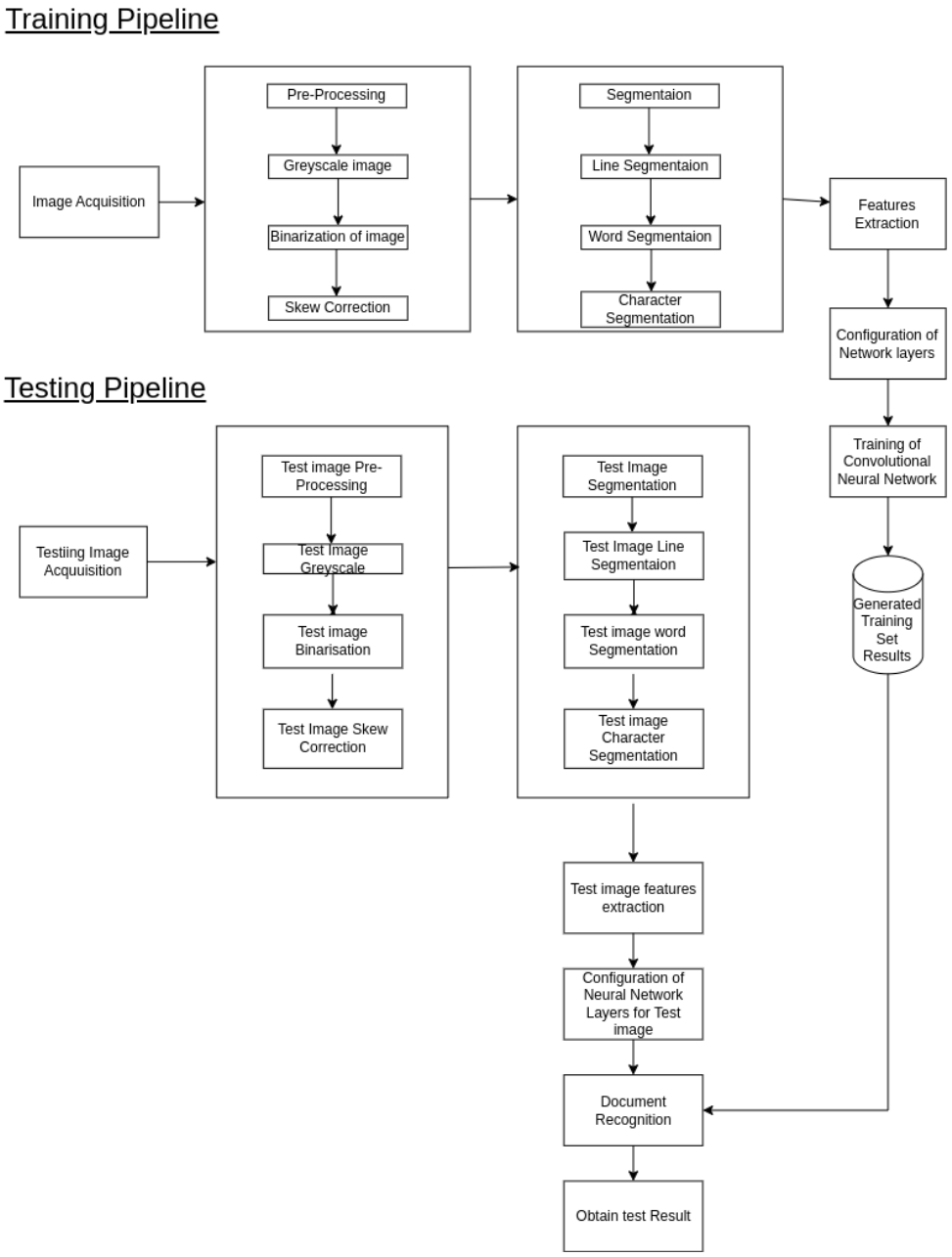

Figure 3.1: Block diagram for the working mechanism of the system

3.1 Working Mechanism

The methodology diagram is shown above fig no. 3.1 which consists of several stages such as image acquisition i.e.testing image, pre-processing (grayscale conversion, binarization, document detection), segmentation (lines, words, and characters), feature extraction, and classification using CNN layers into[**?**] character classes.

3.1.1 Image Acquisition

The first step is to obtain a handwritten script through scanning physical documentation or images using a camera.

3.1.2 Pre-processing

It is essential to enhance the quality and prepare the image for further analysis. This step includes Grayscale Conversion, Binarization, Noise Removal and skew correction. Grayscale conversion is used to reduce complexity of handling multiple color channels (24-bit to 8-bit). Binarization is used to convert grayscale images into black and white to make it easier for segmentation.

Noise removal is used to remove unwanted variations in pixel intensity that degrade image quality. The proposed method uses a median filter to efficiently remove salt and pepper noise. This filter sorts adjacent pixel values and replaces the current pixel value with the median, reducing the noise and improving image.

3.1.3 Document Detection

The document detection is used for auto-cropping the image and correcting the perspective angle of the page of interest. For that quadrilateral points are obtained through the merged process of canny edge detection and hough transformation then KMeans Clustering that clusters the intersection points into 4 groups which are the quadrilaterals. Then through the quadrilaterals, perspective transformation is performed to get a good perspective of the page[**?**].

3.1.4 Segmentation

In the segmentation process, we first tackle Line Segmentation, where the input image is analyzed to identify lines of text. By finding dark centroids in the image, we can mark the positions of each line. We employ a projection profile-based algorithm for this task and address the issue of skewed text by applying skew correction techniques.
Moving on to Word Segmentation, our approach involves morphological dilation, which allows us to connect individual characters in a line, ultimately helping us to extract separate words. Using labeling techniques, we efficiently isolate and extract words from the connected components.
Character Segmentation is a crucial step, particularly for cursive handwriting. To address the challenges of cursive characters, we implement skeletonization and vertical projection techniques to detect potential segmentation points. A distance-based approach is utilized to avoid over-segmentation, ensuring that characters like 'm', 'n', 'u', 'v', 'w', etc., are correctly identified. For untouched characters, a vertical projection is employed to identify spaces between characters, facilitating their separation.
Through these segmentation processes, we break down the handwriting image into lines, words, and individual characters, which is essential for accurate and efficient handwriting recognition.

3.1.5 Features Extraction

Feature extraction involves identifying important image features and saving them for further processing. It bridges the gap between pictorial and data representation in image processing. The proposed method employs a Convolutional Neural Network (CNN) for effective feature extraction, making the process more efficient and accurate.

Training of Convolutional Neural Network

The configured layers of Convolution Neural Network are shown below. To train the neural network, the dataset is divided into two sets: the training data and the testing data. The complex handwritten character dataset has data format of 28x28 ".png" format

picture. The data set has a total of 62 categories of 0-9, a-z and A-Z, corresponding to the files ”0” to ”61” in the order of ”label.txt”. The data set is divided into two parts, the unprocessed original data image is stored in the ”0” to ”61” in the ”V0.3/data” folder, and the binarized data image Stored in ”0” to ”61” in the ”V0.3/data-bin” folder[**?**]. The EMNIST also has 62 character classes divided into digit classes and letter classes

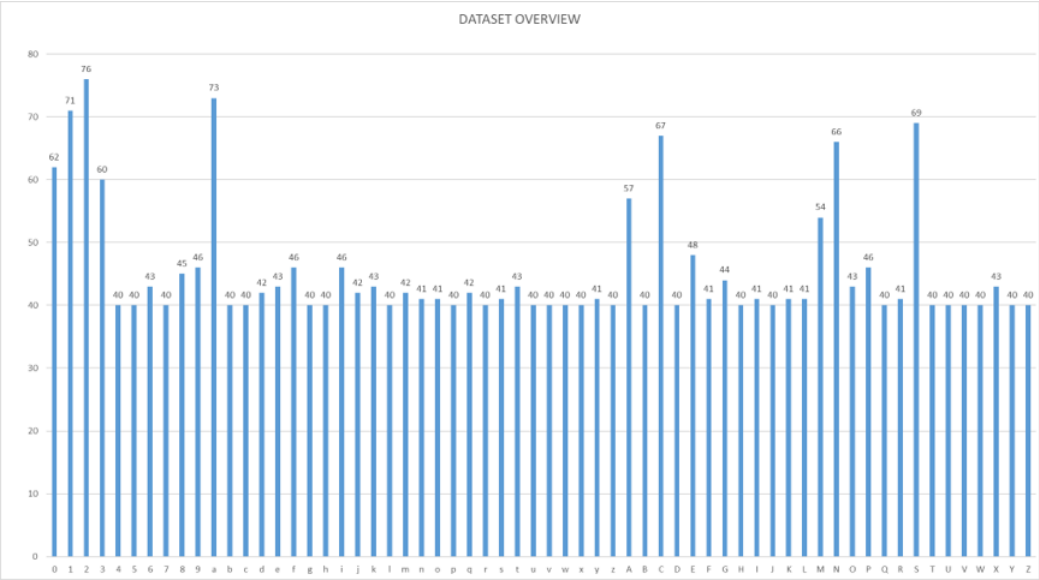

Figure 3.2: Dataset overview for the A good CHoiCe

containing a total 814,255 samples[**?**].

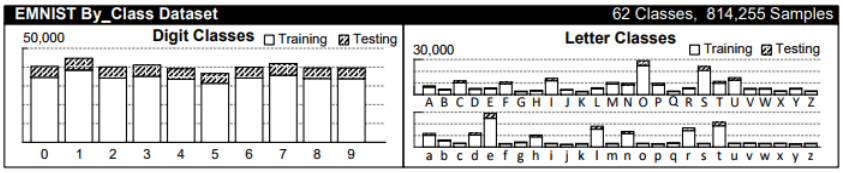

Figure 3.3: Visual representation of E MNIST dataset

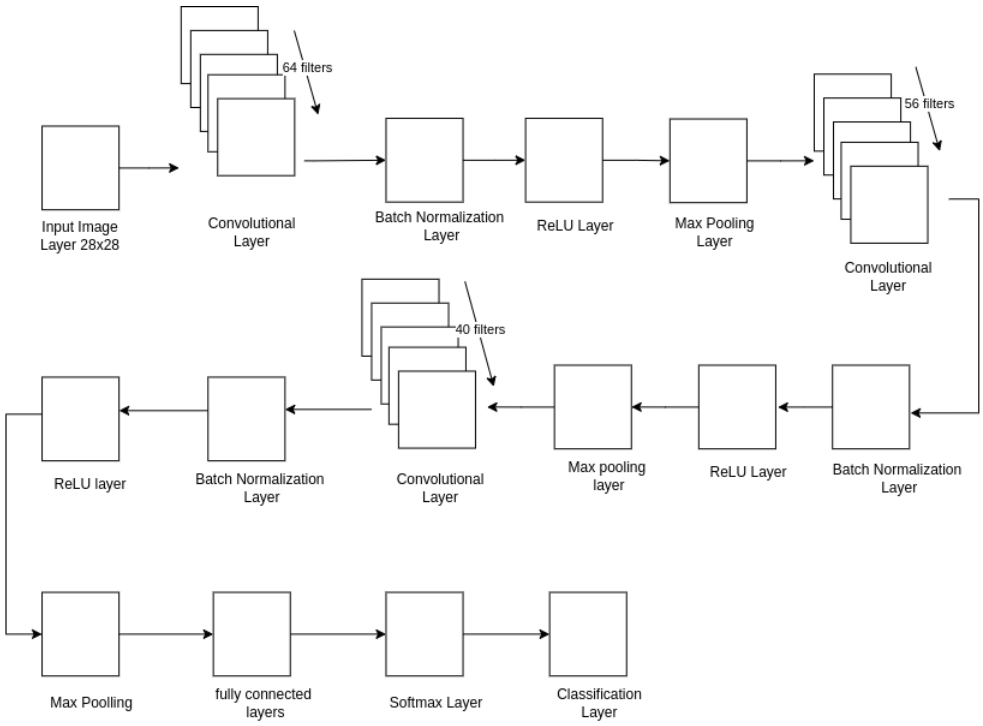

Figure 3.4: Configured layers of Convolution Neural Network

Document Recognition

The neural network used for image recognition involves four primary operations: Convolution, Non-Linearity (ReLU), Pooling, and Fully Connected Layer.

1. Convolution: The convolution operation extracts features from the input image M using a filter matrix N. The resulting feature map $f_{\text{con}}(k, l)$ can be calculated as follows:

$$f_{\text{con}} = \text{convolution}(m, n) \tag{3.1}$$

$$(m \bigotimes n)(k, l) = \sum_{m} \sum_{n} M(m, n) N(k - m, l - n) \tag{3.2}$$

2. Non-Linearity (ReLU): The ReLU function is applied element-wise to the feature map to introduce non-linearity and obtain the rectified feature map frec(xk):

$$f_{\text{Rec}} = \text{ReLU}(x, l) = \max(0, x_k) \tag{3.3}$$

3. Pooling: Pooling reduces feature map dimensionality while preserving important information. The Max pooling technique selects the maximum value from a defined spatial matrix (e.g. 2x2):

$$f_{\text{pool}}(k, l) = \max(x_k) \tag{3.4}$$

4. Fully Connected Layer: In this layer, all neurons are connected to each other, and the output is used for classification.

   Finally, the Softmax Layer outputs probabilities for each class of the handwritten characters. The Softmax function calculates the probability for an input element xk. belonging to label i as follows:

$$p(O_i) = \text{softmax}(x_k) = \frac{\exp(x_k)}{\sum_{g=1}^{y} \exp(x_k)} \tag{3.5}$$

   The Classification Layer then identifies the recognized character based on the highest probability assigned by the Softmax Layer.

3.1.6 Named Entity Extraction using spaCy

NER is a process in which anything that is denoted by a proper name or tag,for example, a location, an organization, or a person is identified as an entity.Named entities include things like geographical location, date, time, or money, and customization of the NER model for user-defined named entities is also possible[**?**].

spaCy is a free, open source library that allows advanced Natural Language Processing in Python. This library will be used to extract the unique word which will be used as tags for the image while searching it. Unique words are generated through named entity recognition for which ”en_core_web_sm” is used, which is a linguistic model.

3.2 Algorithm Used

- **Line Segmentation:** The Line Segmentation algorithm starts with image padding to avoid segmentation errors. It computes the horizontal projection to identify text

regions and detects dark lanes using a threshold in the binary image. The algorithm labels and counts dark regions to determine total lines of text. Centroids are calculated for precise positioning and stored for further processing. x and y coordinates of centroids are extracted. Text regions are cropped based on consecutive y-coordinates for accurate segmentation, enabling proper division of each handwritten script line for improved recognition and analysis.

- **Word Segmentation:** The Word Segmentation algorithm aims to extract individual words from previously segmented lines of handwritten text. It involves three steps: first, dilating the line image to expand the pixels and enhance visual distinction between connected components; second, labeling the connected components to identify distinct words; and finally, separating the labeled components to obtain individual images of each word. These separated images represent the segmented words of the handwritten text.
- **Cursive Character Segmentation:** The Cursive Character Segmentation algorithm separates connected characters in a handwritten text image by skeletonizing the image and computing its vertical projection. It marks segmentation points using the ”Segmentation Points” (SP) array, starting from the first character. For each column and row, it checks if certain conditions are met to identify connected characters. If these conditions are fulfilled, the algorithm stores the column in the SP array based on a threshold distance. After marking the segmentation points, it draws red lines on the word image for visualization and stores 0 in the rows of SP to disconnect the characters. This process paves the way for further character segmentation and recognition.
- **Untouched Character Segmentation:** The Untouched Character Segmentation algorithm effectively segments individual characters in a handwritten text image that have not been separated previously. It begins by padding the image for better segmentation and computes the Vertical Projection (VP) to identify character zones. Applying a threshold to the binary image detects dark lanes between letters. Each dark region is labeled, and the total character count is determined. Centroids are calculated for precise character positioning, and their x and y coordinates are extracted. The algorithm then crops text regions assuming a line between consecutive x-coordinates, ensuring accurate character segmenta-

tion. Cropped regions are stored separately, facilitating proper segmentation for s subsequent recognition and analysis.

3.3 System Diagrams

The usecase diagram and activity diagram for SNAPTAG are given below:

Figure 3.5: UseCase Diagram

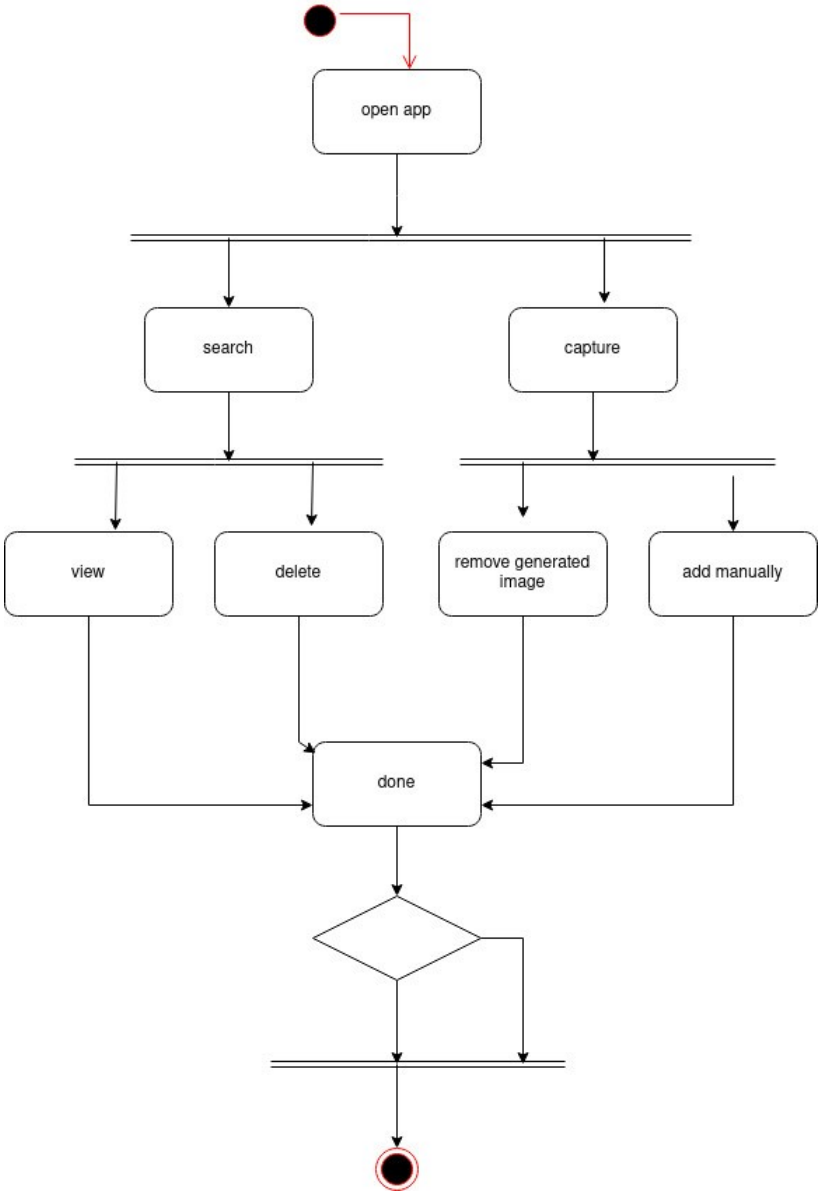

Figure 3.6: Activity Diagram

3.4 Software Development Model

This project is developed using an incremental methodology since it offers a functioning prototype at an early stage of development. The requirements and scope of the project can be altered as necessary by studying the prototype. The rationale for the preference for this software development strategy is the flexibility offered by adopting the incremental technique. In this paradigm, the project goes through several releases or iterations prior to its official release.

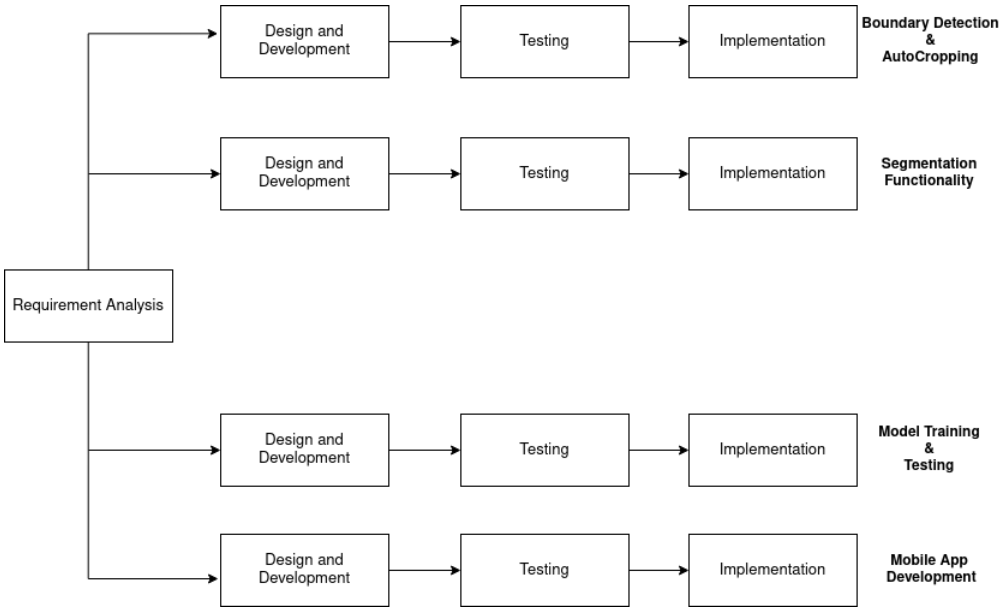

Figure 3.7: Incremental Model for development of SnapTag

CHAPTER 4
RESULT AND DISCUSSION

4.1 Result

We have completed the design and development of the system along with the required output of the project. The project currently allows user to register and log-in into the system. User can post queries, answer, vote, follow other users and search queries. Trending section shows most relevant post using half life decay algorithm.

4.2 Work Done

1. Preprocessing: Common processing steps like Thresholding, De-noising, Resizing were done. In threshold, we used Otsu's Binarization due to its automatic optimal threshold value determination and avoiding having to choose a value. In order to do so, the cv2.threshold() function was used. Processing high resolution images are quite resource intensive. So, images were scaled down for faster processing using cv2.resize() function. And we performed de-noising to remove noise from images. To do this, we used cv2.fastNlMeansDenoising() function because of its faster processing while maintaining the denoising performance.
2. Opening and closing: Opening and Closing are used as morphological operations to manipulate the shapes and structures of objects within an image. Closing and opening was used to improve the quality of the input image for further process of edge detection by reducing noise,smoothing contours .
3. Hough line transformation: Hough transformation was used after using Canny edge detection as this edge description is commonly obtained using Canny may be noisy, i.e. it may contain multiple edge fragments corresponding to a single whole feature. Furthermore, the output of an edge detector like Canny defines only where features are in an image, the work of the Hough transform is to determine both what the features are and how many of them exist in the image.

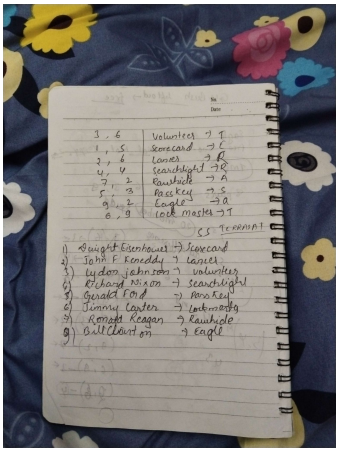

Figure 4.1: Original image

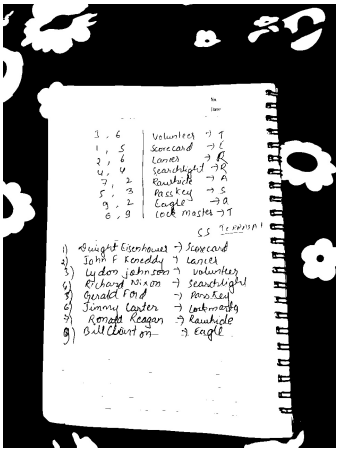

Figure 4.2: Threshold Image

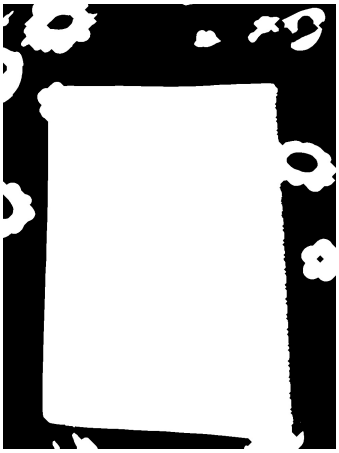

Figure 4.3: Closed Image

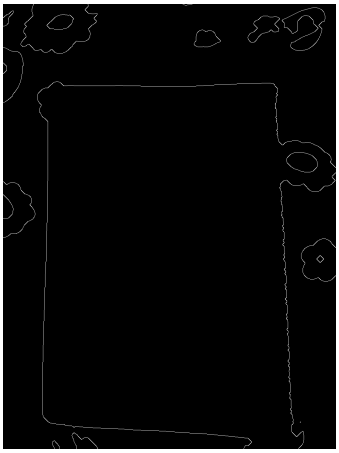

Figure 4.4: Canny Edge

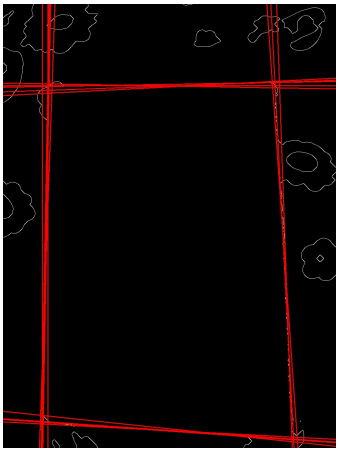

Figure 4.5: Hough Line Image

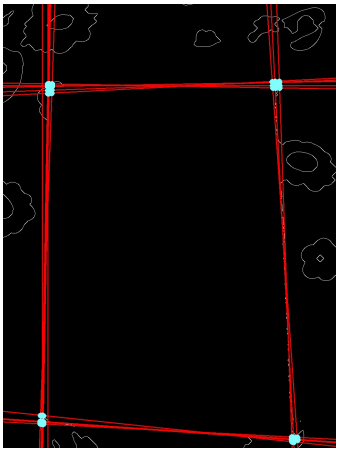

Figure 4.6: Intersection Point Image

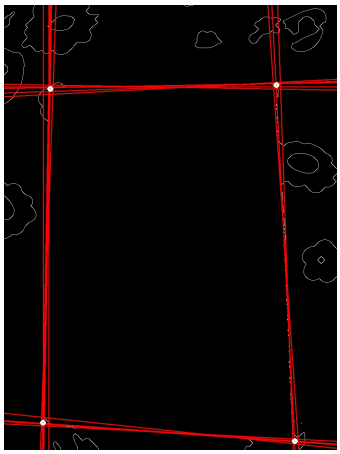

Figure 4.7: K-means Clustering Image

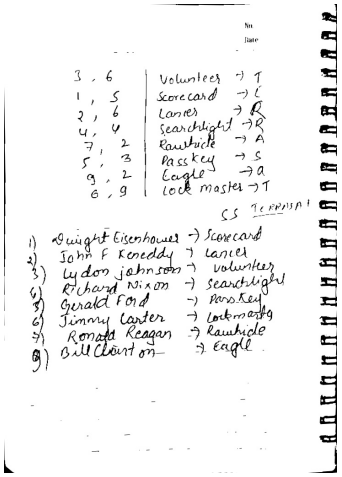

Figure 4.8: Preprocessing Output Image

4. Mobile App: The Mobile app for SnapTag has been developed using Flutter.

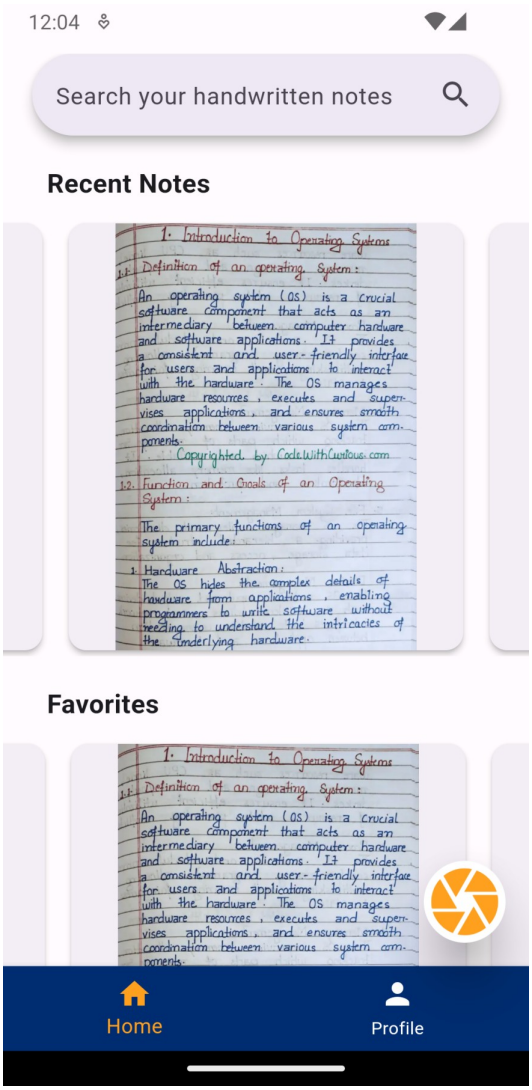

Figure 4.9: Homepage

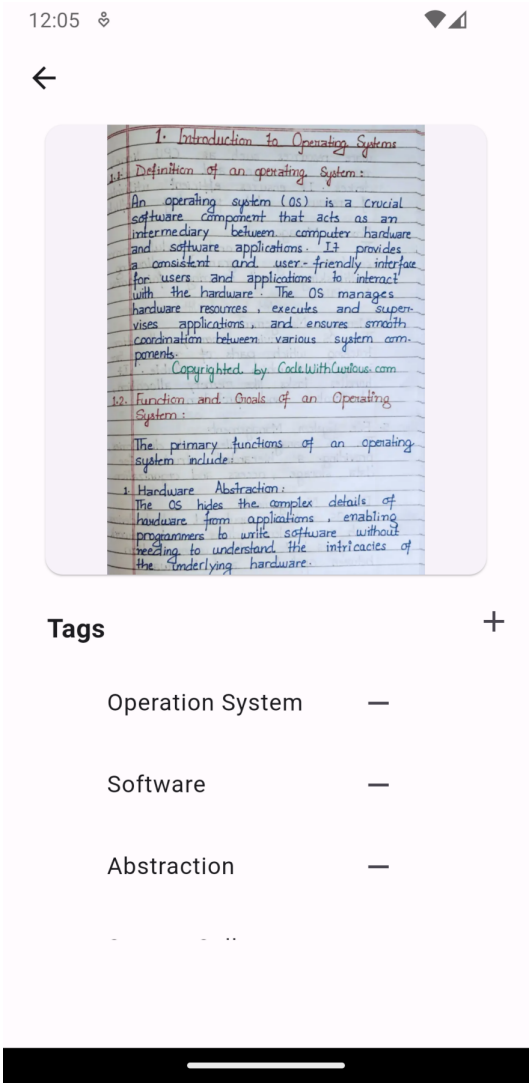

Figure 4.10: Image Capturing

4.3 Work Remaining

1. Segmentation: Till now, Line Segmentation has been done which still needs some improvements. After that, our immediate work will be segmentation words and characters. The following is the result produced using traditional Contour-Based Segmentation for lines.
2. Model Training: The efficiency of the trained model is 82% and we are working on improving the efficiency.