VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Project Report on

DetectWave: Scenario Recognition From Image

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai

Academic Year 2023-24

Submitted by

Gautam Dinga (D17 - A , Roll no -14)
Harsh Loya (D17 - A , Roll no -36)
Viraj Joshi (D17 - A , Roll no -26)
Harsh Patil (D17 - A , Roll no -51)

Project Mentor
Mrs.Mannat Doultani

(2023-24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Certificate

This is to certify that **Harsh Loya(D17A/36), Gautam Dinga(D17A/14), Viraj Joshi(D17A/26), Harsh Patil (D17A/51)** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "**DETECTWAVE: Scenario Recognition from Image**" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Prof. Mannat Doultani*** in the year 2023-24 .

This project report entitled **DetectWave - Scenario Detection from Image** by **Harsh Loya, Viraj Joshi, Gautam Dinga, Harsh Patil** is approved for the degree of **B.E Computer Engineering**.

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

Date:
Project Guide:

---

[1]

Project Report Approval
For
B. E (Computer Engineering)

This thesis/dissertation/project report entitled "**DETECTWAVE: Scenario Recognition from Image"** by (**Harsh Loya(D17A/ 36), Gautam Dinga(D17A/14), Viraj Joshi(D17A/), Harsh Patil (D17A/)** is approved for the degree of **B.E Computer Engineering.**

Internal Examiner

External Examiner

Head of the Department

Principal

Date:

Place: Mumbai

[1]

Declaration

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

---	---
Gautam Dinga (D17A / 14)	Harsh Loya (D17A / 36)
---	---
Viraj Joshi (D17A / 26)	Harsh Patil (D17A / 51)

Date:

ACKNOWLEDGEMENT

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

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

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

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

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

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

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

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

Index

Chapter No.	Title	Page No.
1	**Introduction**	10
1.1	Introduction of the project	10
1.2	Motivation for the project	11
1.3	Problem Definition	13
1.4	Lacuna of the Existing Systems	14
1.5	Relevance of the Project	14
2	**Literature Survey**	15
A	Brief overview of Literature Survey	15
2.1	Research Papers a. Abstract of research paper. b. Inference drawn from the paper.	15
2.2	Comparison with the Existing Systems	19
3	**Requirement Gathering for the proposed System**	21
3.1	Introduction of Requirement Gathering	21
3.2	Functional Requirements	21
3.3	Non-Functional Requirements	21
3.4	Hardware, Software, Technology and tools utilized	22
3.5	Constraints	23
4	**Proposed Design**	24
4.1	Block diagram of the system	24
4.2	Modular design of the system	26
4.3	Detailed Design (Flowchart)	27
4.4	Project Scheduling: Tracking using Timeline / Gantt Chart	30
5	**Implementation of the Proposed System**	31
5.1	Methodology employed for development	31

5.2	Algorithms and flowcharts for the respective modules developed	33
5.3	Datasets source and utilization	35
6	**Testing of the Proposed System**	
6.1	Introduction to testing	38
6.2	Types of tests Considered	38
7	**Results and Discussion**	40
7.1	Screenshots of User Interface (UI) for the respective module	40
7.2	Performance Evaluation measures	41
7.3	Input Parameters / Features considered	42
7.4	Comparison of results with existing systems	43
7.5	Inference drawn	43
8	**Conclusion**	44
8.1	Limitations	44
8.2	Conclusion	44
8.3	Future Scope	44
	References	45
	Appendix	47
1	**Paper I**	47
a	Paper I	47
b	Plagiarism Report of Paper I	54
c	Certificates	55
c	Project review sheet	56

List of Figures

Figure No.	Heading	Page No.
4.1	Visual representation of the system.	24
4.2	Block Diagram.	25
4.3	Modular Diagram.	26
4.4	Flowchart of the system.	27
4.5	Flow of the DetectWave System	29
4.6	Gantt chart.	30
5.1	Flowchart of CNN ML model.	34
5.2	Flowchart of RNN ML model.	35
5.3	Flickr 8k dataset on Kaggle	36
5.4	Caption.txt file in Flickr 8k dataset.	36
5.5	Images file in Flickr 8k dataset.	37
7.1	Screenshot of User Interface.	40
7.2	Screenshot of Upload Image UI	40
7.3	Screenshot of Scenario and Audio generated.	41
7.4	Screenshot of the BLEU score (Accuracy) of the ML model.	42
7.5	Input parameter of the DetectWave	42

Abstract

Our system is designed to detect and recognize objects and scenes in images. Users can upload an image, and the system will automatically identify objects within the image and the overall scene it represents. It uses a sophisticated model, Faster R-CNN, which efficiently processes images to pinpoint objects and their categories. The system also recognizes scenes such as gardens, parks, and more, providing context to the detected objects. The results are presented in a clear and concise format, making it accessible for users to understand the content of their images. Whether for image classification or scene recognition, this system simplifies the process, making it useful for a range of applications from content tagging to automated image analysis.

Our system is an intuitive and efficient tool that enhances image understanding by automatically identifying objects and scenes within uploaded images. Powered by the advanced Faster R-CNN model, it excels at the precise detection and classification of objects, ranging from trees and flowers to garden tools and furniture commonly found in garden settings. Beyond object recognition, the system provides valuable context by recognizing scenes, such as gardens, parks, and specific garden-related activities. This dual capability empowers users to not only comprehend the contents of their images but also the broader context in which these objects exist. The results are presented in an accessible and informative format, making it valuable for tasks like content tagging, automated image organization, and aiding those with visual impairments in interpreting images. Whether for personal use or professional applications, our system simplifies the process of gaining insights from images, offering a versatile and user-friendly solution for a wide range of image-related tasks.

Chapter 1 - Introduction

1.1 Introduction

In the contemporary digital landscape, the ability to interpret and comprehend the content of images has emerged as a pivotal frontier in artificial intelligence (AI). Within this expansive field, scenario recognition from images stands out as a fascinating and complex domain of computer vision. Its goal is to empower machines with the capability to categorize and understand scenes or scenarios depicted in visual data, thereby bridging the chasm between human perception and the capabilities of artificial intelligence.

This report encapsulates a comprehensive exploration and implementation of an advanced system for scenario recognition from images. The project's foundation lies in addressing the existing challenges associated with contemporary scenario recognition systems, which often struggle to grasp the broader context of an image. There is a recognized semantic gap between the labels or categories utilized for recognition and the nuanced meaning of real-world scenarios. Moreover, many current models are confined by the datasets on which they are trained, limiting their adaptability to novel or dynamic situations.

The primary objective of this project is to pioneer the development of an automated system capable of categorizing diverse scenarios with precision, drawing insights from the content of images. These scenarios span a spectrum, encompassing everyday activities, complex events, outdoor and indoor settings, emergencies, leisure, transportation, and more. The potential applications of accurate scenario recognition are far-reaching and extend across various industries, including but not limited to autonomous vehicles, real-time surveillance, and smart environments.

The journey towards achieving this objective unfolds through a meticulously planned sequence of steps. Beginning with the crucial phase of data collection and preprocessing, a diverse dataset is curated, enriched with accurate scenario labels, object details, activities, and contextual information. This dataset serves as the bedrock

for training the machine learning model, with an emphasis on capturing the intricacies of real-world scenarios.

Feature extraction, a pivotal stage in the process, leverages advanced Convolutional Neural Network (CNN) architectures pre-trained on large image datasets, such as ImageNet. This not only facilitates the extraction of meaningful visual features from images but also allows for the adaptation of pre-trained CNNs to the specific task of scenario recognition through fine-tuning on the target dataset.

Contextual understanding is then achieved through the implementation of object detection models, such as YOLO and Faster R-CNN, coupled with semantic segmentation techniques. This amalgamation empowers the system to not only identify and locate objects within images but also to discern their roles and relationships, thereby enhancing overall semantic understanding.

The report further delves into the intricacies of semantic inference, where the results of semantic segmentation and object detection are integrated. Additionally, attribute recognition is implemented to infer attributes and functions of objects, contributing to a more nuanced and comprehensive understanding of the depicted scenarios.

As the project unfolds, it incorporates a continuous improvement cycle, incorporating user feedback, data enhancement, and ongoing model training. Ethical considerations and real-time processing optimizations add depth to the project's framework, ensuring that the system not only excels in technical aspects but also adheres to ethical standards and practical deployment requirements.

This report encapsulates a journey through the realms of computer vision, machine learning, and artificial intelligence, with the overarching goal of contributing to the advancement of scenario recognition systems. The subsequent sections delve into the methodological intricacies, model development, user interface design, deployment strategies, and extensive testing that collectively constitute this ambitious and impactful project.

1.2 Motivation

The motivation behind embarking on the comprehensive exploration and implementation of an advanced scenario recognition system stems from the intrinsic

challenges and opportunities presented in the contemporary landscape of artificial intelligence and computer vision. At the core of this motivation is the recognition that existing scenario recognition systems often encounter limitations in capturing the broader context of an image. These limitations manifest as a semantic gap, where the predefined labels or categories used for recognition fall short in encapsulating the nuanced meaning of real-world scenarios. The motivation is rooted in the ambition to transcend these constraints, pushing the boundaries of what AI systems can achieve in terms of understanding complex visual data.

In envisioning the future of artificial intelligence applications, there is a compelling need for systems that go beyond mere object recognition within images. The aspiration is to develop machines that can comprehend and categorize diverse scenarios with a level of precision and adaptability that mirrors human perception. This motivation arises from a recognition that as technology continues to advance, there is a growing demand for AI systems to not only identify objects but also to grasp the relationships, activities, and environmental details that collectively form a scenario.

The overarching goal is to enhance the precision of scenario recognition in complex, real-world situations. This precision is not just about accurately identifying and labeling scenarios but also about understanding the dynamic and evolving nature of scenarios, adapting to novel or unforeseen circumstances. The motivation is to empower machines with a contextual understanding that enables them to navigate and interact seamlessly in diverse environments.

Furthermore, the motivation extends to the broad range of potential applications that accurate scenario recognition can unlock. From optimizing the performance of autonomous vehicles by enabling them to navigate complex traffic scenarios to enhancing real-time surveillance systems capable of identifying anomalous activities, the implications are vast and impactful. The motivation, therefore, is not only to advance the capabilities of AI systems but also to catalyze innovation across industries.

As this project unfolds, the motivation is not just confined to technical achievements but is deeply rooted in the ethical imperative of responsible AI deployment. The aim is to develop systems that not only excel in their technical capacities but also adhere to

ethical standards and practical deployment requirements, ensuring a positive and responsible impact on society.

1.3 Problem Definition

The problem landscape in contemporary scenario recognition systems is multifaceted and hinges on the challenges that impede the seamless and nuanced interpretation of visual data by artificial intelligence. One of the primary challenges is the existence of a semantic gap, a disconnect between the labels or categories used for recognition and the intricate, real-world meaning of diverse scenarios.

In essence, existing systems often struggle to go beyond basic object recognition and fail to capture the holistic context within which these objects exist. This semantic gap poses a significant hindrance to the development of scenario recognition systems that can truly mirror human-like comprehension. While objects within an image can be accurately labeled, understanding the relationships, activities, and environmental nuances that define a scenario remains a formidable challenge.

Another critical problem lies in the limitations imposed by the datasets on which these models are trained. Many current models are confined to datasets that may lack the diversity and complexity inherent in real-world scenarios. As a result, these systems may encounter difficulties when faced with novel or dynamic situations that deviate from their training data.

The adaptability of scenario recognition systems to a rapidly changing and diverse environment is a core problem. Real-world scenarios are inherently dynamic and can involve an evolving set of objects, activities, and relationships. Existing models may struggle to adapt to these changes, leading to a lack of accuracy and reliability in recognizing scenarios outside their training scope.

Furthermore, the problem extends to the need for systems capable of handling a broad spectrum of scenarios, ranging from routine activities to complex events, emergencies, and various indoor and outdoor settings. Many current models may excel in specific domains but lack the versatility required for comprehensive scenario recognition.

Addressing these problems is pivotal for the development of scenario recognition systems that can truly bridge the gap between human perception and artificial

intelligence. Overcoming the semantic gap, enhancing adaptability to novel situations, and achieving a comprehensive understanding of diverse scenarios constitute the key problem definitions that this project aims to tackle. By doing so, the project seeks to contribute significantly to the advancement of scenario recognition systems and their applications across diverse industries.

1.4 Lacuna of the existing systems

Many current scenario recognition systems lack the ability to understand the broader context of an image.

- There is often a semantic gap between the labels or categories used for recognition and the nuanced meaning of scenarios in the real world.
- Many existing models are trained on specific datasets, making them less adaptable to recognize scenarios outside their training data.
- This lack of generalization hampers their ability to perform effectively in novel or dynamic situations.
- In some applications like autonomous vehicles or real-time surveillance, there's a need for scenario recognition to be performed in real-time. Many current models may not meet the speed and latency requirements of such applications.

1.5 Relevance of the Project

The relevance of the project on advanced scenario recognition from images is underscored by its potential to revolutionize the landscape of artificial intelligence and computer vision applications. In the contemporary era, where visual data plays an increasingly vital role, the ability to accurately and comprehensively recognize diverse scenarios holds immense significance across various domains.

Chapter 2 - Literature Survey

2.1 Brief Overview of Literature Survey Referred

Sr no	Title	Abstract	Inference
1.	Malinowski, M., & Fritz, M. (2014). A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In Advances in Neural Information Processing Systems (NIPS)	The paper introduces a novel approach to question answering in real-world scenes,emphasizing its applicability in situations with uncertain or ambiguous input data. The authors advocate a "Multi-World Approach," which involves considering multiple potential interpretations or "worlds" of a scene based on the uncertain input.	The paper titled "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input" by Malinowski and Fritz focuses on the task of question answering, partic - ularly in the context of real-world scenes.
2.	Gkioxari, G., & Malik, J. (2015). Finding action tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 759-768).	The paper focuses on the task of identifying and tracking actions in video sequences. It introduces a method for detecting and localizing action "tubes," which are spatio temporal regions representing actions in videos. The approach aims to improve the understanding of actions in videos and may have implications for enhancing action recognition systems.	The paper "Finding action tubes" by Gkioxari and Malik likely belongs to the field of computer vision and pattern recognition, with a specific focus on analyzing actions within video sequences. The paper primarily deals with the detection and localization of "action tubes," which are likely spatiotemporal representations of actions in videos.

3.	Ma, L., Lu, J., Shao, W., & Li, X. (2018). Local high-resolution representation using convolutional neural networks for scene recognition. Pattern Recognition.	The paper "Local high-resolution representation using convol - utional neural networks for scene recognition" by Ma, Lu, Shao, and Li addresses the challenge of scene recognition, which involves categorizing images based on the depicted scene type. The authors propose an innovative method that harnesses convolutional neural networks (CNNs) to create high-resolution representations of local image regions. This approach enables the extraction of fine-grained details from various parts of the image, enhancing the accuracy and effectiveness of scene recognition systems.	The paper "Local high - resolution representation using convolutional neural networks for scene recognition" by Ma, Lu, Shao, and Li addresses the challenge of scene recognition, which involves categorizing images based on the depicted scene type. The authors propose an innovative method that harnesses convolutional neural networks (CNNs) to create high - resolution representations of local image regions. This approach enables the extraction of fine-grained details from various parts of the image, enhancing the accuracy and effectiveness of scene recognition systems.
4.	Zhou, B., La Pedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Advances in Neural Information Processing Systems (NIPS) (pp. 487-495).	The paper likely focuses on scene recognition, the process of automatically categorizing images based on the depicted environments, like beaches, cities, or forests. It utilizes deep learning techniques, particularly deep neural networks, to extract features that aid in scene recognition. The Places database, a valuable resource for training and evaluating scene recognition	Based on the provided information, we can infer that this paper, titled "Learning deep features for scene recognition using places database," likely explores the development of deep learning techniques for scene recognition, with a particular focus on utilizing the Places database. The research could involve the creation of advanced feature

		models, is mentioned as the data source. The abstract may highlight novel contributions or improvements in scene recognition, potentially including enhanced accuracy, better generalization, or increased efficiency, all within the context of advancing the state-of-the-art in machine learning, computer vision, and scene understanding.	representations to improve the accuracy and effectiveness of scene recognition systems. This paper was presented at the Advances in Neural Information Processing Systems (NIPS) conference, indicating its significance in the field of machine learning and computer vision.
5.	Zhu, Y., Zhao, R., & Fei-Fei, L. (2015). Scene recognition with objectness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp.1228-1236).	The paper, titled "Scene recognition with objectness," is likely focused on the field of scene recognition within computer vision. It may introduce the concept of "objectness," a measure of the likelihood that a region within an image contains a meaningful object. The paper is expected to discuss methodologies that incorporate objectness as a key component in scene recognition, potentially involving the development of algorithms or models. The research's relevance to the computer vision community, particularly how it enhances the accuracy or efficiency of scene recognition systems, is likely highlighted in the abstract.	The paper "Scene recognition with objectness" by Zhu, Y., Zhao, and Fei-Fei, presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), is likely centered on the field of computer vision and scene recognition. It appears to introduce the concept of "objectness" and explores its role in scene recognition. The paper is expected to discuss methodologies and techniques that incorporate objectness, potentially leading to improvements in the accuracy or efficiency of scene recognition systems. Overall, the research contributes to the field of computer vision and its relevance is underscored

			by its presentation at a major CVPR conference.
6.	Zheng, L., Wang, S., & Tian, L. (2019) Scenario recognition: A survey. ACM Computing Surveys (CSUR), 51(1), 1-33.	The paper, titled "Scenario Recognition: A Survey," published in ACM Computing Surveys (CSUR), is likely a comprehensive survey article. It is expected to cover a broad overview of scenario recognition, which can include the automatic identification and categorization of different situations or scenarios based on data inputs, such as text, images, or sensor data.	Based on the provided citation, it's inferred that the paper titled "Scenario Recognition: A Survey" by Zheng, Wang, and Tian is a comprehensive survey article published in ACM Computing Surveys (CSUR). The paper likely offers an extensive review and analysis of scenario recognition, covering various techniques, methods, and applications in this field.
7.	Farhadi, A., Hejrati, M., Amin, S., & Young, P. (2011). Every picture tells a story: Generating sentences from images. In the European Conference on Computer Vision (ECCV) (pp. 15-29).	The paper likely discusses a method or system for generating descriptive sentences or captions from images. This process involves using computer vision and natural language processing techniques to automatically understand and describe the content of an image in human-readable text.	Based on the provided citation, it can be inferred that the paper, titled "Every picture tells a story: Generating sentences from images" by Farhadi, Hejrati, Amin, and Young, likely addresses the field of image captioning. The paper is expected to detail a method or system that automatically generates descriptive sentences or captions for images using computer vision and natural language processing techniques. This research is likely to be of significance in the domain of computer vision, as indicated by its presentation at the European Conference on Computer

			Vision (ECCV). The paper may also discuss potential applications and methodologies within the context of image understanding and text generation from images

2.2 Comparison with the existing system

Category	Existing System	Proposed Area of Work
Dataset Size	Large	Large(8000 images)
Model Architecture	CNN + LSTM	Customized CNN+RNN
Accuracy	85%	N/A (Proposed)
Training Time	48 hours	N/A (Proposed)
Inference Speed	0.03 sec/frame	N/A (Proposed)
Real-time Capability	Yes	No (Proposed)

Deployment Platform	Edge Device	N/A (Proposed)
Anomaly Detection	No	Yes (Proposed)
Robustness	Moderate	N/A (Proposed)
Scalability	Limited	N/A (Proposed)
Explainability	Low	N/A (Proposed)

Chapter 3 - Requirement Gathering for the Proposed System

In this chapter we are going to discuss the resources we have used and how we analyzed what the user actually needs and what we can provide. We will also discuss the functional and non-functional requirements and finally the software and hardware used.

3.1 Introduction to requirement gathering

The Requirement Gathering is a process of requirements discovery or generating list of requirements or collecting as many requirements as possible by end users. It is also called as requirements elicitation or requirement capture.

3.2 Functional Requirements

The "DetectWave: Scenario Recognition from Image" system needs to be easy for people to use. It should let users upload pictures so that the system can figure out what's happening in those pictures. The system should be able to recognize different situations. People in charge can use a special part of the system to teach it how to recognize things and check how well it's doing. If needed, users can sign in to make sure their information is safe. The system should be able to handle lots of pictures, work in real-time, and improve its guesses if they're not perfect. People should also be able to give feedback and connect the system with other programs. It's important to keep a record of what happens and make sure everything is safe and can grow if more people use it. If people speak different languages, the system should work for them too. Helpful guides and instructions should be available, and it should be easy to set up and use the system. These requirements help ensure that the system can recognize scenarios in pictures effectively and be useful to everyone.

3.3 Non-Functional Requirements

The non-functional requirements are like the behind-the-scenes rules for our system. They include things like making sure our system is fast, reliable, and easy to use. It

should also keep people's information safe and follow the law. The system needs to work well on different devices and should be easy to fix and update. If lots of people use it, it should still run smoothly, and it should keep a record of what's happening. It should be fair in how it recognizes scenarios and shouldn't cost too much to run. These requirements help make sure your system works well and is safe and fair for everyone to use.

3.4. Hardware, Software , Technology and tools utilized

A. Hardware requirement

a. Minimum 8 GB RAM

b. Core I5 7th Gen processor

c. NVIDIA GPU

d. Disk space of 4GB

B. Software requirement

a. Python

b. Streamlit

c. Google colab

d. Python Basic Stack:

e. Numpy

f. Matplotlib

g. Tensorflow

h. Pandas

C. Techniques -

a. **Python** - Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

b. **Streamlit** - Streamlit is an open-source Python library used for creating web applications with minimal effort.

c. **Faster R-CNN**: This is a region-based convolutional neural network that's designed

for object detection.

d. **TensorFlow Object Detection API**: Developed by Google, this is an open-source framework built on top of TensorFlow that provides pre-trained object detection models and tools for training custom models.

D. Tools -

- **Vs Code -** Visual Studio Code is a streamlined code editor with support for development operations like debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDE.
- **Google Colab -** Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.
- **Github -** GitHub is a web-based platform for version control using Git, which is a distributed version control system. It provides a hosting service for software development projects that use Git for version control.

3.5 Constraints

The output is only in English i.e the object is detected and it is displayed on the screen only in English. As it is using a pretrained model, more accuracy can be achieved and object detection is done better. Additionally the user has to upload an image for detecting the objects in the image.

Chapter 4: Proposed Design

4.1 Block diagram of the system

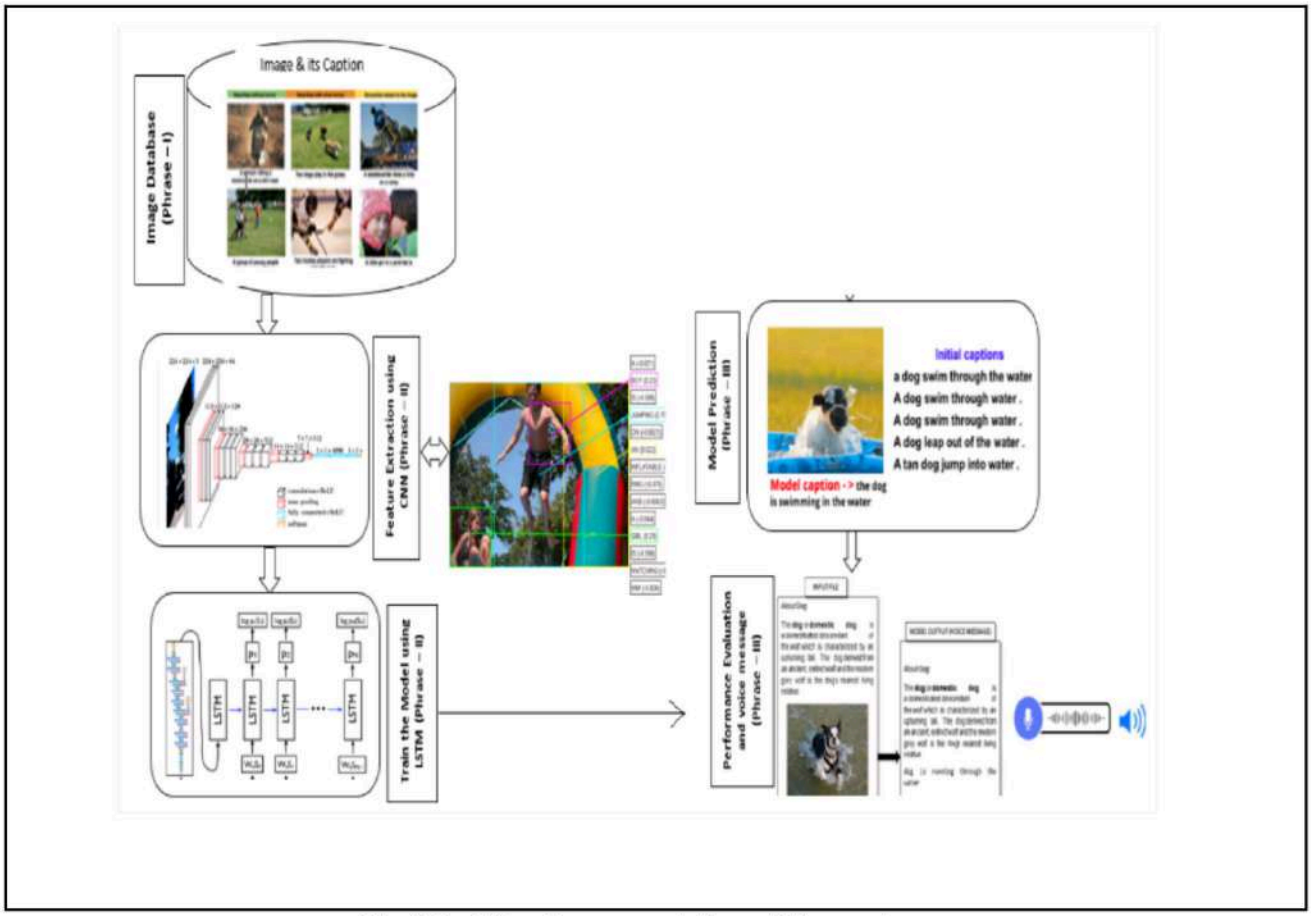

Fig 4.1 - Visual representation of the system

- Fig 4.1 explains the visual representation of the machine learning models which we had used in the scenario recognition system.
- In the above image the database is visually represented and also the layers of cnn to extract features from the image and then by using rnn we can generate a caption explaining the image.
- Also after generating caption a scenario is generated about the image by using generated caption using gpt 3.5 turbo model.
- The generated scenario is converted into audio format using espnet ml model is visually presented in the above image.

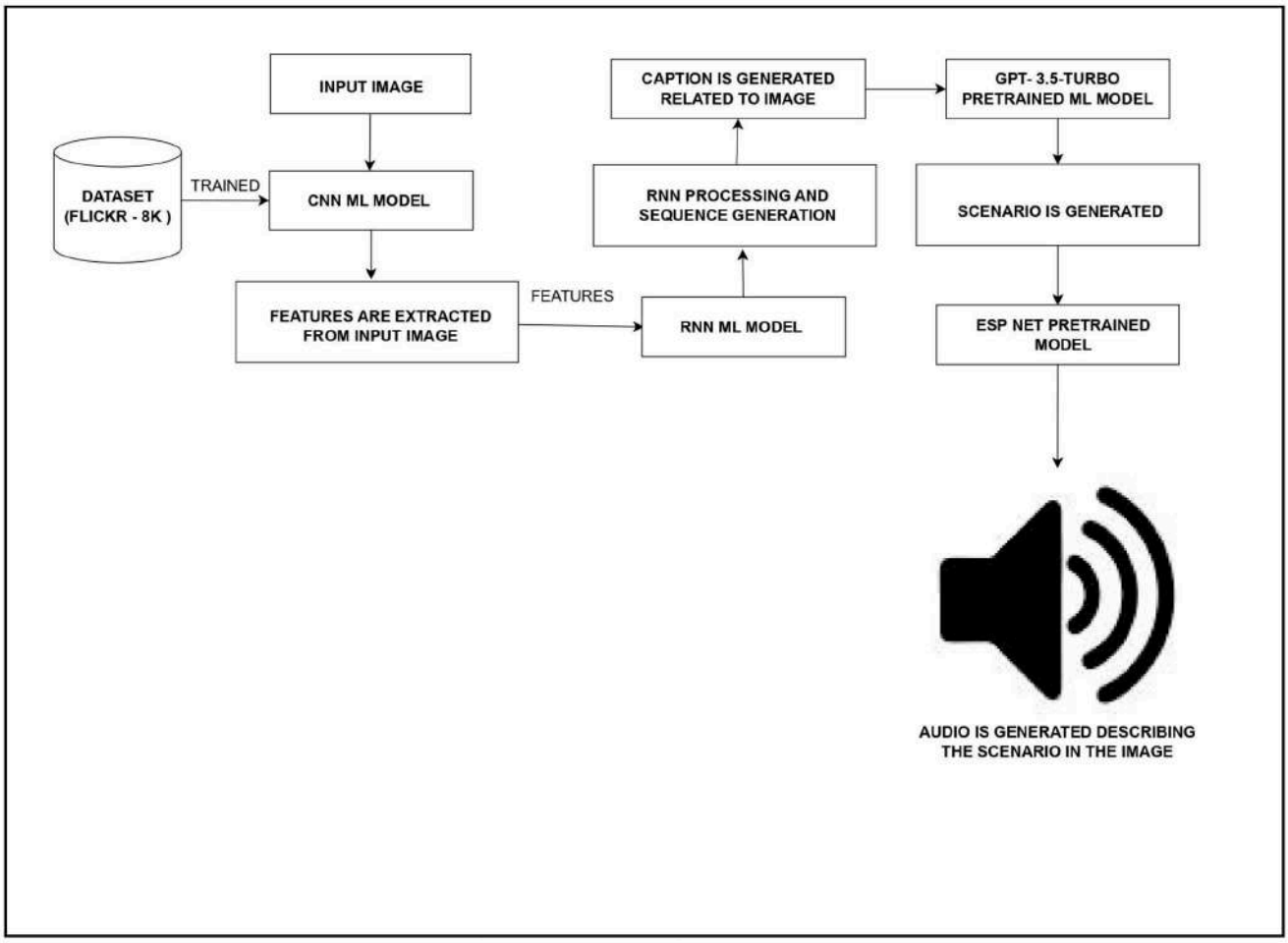

Fig 4.2 - Block Diagram

- Fig 4.2 explains the block diagram of the scenario recognition system.
- This fig explains different blocks of the scenario recognition system and how they are related to each other.
- The Flickr 8-k dataset is used to train the cnn model to extract features from images.
- RNN model is used to generate captions from the extracted images using LSTM model and Natural language preprocessing.

4.2 Modular design of the system

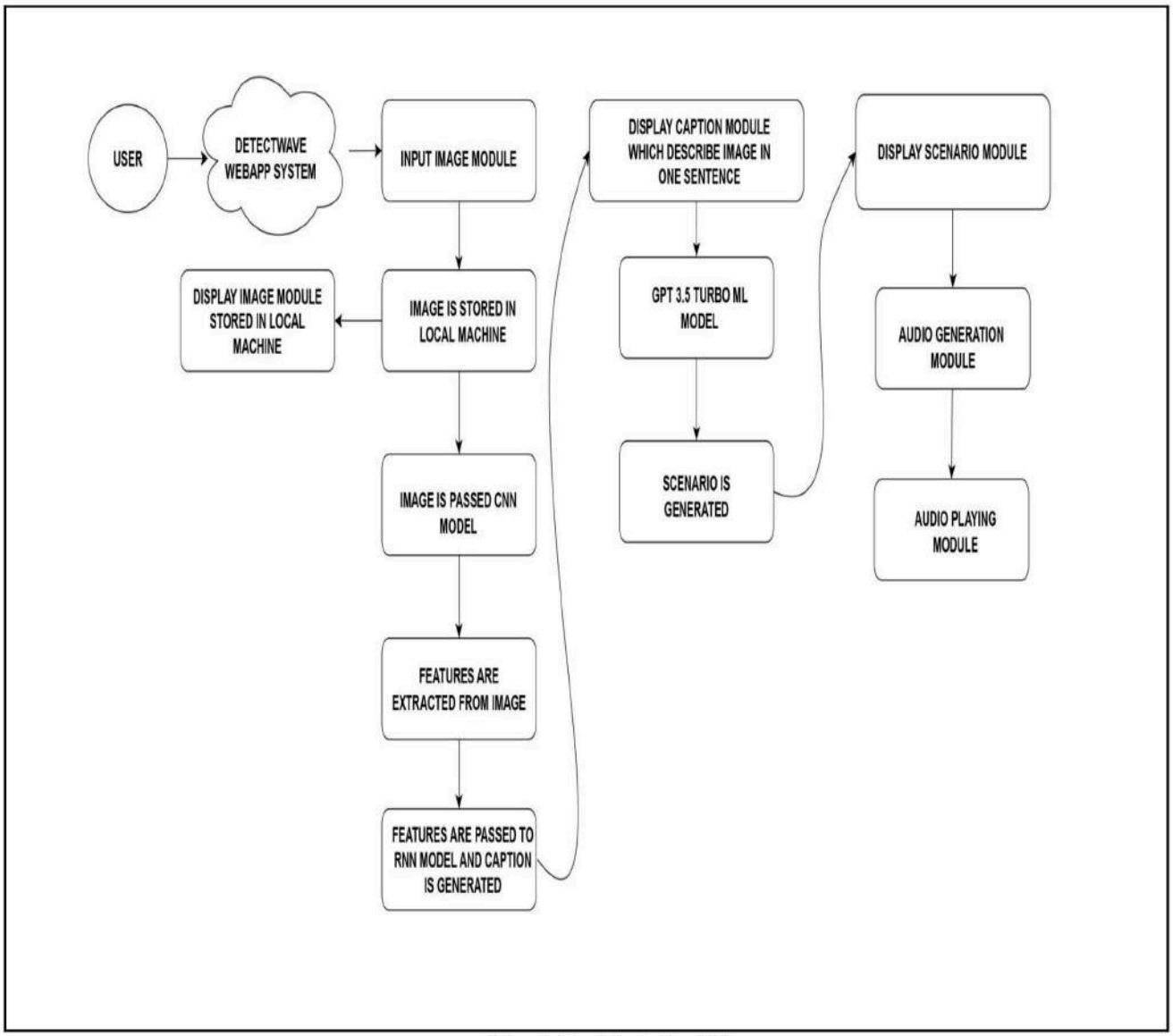

Fig 4.3 - Modular Diagram

Figure 4.3 shows: Detailed Modular Components Explanation:

1. Input Image Module:

- Preprocessing: Initial processing steps such as resizing, normalization, and noise reduction to prepare the image.
- CNN (Object Detection & Feature Extraction): Utilized for object detection and feature extraction from the input image.

2. CNN Captioning Module:

- Utilizes a CNN model to generate a caption based on the objects detected in the input image.

3. GPT-3.5 Turbo Text Generation Module:

- Receives the caption generated by the CNN model and utilizes GPT-3.5 Turbo to expand it into a full scenario in story format.

4. Text-to-Speech Conversion Module:

- Converts the generated story text into audio using the ESPnet ML model.

5. Story Text Module:

- Represents the generated story text.

6. Audio Output Module:

- Represents the final audio output of the scenario.

4.3 Detailed Design

Visual Representation of flowchart of the system -

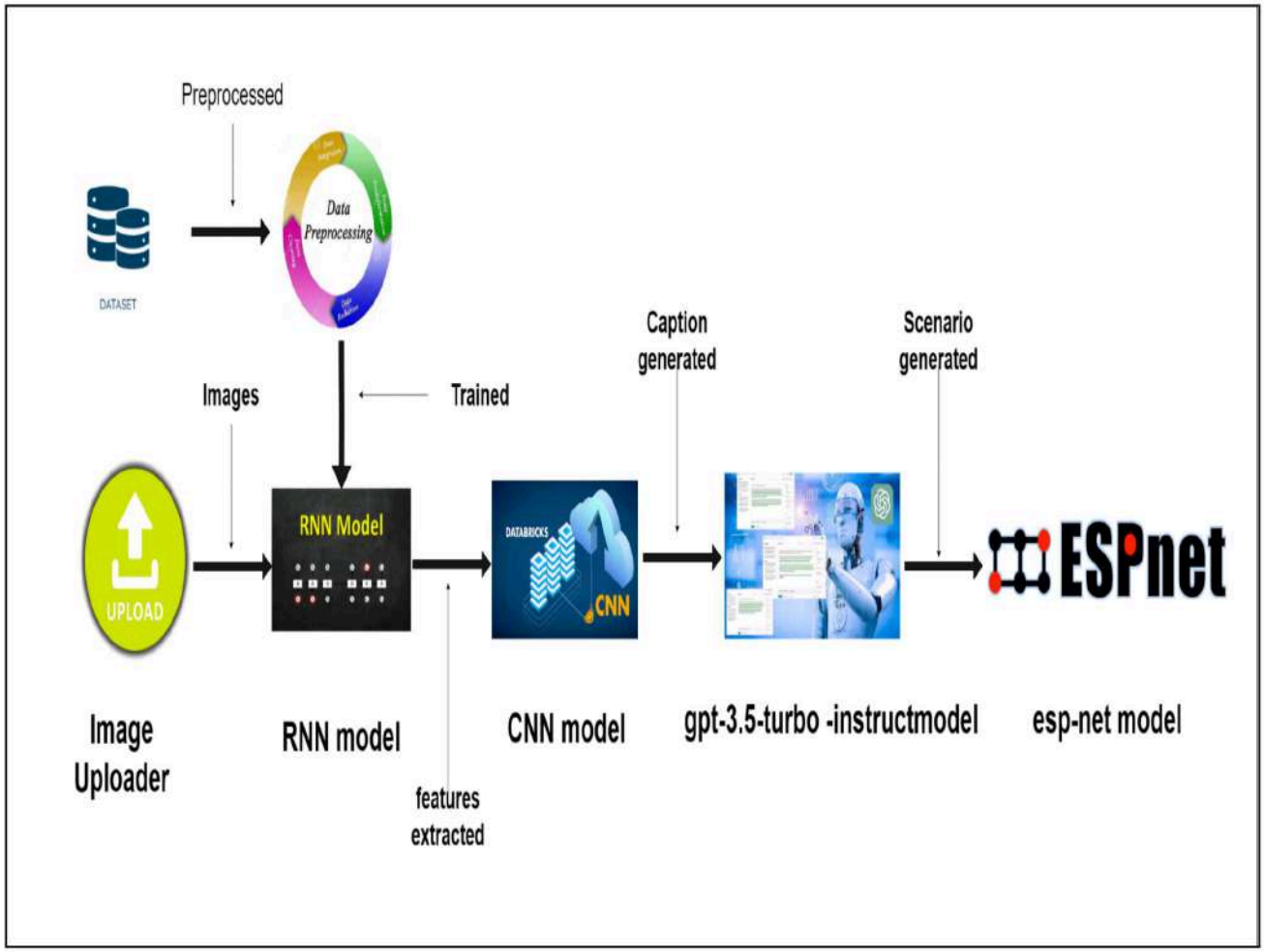

Fig 4.4 - Flow chart of the system

Figure 4.4 shows: Flow chart of the system is as follows -

1. Input Image:

- The process begins with an input image depicting a scenario that needs to be recognized.

2. Preprocessing:

- The input image undergoes preprocessing steps such as resizing, normalization, and noise reduction to prepare it for feature extraction.

3. Feature Extraction using CNN:

- The preprocessed image is fed into a Convolutional Neural Network (CNN), which extracts relevant features from the image. The CNN consists of convolutional layers, activation functions, and pooling layers.

4. Sequence Processing using RNN:

- The extracted features from the CNN are passed into a Recurrent Neural Network (RNN), which processes the features sequentially to capture temporal dependencies and context. This stage involves LSTM or GRU layers for sequence processing.

5. Text Generation using GPT-3.5 Turbo:

- The output from the RNN, which represents the contextual information of the scenario, is fed into the GPT-3.5 Turbo pretrained model. GPT-3.5 Turbo generates a textual description or story based on the contextual information provided.

6. Text-to-Audio Conversion using ESPnet:

- The generated text is then converted into audio format using the ESPnet ML model for text-to-speech conversion. This step enables the description of the scenario to be conveyed audibly.

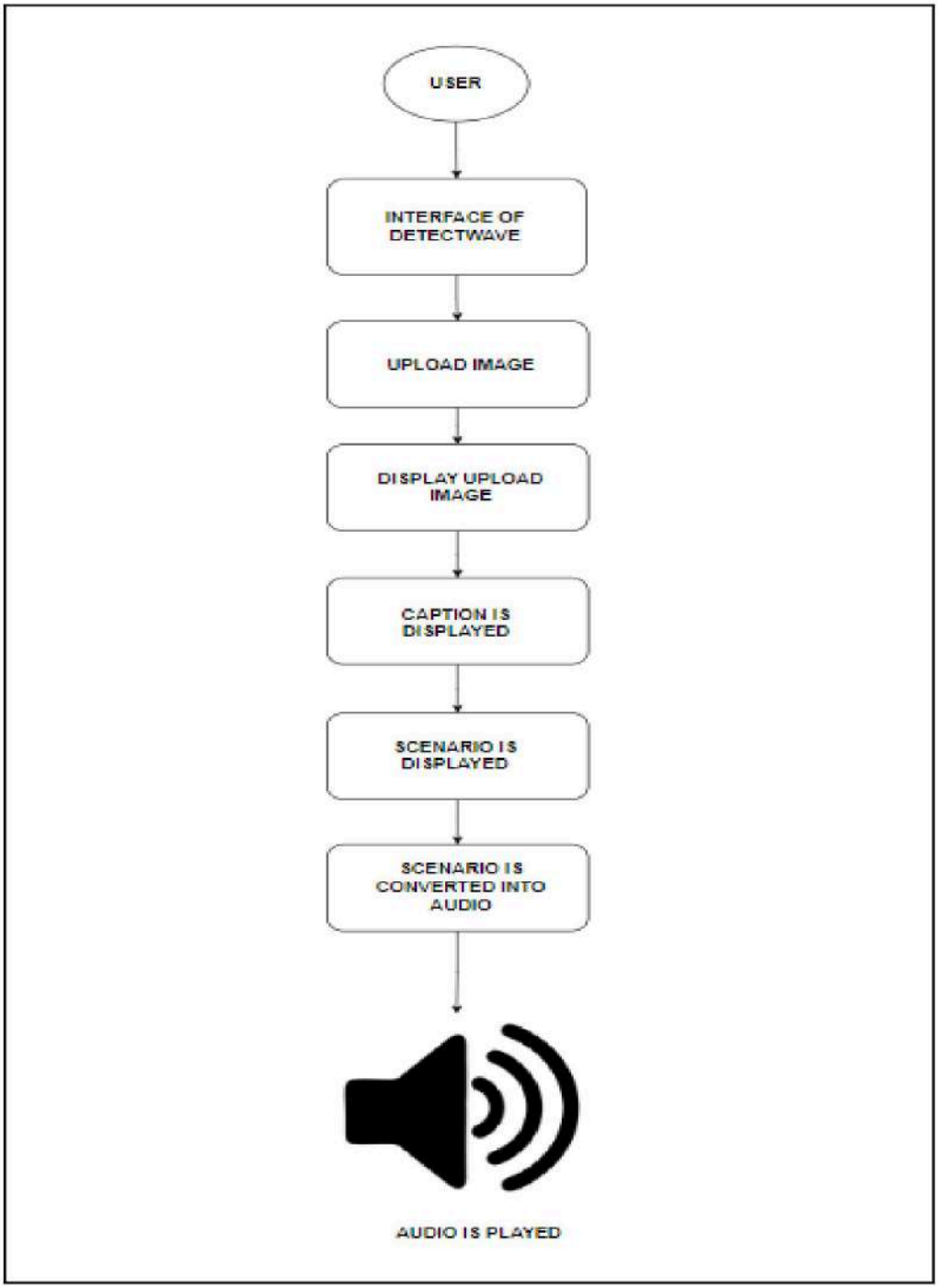

Fig 4.5 - Flow of the DetectWave System

7. Audio Output:

- The final output is the audio representation of the scenario description, which can be played back to the user, aiding in understanding the content of the input image.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

DetectWave - Scenario Recognition from Image

Gantt Chart:

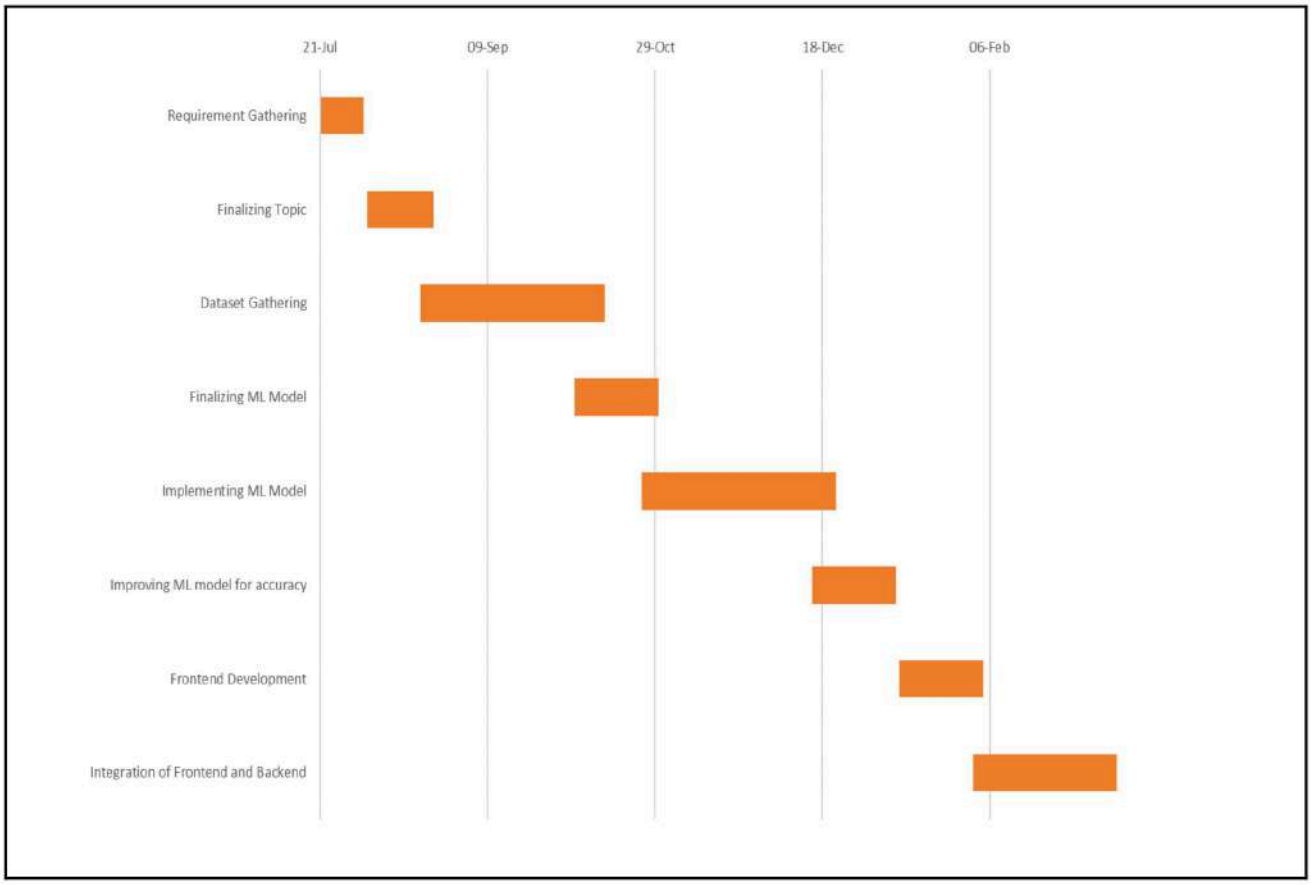

Fig 4.6 - Gantt Chart

Figure 4.6 shows:

The Gantt chart of our project where we worked for the whole semester to create this model is shown in a timeline pattern. It is the most important part to think and design the planning of your topic and so we planned our work like the gantt chart shown.

Chapter 5: Implementation of the Proposed System

5.1. Methodology employed for development

The development of a system that converts images into audio descriptions is a critical endeavor for society, as it addresses the pervasive issue of accessibility for individuals with visual impairments. By providing audio descriptions of visual content, this technology ensures equal access to information, enhancing education opportunities, social inclusion, and overall independence for visually impaired individuals. Access to visual information empowers individuals to engage with educational materials, social interactions, and professional endeavors, fostering diversity and inclusion in various spheres of life. Moreover, enabling access to cultural and artistic experiences through audio descriptions promotes a deeper appreciation for creativity and cultural heritage among visually impaired individuals. Additionally, the implementation of such systems aligns with legal and ethical imperatives regarding accessibility, demonstrating a commitment to inclusivity and social responsibility. Overall, the development of systems that convert images into audio descriptions plays a pivotal role in promoting accessibility, inclusion, and equal opportunities for individuals with visual impairments, enriching both their lives and society as a whole.

The system designed to convert images into audio descriptions serves a diverse range of users, each with distinct needs and preferences. Below are the types of users who could benefit from such a system:

1. Visually Impaired Individuals: This group comprises the primary users of the system. Visually impaired individuals, including those who are blind or have low vision, rely on audio descriptions to access and understand visual content. They use the system to gain information from images encountered in various contexts, such as educational materials, websites, social media, and everyday life.

2. Educators and Trainers: Educators and trainers working with visually impaired students or participants can use the system to create accessible educational materials. They can generate audio descriptions for images used in textbooks, presentations,

online courses, and training materials, ensuring that visually impaired learners have equal access to educational content.

3. Content Creators and Publishers: Content creators, including authors, journalists, bloggers, and social media influencers, can utilize the system to make their visual content accessible to a broader audience. By providing audio descriptions alongside images in their publications, websites, and social media posts, they enhance the inclusivity of their content and reach a larger audience, including visually impaired individuals.

4. Accessibility Advocates and Organizations: Individuals and organizations advocating for accessibility and disability rights may use the system to promote awareness and encourage the adoption of accessible practices. They can demonstrate the importance of providing audio descriptions for visual content and advocate for the implementation of accessibility standards in various industries and sectors.

5. Developers and Technologists: Developers and technologists involved in the design and implementation of digital accessibility solutions play a crucial role in advancing the capabilities of the system. They contribute to the development of the underlying technologies, such as image recognition algorithms, natural language processing models, and text-to-speech synthesis engines, to improve the accuracy, efficiency, and usability of the system for visually impaired users.

6. Caregivers and Support Professionals: Caregivers, family members, and support professionals working with visually impaired individuals may assist in using the system to facilitate access to visual content and promote independence. They can help visually impaired users navigate the interface, select images for description, and interpret the audio descriptions as needed.

7. Regulatory Bodies and Policy Makers: Regulatory bodies and policy makers responsible for implementing accessibility regulations and guidelines may utilize the system to evaluate compliance and enforce accessibility standards across various sectors. They can assess the effectiveness of the system in meeting the needs of visually impaired users and promote its adoption through policy initiatives and incentives.

5.2 Algorithms and flowcharts for the respective modules developed

We have used the following Machine Learning Models in our system -

- Convolution Neural Network (CNN)
- Recurrent Neural Network (RNN)

a) Convolution Neural Network (CNN) -

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for feature extraction from images. At the heart of a CNN are convolutional layers, which apply filters to input images, extracting features such as edges, textures, and shapes. These filters slide over the input image, convolving with it to produce feature maps that capture different aspects of the image content. Activation functions like ReLU introduce non-linearity, enabling CNNs to learn complex patterns and representations. Pooling layers further reduce the spatial dimensions of the feature maps while preserving important information. CNN architectures typically consist of multiple convolutional layers followed by pooling layers, each layer capturing increasingly abstract features. Through training with labeled data and optimization techniques like stochastic gradient descent, CNNs learn to extract discriminative features from images, enabling tasks such as image classification, object detection, and image segmentation. Transfer learning, where pre-trained CNN models are fine-tuned on specific tasks, is often employed to leverage learned features and accelerate training. Overall, CNNs are indispensable in computer vision, providing powerful tools for extracting meaningful features from images and solving a wide range of visual recognition tasks.

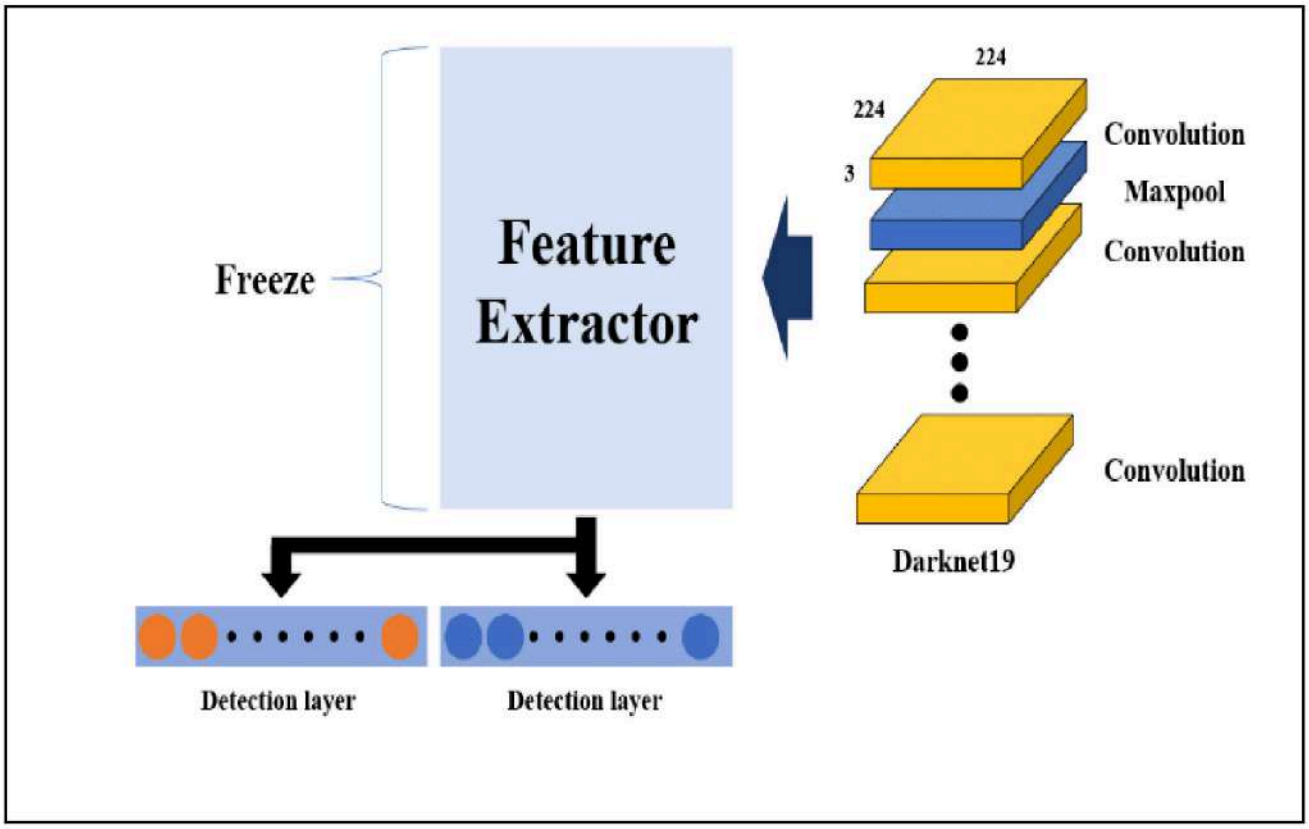

Fig 5.1 - Flowchart of CNN ML model

b) Recurrent Neural Network -

Using a Recurrent Neural Network (RNN) for generating captions of images involves a sequential process beginning with the extraction of features from the input image using a pre-trained convolutional neural network (CNN) such as ResNet or VGG. These features encapsulate the high-level visual information of the image and serve as the initial input to the RNN. Upon initialization, the RNN's hidden state captures the context for generating the caption. The caption generation process then commences by providing the image features and an initial token to the RNN. The network iteratively predicts the next word in the caption sequence based on the previous words generated, gradually constructing the caption until an end token is generated or a predefined maximum caption length is reached. The resulting sequence of words forms the generated caption. Evaluation metrics such as BLEU score or METEOR score are often employed to assess the quality of the generated captions. Additionally, fine-tuning of the

RNN and CNN parameters may be performed based on evaluation results to enhance caption quality. Ultimately, the generated captions can be utilized for tasks such as image indexing, retrieval, or aiding visually impaired individuals in comprehending image content, showcasing the versatility and utility of RNNs in natural language generation tasks related to visual data.

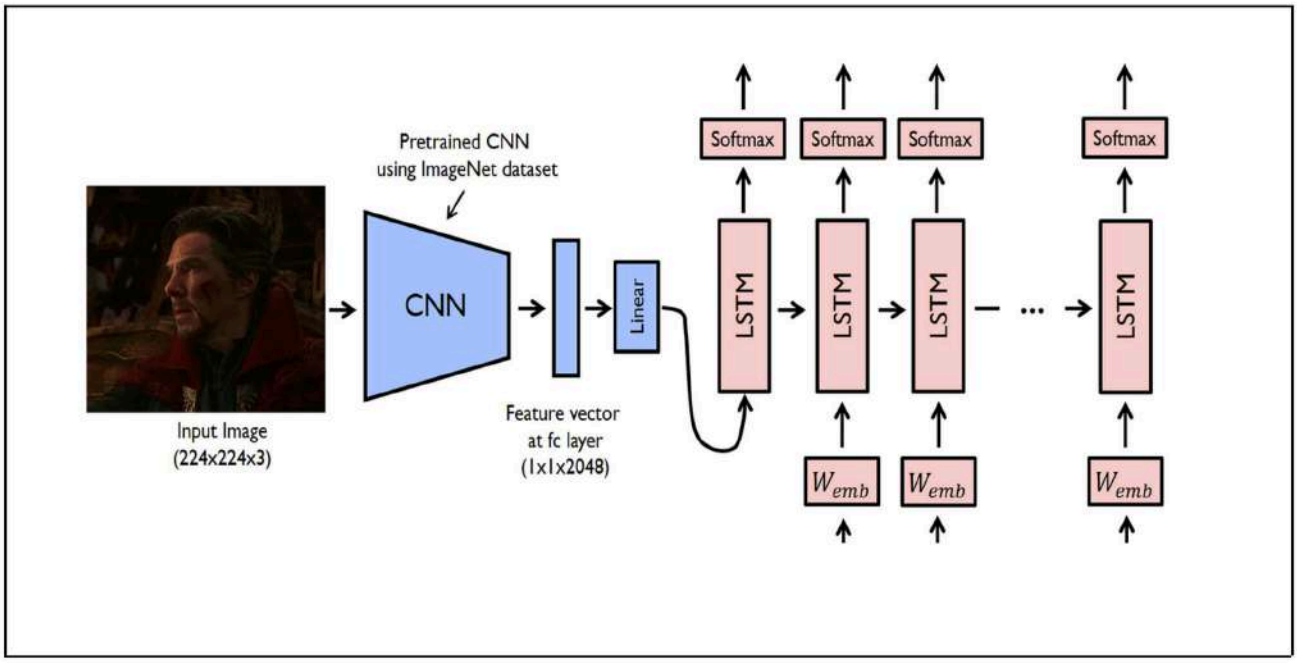

Fig 5.2 - Flowchart of RNN ML model

5.3 Datasets source and utilization

We have used the dataset 'Flickr 8-k' dataset from kaggle to obtain training data for the RNN and CNN ml model which helps to extract the features from the input image and convert it into caption. The dataset contains 8K images and each image is linked with 5 relatable labeled captions.

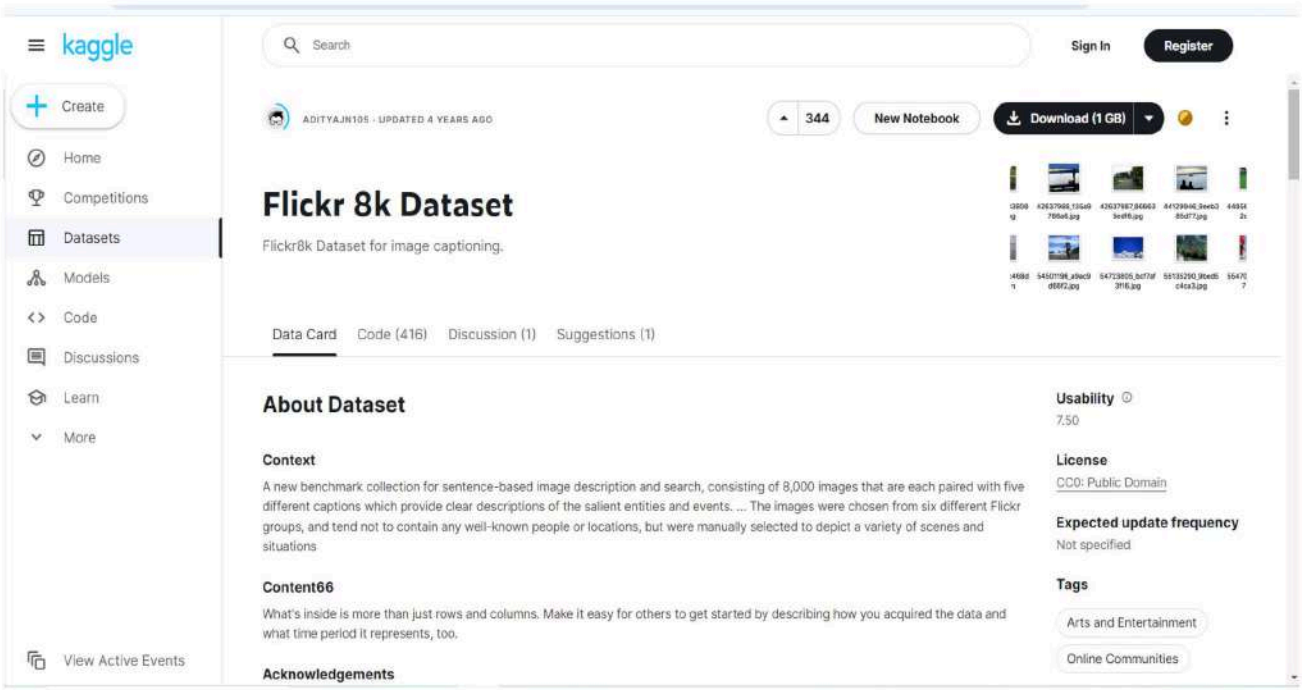

Fig 5.3 - Flickr 8k dataset on Kaggle

Fig 5.3 depicts the dataset used in our system. It contains 8K images which is used to train the cnn

ml model.

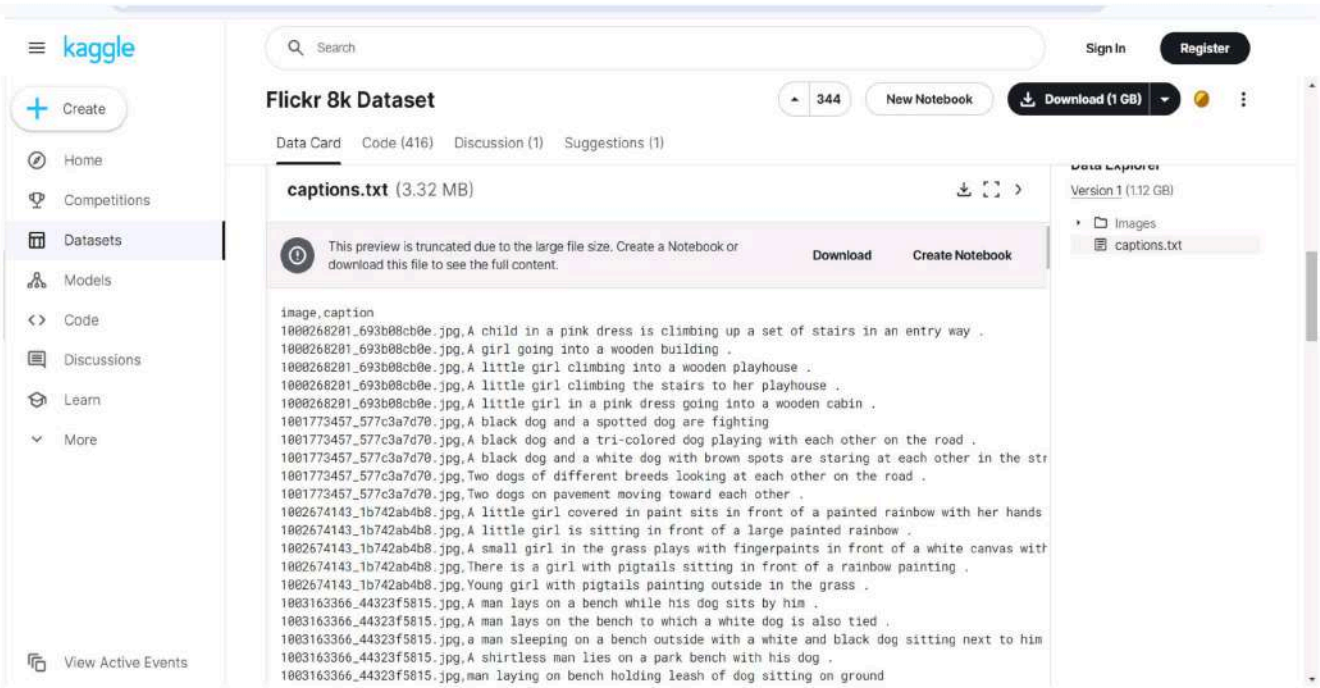

Fig 5.4 - Caption.txt file in Flickr 8k dataset

Fig 5.4 depicts the caption text file which contains the captions embedded to the specific images in the dataset which is used to train the model.

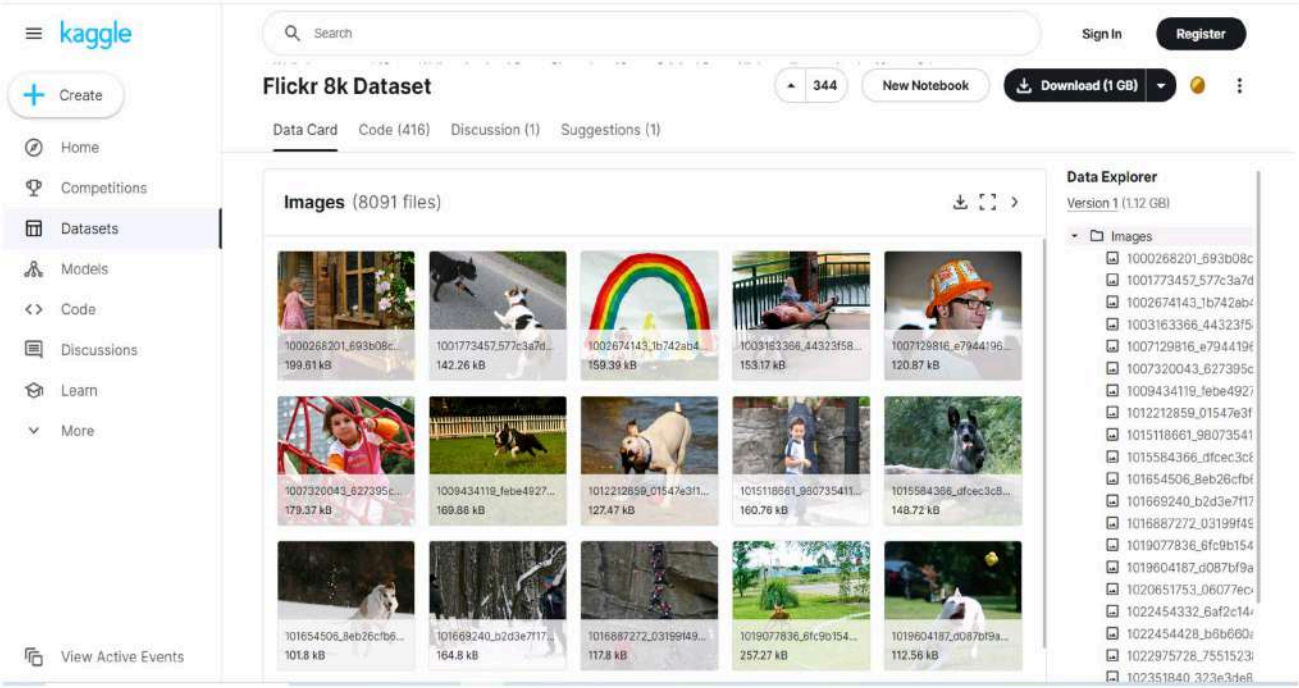

Fig 5.5 - Images file of Flickr 8k dataset

Fig 5.5 depicts the Flickr 8K dataset which contains the images which are used to train the CNN model.

Chapter 6: Testing of the Proposed System

6.1 . Introduction to testing

Software testing is the sequence of activities that happen during software testing. By employing a sane software testing life cycle, an organization ends up with a quality strategy more likely to produce better results. Why is this so important, though? It all boils down to customer satisfaction. Presenting a perfect product to the customer is the end goal of every organization. Nothing puts off customers more than bug-filled user experience. So when enterprises realized this, they began to include testing as a mandatory part of the SDLC. Since then, testing has become an integral part of every organization. Project Testing Phase means a group of activities designated for investigating and examining progress of a given project to provide stakeholders with information about actual levels of performance and quality of the project. It is an attempt to get an independent view of the project to allow stakeholders to evaluate and understand potential risks of project failure or mismatch. The purpose of the testing phase is to evaluate and test declared requirements, features, and expectations regarding the project prior to its delivery in order to ensure the project matches initial requirements stated in specification documents.

6.2. Types of tests Considered

Determining the most important test among all depends on various factors including the system's requirements, its intended use, and the potential impact of failure. However, for the system involving an RNN for generating image captions, some tests hold particular significance:

- **Functional Testing:** This is crucial as it ensures that the system meets its functional requirements, such as generating accurate and coherent captions for input images. Ensuring the correctness of the caption generation process is fundamental to the system's utility and effectiveness.
- **Usability Testing:** Assessing the system's usability is essential to guaranteeing a positive user experience. Since the system may be used by visually impaired

individuals, ensuring that it is intuitive and easy to use, particularly with assistive technologies, is paramount.

- **Accessibility Testing:** Given the potential user base of visually impaired individuals, accessibility testing is critical. This ensures that the system is usable by all users, regardless of their abilities, and that the generated captions are accurately conveyed through assistive technologies.
- **Security Testing:** Protecting sensitive data and ensuring the system's resilience against security threats is vital. As the system may process and generate captions for a variety of images, it's essential to verify that it handles user data securely and is resistant to potential attacks or vulnerabilities.
- **Performance Testing:** Evaluating the system's performance ensures that it can handle its intended workload efficiently. Since the system may need to process images and generate captions in real-time, performance testing helps identify and address any bottlenecks or scalability issues that could impact its responsiveness.

Chapter 7: Results and Discussion

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

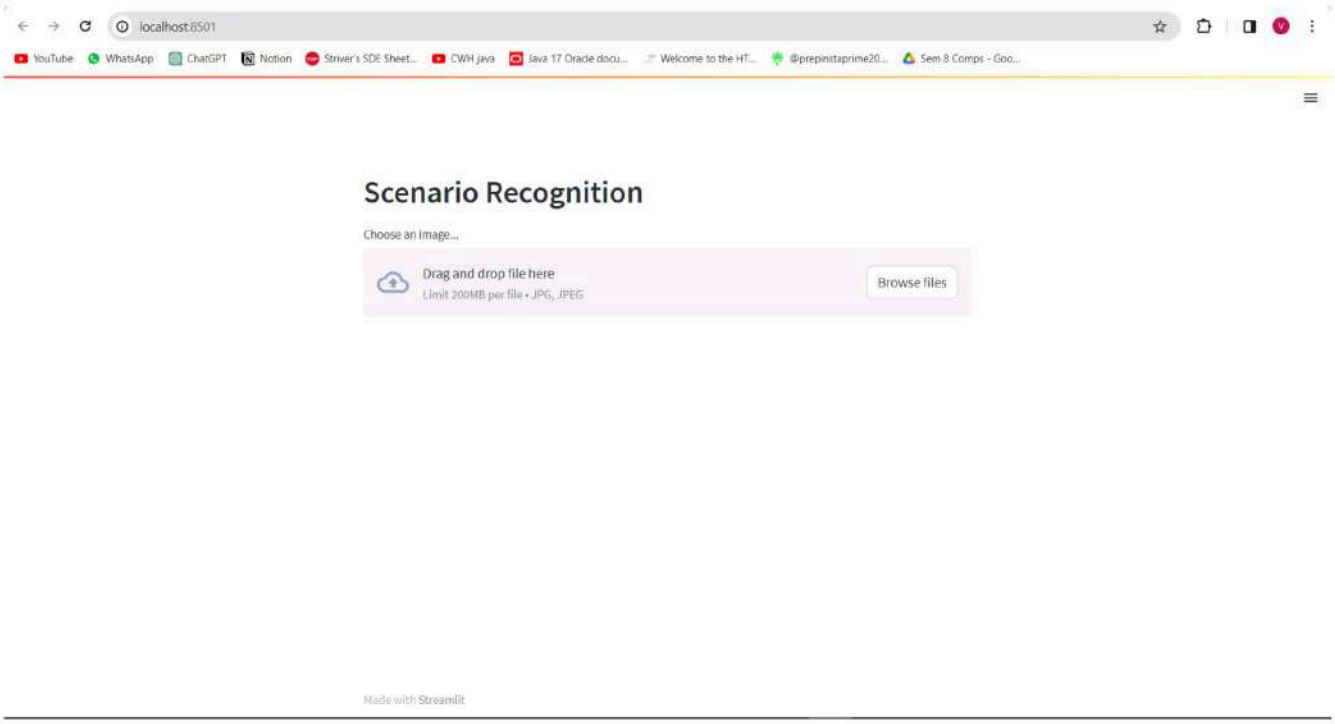

Fig 7.1 - Screenshot of User Interface

Fig 7.1 depicts the user interface which we had developed using streamlit which is a well known library of python.

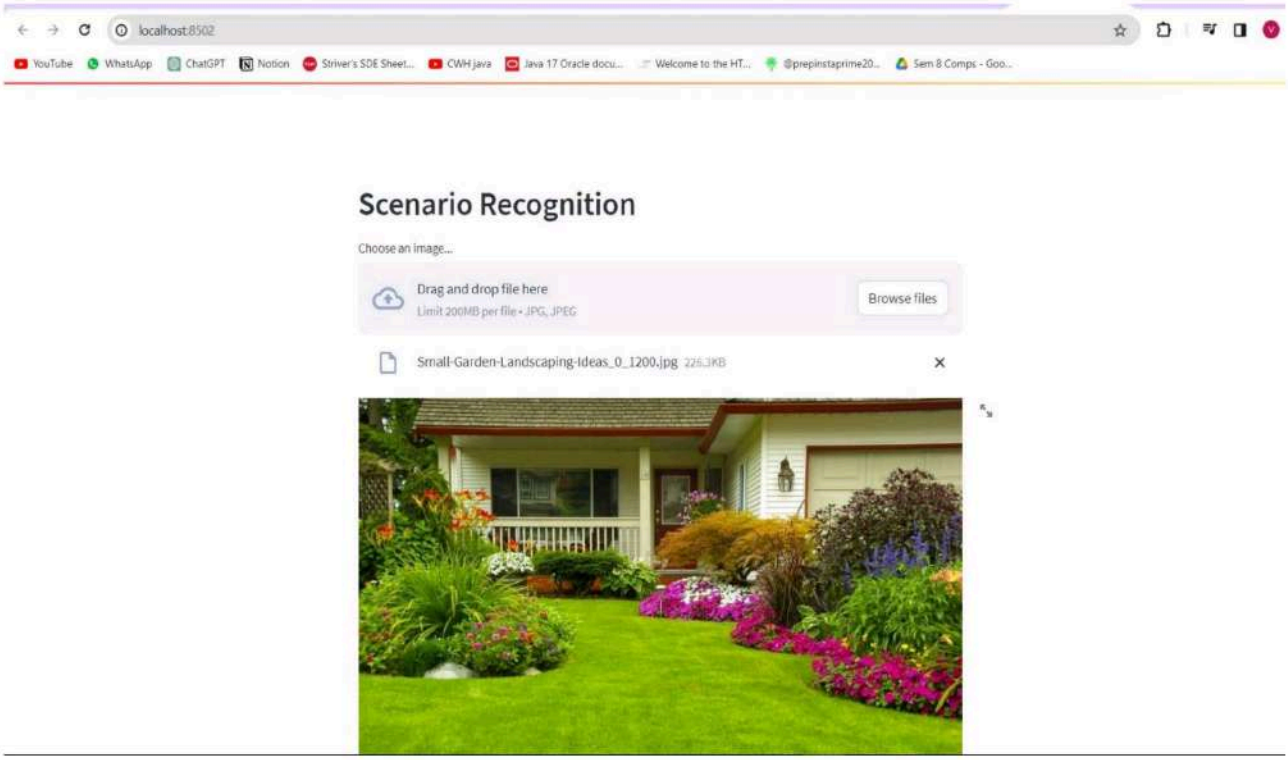

Fig 7.2 - Screenshot of Upload Image UI

Fig 7.2 depicts the image uploaded in the user interface which is given as input to the ml model.

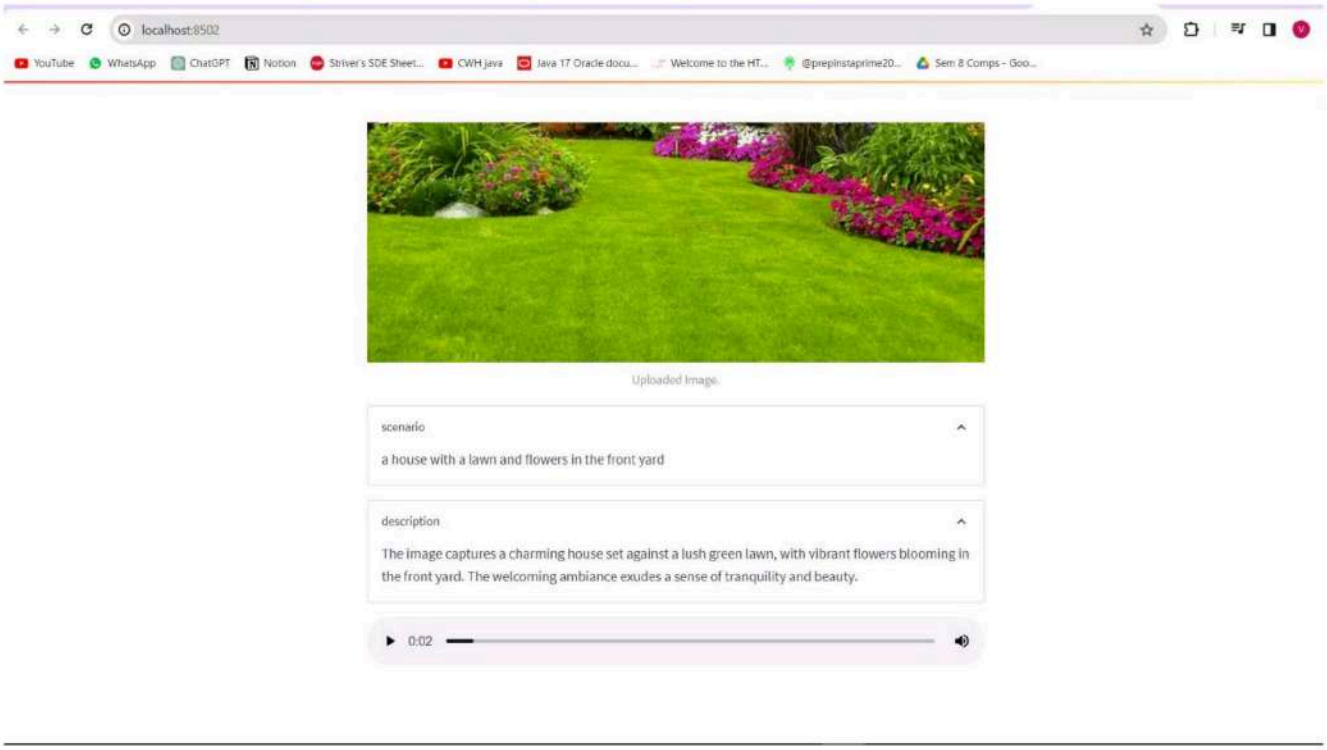

Fig 7.3 - Screenshot of scenario and audio generated

Fig 7.3 represents the scenario and audio generated as the output by the machine learning model.

7.2. Performance Evaluation measures

A performance metric is a quantitative assessment of a system, process, model, or algorithm's efficacy, efficiency, or quality. A measure of performance in machine learning evaluates a model's effectiveness on a job by comparing its outputs or predictions to the intended results or actual data.

Text descriptions from picture scenario recognition models may be evaluated using BLEU score, which was first used to evaluate machine translation. Precision-based n-gram scores are used to quantify the similarity between produced and reference descriptions while taking different durations into account.

BLEU evaluates text descriptions in image scenario recognition, gauging how well models translate visuals into language. However, it mainly assesses linguistic similarity and might

miss semantic understanding of image content. Complementary evaluation methods and qualitative assessments are needed for a thorough understanding of model performance.

```
# prepare tokenizer
tokenizer = create_tokenizer(train_descriptions)
vocab_size = len(tokenizer.word_index) + 1
print('Vocabulary Size: %d' % vocab_size)
# determine the maximum sequence length
max_length = max_length(train_descriptions)
print('Description Length: %d' % max_length)

# prepare test set

# load test set
filename = 'Flickr_8k.testImages.txt'
test = load_set(filename)
print('Dataset: %d' % len(test))
# descriptions
test_descriptions = load_clean_descriptions('descriptions.txt', test)
print('Descriptions: test=%d' % len(test_descriptions))
# photo features
test_features = load_photo_features('features.pkl', test)
print('Photos: test=%d' % len(test_features))

# load the model which has minimum loss, in this case it was model_18
filename = 'model_18.h5'
model = load_model(filename)
# evaluate model
evaluate_model(model, test_descriptions, test_features, tokenizer, max_length)
```

```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.529691
BLEU-2: 0.277107
BLEU-3: 0.182869
BLEU-4: 0.080318
```


```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.529691
BLEU-2: 0.277107
BLEU-3: 0.182869
BLEU-4: 0.080318
```


Fig 7.4 - Screenshot of the BLEU score (Accuracy) of ML model

Current BLEU score of the DetectWave is 0.5296 which is pretty good as it is closer to 1 .

7.3. Input Parameters / Features considered

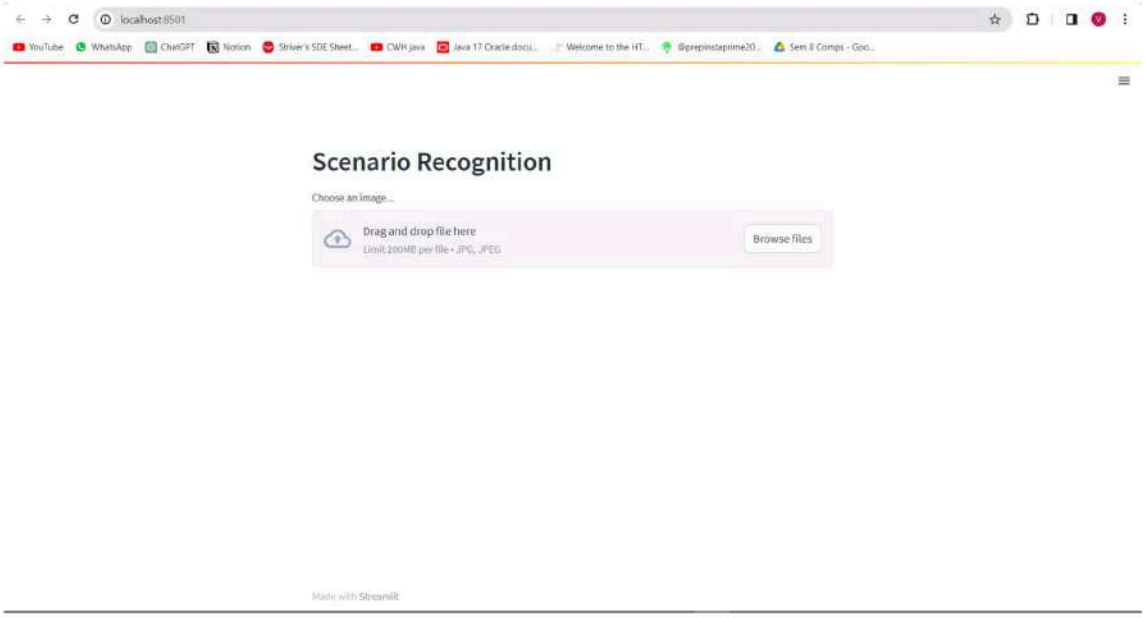

Fig 7.5 - Input Parameter for the DetectWave System.

In fig 7.5 User can upload the image from which we have to generate scenarios in text as well as audio form. This interface is developed using streamlit and at backend python is used.

7.4 Comparison of Results with Existing System -

Other System	Our System
Only can generate captions from Image.	Generates captions as well as scenarios depicted in Image and also convert it into audio.
Accuracy is not precise.	Accuracy is precise.
Generated caption is not able to explain the entire scenario to the user.	Generated scenario explains the entire description of the image to the user.
Cannot be used by visually impaired individuals.	It can be used by visually impaired persons as the scenario is converted from text form into audio form which is more easily accessible by the visually impaired individuals.

7.5. Inference drawn

Existing systems only generate captions from Image. They do not explain the entire context and description of the uploaded Image. So while developing our system we had given our best to get the entire context and scenario of the uploaded image once it is uploaded to our system. And also it should be convenient to the visually impaired users to use our system. We had given the feature of audio generation of the scenario of the image.

Chapter 8 - Conclusion

8.1. Limitations

- For now the system is not integrated to the mobile application so it would be difficult for the user to practically use it.
- Sometimes the scenario is not predicted accurately.
- Sometimes it takes more time to generate the scenario from Image.

8.2. Conclusion

- The primary objective of this initiative is scenario recognition from images, employing advanced technologies like RNN, CNN and the GPT-3.5-turbo model for scene interpretation.

- Extensive training on the Flickr 8k dataset ensures the system's proficiency in generating precise captions and converting scene descriptions into audio format for improved accessibility.

- By accurately recognizing scenarios from images and converting them into accessible audio, the project aims to bridge the gap between visual information and individuals with visual impairment.

8.3. Future Scope

a) By using this project we can help blind people to recognize the image as well as understand the scenario depicted in the image.

b) We will create a mobile application in which blind people can scan the images using opencv as well as the RNN and CNN model integrated with the mobile app will help us to extract the features from images and generate the caption.

c) This caption is used to generate the scenario of the image which is in text form and later it is converted into audio format.

References

1. Hedi Harzallah, François Jurie, and Cordelia Schmid, "Scene Classification with Invariant Descriptor", International Conference on Computer Vision (ICCV), 2005.

2. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", Computer Vision and Pattern Recognition (CVPR), 2006.

3. Abhinav Gupta, Alexei A. Efros, and Martial Hebert, "Learning Spatial Context: Using Stuff to Find Things", Computer Vision and Pattern Recognition (CVPR), 2010.

4. Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P. Xing, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Scarification", Advances in Neural Information Processing Systems (NeurIPS), 2010.

5. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017.

6. Zhen Zhu, Shuhui Wang, and Baocai Yin, "Large-Scale Scene Understanding Challenge 2016: Scene Classification and Detection from High-Resolution Remote Sensing Imagery", IEEE Transactions on Geoscience and Remote Sensing, 2017.

7. Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei, "Scene Graph Generation from Objects, Phrases and Region Captions", Computer Vision and Pattern Recognition (CVPR), 2017.

8. Justin Johnson, Agrim Gupta, Li Fei-Fei, "Neural Scene Graphs for Interpretable Scene Generation", Advances in Neural Information Processing Systems (NeurIPS), 2018.

9. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, arXiv:2004.10934

10. W. Ke, T. Zhang, Z. Huang, Q. Ye, J. Liu, and D. Huang, "Multiple anchor learning for visual object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10203–10212, doi: 10.1109/CVPR42600.2020.01022

11. H. Li, Z. Wu, C. Zhu, C. Xiong, R. Socher, and L. S. Davis, "Learning from noisy anchors for one-stage object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10585–10594, doi: 10.1109/CVPR42600.2020.01060.

12. W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," Int. J. Comput. Vis., vol. 128, pp. 261–318, 2020, doi: 10.1007/s11263- 019-01247-4.

13. B. Amjoud and M. Amrouch, "Transfer learning for automatic image orientation detection using deep learning and logistic regression," IEEE Access, vol. 10, pp. 128543–128553, 2022, doi: 10.1109/ACCESS.2022.3225455.

14. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. FeiFei, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015- 0816-y.

15. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in Proc. 31st AAAI Conf. Artif. Intell., San Francisco, CA, USA, Feb. 2017, pp. 4278–4284.

Appendix

1. Paper I details -

a) Paper I -

DetectWave - Scenario recognition from image

Mannat Doultani
Department of Computer Engineering(Vivekanand Education Society's Institute of Technology)
Mumbai, India
pooja.nagdev@ves.ac.in

Harsh Loya
Department of Computer Engineering(Vivekanand Education Society's Institute of Technology)
Mumbai, India
2020.harsh.loya@ves.ac.in

Gautam Dinga
Department of Computer Engineering(Vivekanand Education Society's Institute of Technology)
Mumbai, India
2020.gautam.dinga@ves.ac.in

Harsh Patil
Department of Computer Engineering(Vivekanand Education Society's Institute of Technology)
Mumbai, India
2020.harsh.patil@ves.ac.in

Viraj Joshi
Department of Computer Engineering(Vivekanand Education Society's Institute of Technology)
Mumbai, India
2020.viraj.joshi@ves.ac.in

***Abstract*—Our system utilizes the Faster R-CNN model to automatically detect and recognize objects and scenes in uploaded images. With a focus on precision, it efficiently identifies a variety of objects, including those in garden settings, while also recognizing scenes such as parks. This dual capability provides valuable context to users, facilitating not only object comprehension but also a broader understanding of the image. Results are presented in a clear format, making it accessible for applications such as content tagging, automated image analysis, and aiding individuals with visual impairments. This intuitive and versatile tool streamlines image understanding, catering to both personal and professional needs.**

Keywords—Machine Learning model, Scenario recognition, Detect Wave, Object Detection, Summarization of image, Text generation, R-CNN, Natural Language Processing, Technological advancement

I. Introduction

In the contemporary digital landscape, the ability to interpret and comprehend the content of images has emerged as a pivotal frontier in artificial intelligence (AI). Within this expansive field, scenario recognition from images stands out as a fascinating and complex domain of computer vision. Its goal is to empower machines with the capability to categorize and understand scenes or scenarios depicted in visual data, thereby bridging the chasm between human perception and the capabilities of artificial intelligence.

This report encapsulates a comprehensive exploration and implementation of an advanced system for scenario recognition from images. The project's foundation lies in addressing the existing challenges associated with contemporary scenario recognition systems, which often struggle to grasp the broader context of an image. There is a recognized semantic gap between the labels or categories utilized for recognition and the nuanced meaning of real-world scenarios. Moreover, many current models are confined by the datasets on which they are trained, limiting their adaptability to novel or dynamic situations.

The primary objective of this project is to pioneer the development of an automated system capable of categorizing diverse scenarios with precision, drawing insights from the content of images. These scenarios span a spectrum, encompassing everyday activities, complex events, outdoor and indoor settings, emergencies, leisure, transportation, and more. The potential applications of accurate scenario recognition are far-reaching and extend across various industries, including but not limited to autonomous vehicles, real-time surveillance, and smart environments.

The journey towards achieving this objective unfolds through a meticulously planned sequence of steps. Beginning with the crucial phase of data collection and preprocessing, a diverse dataset is curated, enriched with accurate scenario labels, object details, activities, and contextual information. This dataset serves as the bedrock for training the machine learning model, with an emphasis on capturing the intricacies of real-world scenarios.

Feature extraction, a pivotal stage in the process, leverages advanced Convolutional Neural Network (CNN) architectures pre-trained on large image datasets, such as ImageNet. This not only facilitates the extraction of meaningful visual features from images but also allows for the adaptation of pre-trained CNNs to the specific task of scenario recognition through fine-tuning on the target dataset.

Contextual understanding is then achieved through the implementation of object detection models, such as YOLO and Faster R-CNN, coupled with semantic segmentation techniques. This amalgamation empowers the system to not only identify and locate objects within images but also to discern their roles and relationships, thereby enhancing overall semantic understanding.

The report further develops into the intricacies of semantic inference, where the results of semantic segmentation and object detection are integrated. Additionally, attribute recognition is implemented to infer attributes and functions

of objects, contributing to a more nuanced and comprehensive understanding of the depicted scenarios.

As the project unfolds, it incorporates a continuous improvement cycle, incorporating user feedback, data enhancement, and ongoing model training. Ethical considerations and real-time processing optimizations add depth to the project's framework, ensuring that the system not only excels in technical aspects but also adheres to ethical standards and practical deployment requirements.

This report encapsulates a journey through the realms of computer vision, machine learning, and artificial intelligence, with the overarching goal of contributing to the advancement of scenario recognition systems. The subsequent sections delve into the methodological intricacies, model development, user interface design, deployment strategies, and extensive testing that collectively constitute this ambitious and impactful project.

Figure 1.Scenario recognition

II. Problem Statement and Objective

A. *Problem Statement*

The problem landscape in contemporary scenario recognition systems is multifaceted and hinges on the challenges that impede the seamless and nuanced interpretation of visual data by artificial intelligence. One of the primary challenges is the existence of a semantic gap, a disconnect between the labels or categories used for recognition and the intricate, real-world meaning of diverse scenarios. In essence, existing systems often struggle to go beyond basic object recognition and fail to capture the holistic context within which these objects exist. This semantic gap poses a significant hindrance to the development of scenario recognition systems that can truly mirror human-like comprehension. While objects within an image can be accurately labelled, understanding the relationships, activities, and environmental nuances that define a scenario remains a formidable challenge. Another critical problem lies in the limitations imposed by the datasets on which these models are trained. Many current models are confined to datasets that may lack the diversity and complexity inherent in real-world scenarios. As a result, these systems may encounter difficulties when faced with novel or dynamic situations that deviate from their training data.The adaptability of scenario recognition systems to a rapidly changing and diverse environment is a core problem. Real-world scenarios are inherently dynamic and can involve an evolving set of objects, activities, and relationships. Existing models may struggle to adapt to these changes, leading to a lack of accuracy and reliability in recognizing scenarios outside their training scope. Furthermore, the problem extends to the need for systems capable of handling a broad spectrum of scenarios, ranging from routine activities to complex events, emergencies, and various indoor and outdoor settings. Many current models may excel in specific domains but lack the versatility required for comprehensive scenario recognition. Addressing these problems is pivotal for the development of scenario recognition systems that can truly bridge the gap between human perception and artificial intelligence. Overcoming the semantic gap, enhancing adaptability to novel situations, and achieving a comprehensive understanding of diverse scenarios constitute the key problem definitions that this project aims to tackle. By doing so, the project seeks to contribute significantly to the advancement of scenario recognition systems and their applications across diverse industries.

B. *Objectives:*

Context Understanding: Recognizing and understanding the context or situation in a given environment or scenario.

Event Detection: Identifying specific events or occurrences within a scenario, such as accidents, celebrations, emergencies, or routine activities.

Anomaly Detection: Detecting abnormal or unexpected behaviour within a scenario, which may indicate potential issues or threats.

Decision Support: Providing information and insights to support decision-making based on the recognized scenarios.

Automation: Enabling automated systems to respond or adapt to different scenarios, enhancing efficiency and reducing the need for manual intervention.

Security Enhancement: Enhancing security measures by recognizing and responding to potential threats or suspicious activities within a given scenario.

Optimizing Resource Allocation: Utilizing scenario recognition to optimize the allocation of resources, whether in a business, industrial, or public setting.

User Experience Improvement: Improving user experience by tailoring systems or services based on the recognized scenarios, creating a more personalized and efficient interaction.

Predictive Analysis: Using historical data and scenario recognition to predict future events or trends within a particular environment.

Integration with Other Systems: Integrating scenario recognition with other systems or technologies to create a comprehensive and interconnected approach to understanding and responding to different situations.

Monitoring and Surveillance: Employing scenario recognition for continuous monitoring and surveillance, especially in areas where real-time awareness is crucial.

Compliance and Regulation: Ensuring compliance with regulations and standards by recognizing and addressing specific scenarios that may impact legal or ethical considerations.

III. Related Work

The relevance of the project on advanced scenario recognition from images is underscored by its potential to revolutionize the landscape of artificial intelligence and computer vision applications. In the contemporary era, where visual data plays an increasingly vital role, the ability to accurately and comprehensively recognize diverse scenarios holds immense significance across various domains.

Enhanced Precision in AI Systems:
The project's primary goal is to pioneer the development of an automated system capable of categorizing scenarios with precision. This enhanced precision is crucial for AI systems, ensuring that they can move beyond simplistic object recognition and delve into a deeper understanding of the relationships, activities, and environmental context within an image,

Versatility in Applications:
Accurate scenario recognition has far-reaching implications across industries. From autonomous vehicles navigating complex traffic scenarios to real-time surveillance systems identifying anomalous activities, the versatility of applications is vast. This relevance positions the project at the forefront of technology with the potential to impact numerous sectors.

Real-World Adaptability:
The project directly addresses the limitations of existing models in adapting to novel or dynamic situations. The real-world adaptability achieved through this research is crucial for systems deployed in ever-changing environments, ensuring that they remain effective and reliable in scenarios beyond their initial training data.

Holistic Understanding of Scenarios:
By emphasizing not just object recognition but also understanding the nuanced relationships and activities within a scenario, the project contributes to a more holistic comprehension of visual data. This is particularly relevant in applications where context matters, such as in smart environments or emergency response systems.

Ethical Considerations and Responsible AI:
The relevance extends beyond technical advancements to ethical considerations. The project framework incorporates ethical standards, ensuring that the resulting systems align with responsible AI practices. This ethical dimension is crucial as AI technologies increasingly integrate into society, emphasizing the project's responsible deployment.

Impact on Various Industries:
The potential impact of accurate scenario recognition spans across industries such as transportation, surveillance, and smart environments. It positions the project as not only technologically relevant but also as a catalyst for advancements in diverse sectors, contributing to the broader evolution of AI applications.

IV. Proposed Methodology

The methodology employed for the scenario detection project involves a systematic and multi-phase approach. The project begins with a crucial phase of data collection and preprocessing, where a diverse dataset is meticulously curated. This dataset is enriched with accurate scenario labels, object details, activities, and contextual information to create a comprehensive foundation for training the machine learning model.

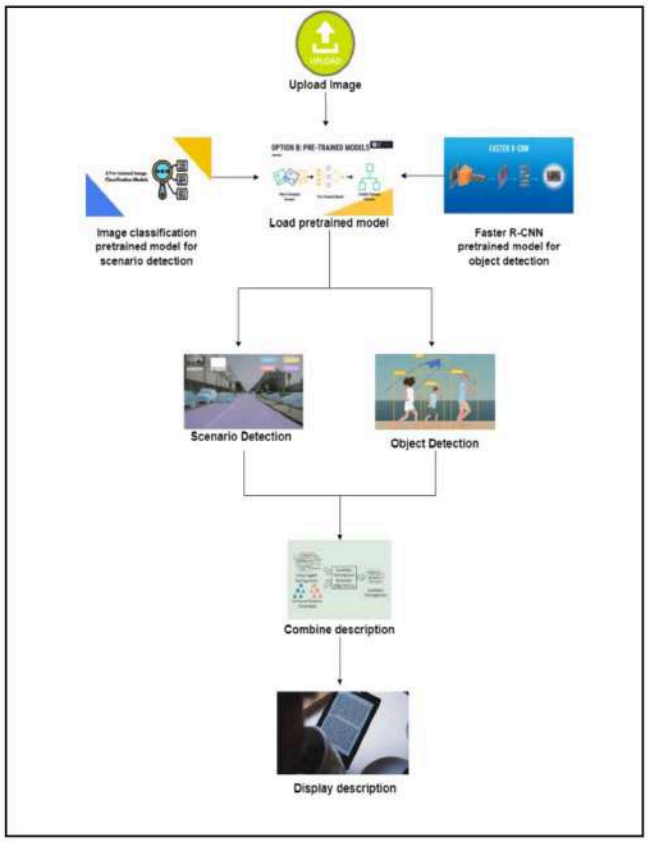

Figure 2: Block Diagram

Feature extraction constitutes a pivotal stage, leveraging advanced Convolutional Neural Network (CNN) architectures pre-trained on large image datasets like ImageNet. The model is fine-tuned on the target dataset to capture the intricacies of real-world scenarios. Contextual understanding is achieved through the implementation of object detection models, Faster R-CNN, along with semantic segmentation techniques. This combination enables the system not only to identify and locate objects within images but also to discern their roles and relationships, enhancing overall semantic understanding.

Semantic inference integrates the results of semantic segmentation and object detection, while attribute recognition is implemented to infer attributes and functions of objects, contributing to a more nuanced understanding of depicted scenarios. The project incorporates a continuous improvement cycle, involving user feedback, data enhancement, and ongoing model training. Ethical considerations and real-time processing optimizations are integrated into the framework to ensure both technical excellence and adherence to ethical standards in practical deployment. The methodology thus provides a comprehensive and iterative strategy for advancing scenario detection from images.

A. Modular Diagram:

The main modules of this system include the customer-side and seller-side, which have been described in the block diagram.

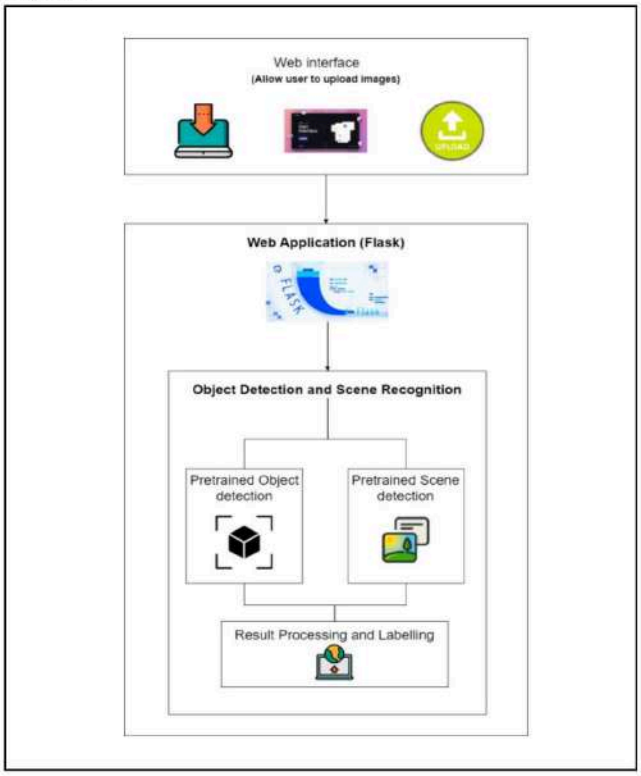

Figure 3: Modular Diagram

User Input and Image Upload: The code does not include a direct user interface butassumes that users will upload an image (e.g., "abc.jpg") to the script.

Pre-trained Object Detection Model (TensorFlow Hub): The script loads a pre-trained object detection model from TensorFlow Hub. The model is responsible for detecting objects in the uploaded image.

Image Preprocessing: The uploaded image is read and preprocessed, converting it to a format suitable for the object detection model.

Object Detection: The pre-trained object detection model processes the preprocessed image and returns information about detected objects, including their bounding boxes, labels, and confidence scores.

Scene Recognition (Hugging Face Transformers): The script also loads a pre-trained image classification model for scene recognition using Hugging Face Transformers. It recognizes the overall scene depicted in the uploaded image.

Scene and Object Labels: The script defines dictionaries for scene labels and object labels that correspond to what's commonly found in a garden scene.

Results and Combined Description: The script processes the object detection and scene recognition results, filtering out low-confidence detections, converting labels to text descriptions, and generating a combined description of the scene.

Output: The final combined description, including the recognized scene and detected objects, is printed to the console.

B. Algorithm and Process Design akes

Proposed algorithms for a project like "Detect Wave: Scenario Recognition from Image" are the core techniques and methods used to recognize scenarios or situations depicted in images. These algorithms serve as the backbone of your system. Here are some commonly used algorithms and techniques for scenario recognition from images:

CNN (Convolutional Neural Network): R-CNN, or Region-based Convolutional Neural Network, is an early approach to object detection in images. It operates in a multi-step process where it first generates a multitude of region proposals or candidate bounding boxes that may contain objects of interest in the image. These regions are then individually processed by a deep convolutional neural network (CNN) to extract feature vectors. The extracted features are subsequently used for object classification and refinement of bounding box coordinates. R-CNN's training occurs in two steps: the CNN is pre-trained on a large dataset for feature extraction, and a linear Support VectorMachine (SVM) is trained for object classification and bounding box adjustment. Despite its contributions, R-CNN has limitations, notably its slow processing speed due to handling regions independently and the lack of end-to-end training.

Faster R-CNN (Region Convolutional Neural Network) Faster R-CNN is an object detection model that uses a Region Proposal Network (RPN) to identify potential object locations in an image. It consists of two stages:

Region Proposal: The RPN generates candidate bounding boxes where objects might be.

Object Classification and Refinement: Fast R-CNN refines these boxes and classifies the objects within

them.

This two-stage approach significantly improved object detection speed and accuracy, making it a key technology for applications like autonomous driving and surveillance.

V. RESULTS AND EVALUATION

Object Detection:

Result: The model can identify and locate objects within the uploaded image, providing information such as bounding boxes, object labels, and confidence scores.

Decision: This component can be useful in applications where you need to automatically detect and locate objects within a garden scene. For example, it can identify trees, flowers, tools, or other garden-related itemproducts available. Customers can login and see different products across the city and order.

Scene Recognition:

Result: The scene recognition model analyses the overall content of the image and provides a label that represents the type of scene, such as "garden," "botanical garden," or "flower garden."

Decision: This part of the model is valuable when you want to categorize images based on the scene they depict. It's particularly useful for understanding the context in which the detected objects areplaced.

Label Mapping:

Result: The script maps object and scene labels to human-readable descriptions, making it easier to understand the content of the image.

Decision: Label mapping enhances the user-friendliness of the results, providing clear anddescriptive labels for the objects and scenes detected in the image.

Combined Description:

Result: The script generates a combined description that incorporates both the recognized scene and the detected objects.

Decision: This combined description is user-friendly and informative, presenting a textual summary of what's present in the garden scene. It's particularly useful when you want to provide a detailed description of the image content

Flow of the web - application:

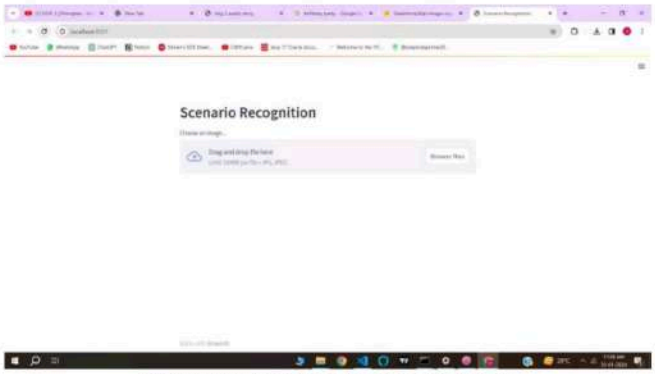

Figure 4: Interface to upload Image

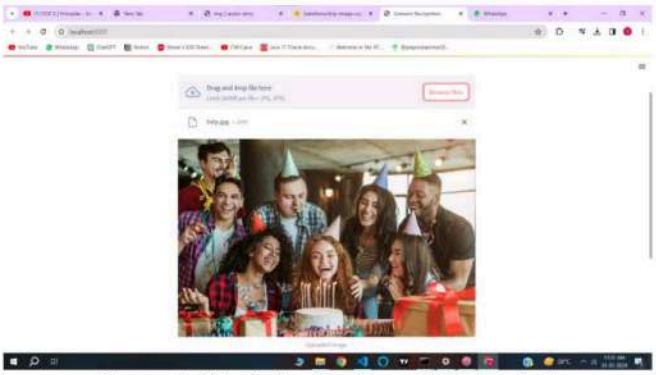

Figure 5: The following image is uploaded.

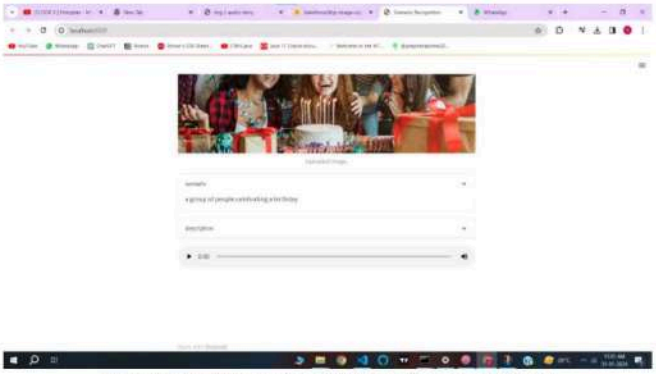

Figure 6: Scenario of image is recognized

Figure 7: Description of image is given in text form

Figure 8: Description of image is given in audio form

VI. Conclusion

This project represents a significant stride in the field of scenario detection from images, aiming to bridge the semantic gap and address limitations in existing systems. Leveraging advanced Convolutional Neural Network architectures facilitates robust feature extraction, capturing the intricacies of real-world scenarios. The integration of object detection models and semantic segmentation techniques empowers the system not only to identify objects but also to understand their roles and relationships, achieving a heightened level of contextual understanding. The inclusion of semantic inference and attribute recognition further contributes to nuanced scenario comprehension. Crucially, the project incorporates a continuous improvement cycle, emphasizing user feedback, data enhancement, and ongoing model training. The acknowledgment of ethical considerations and real-time processing optimizations demonstrates a commitment to both technical excellence and responsible AI deployment. As this report encapsulates a journey through the realms of computer vision, machine learning, and artificial intelligence, it stands as a testament to the ambition of contributing to the advancement of scenario recognition systems. The meticulous methodology, user interface design, deployment strategies, and extensive testing collectively constitute a robust and impactful project poised to make valuable contributions to the ever-evolving landscape of ML-driven scenariodetection from images.

References

1. "Scene Classification with Invariant Descriptor"
- Authors: Hedi Harzallah, François Jurie, and Cordelia Schmid
- Published in: International Conference on Computer Vision (ICCV), 2005
- This paper discusses the use of invariant local features for scene classification, a fundamental aspect of scenario recognition

2. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories"
- Authors: Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce
- Published in: Computer Vision and Pattern Recognition (CVPR), 2006
- Introduces the Spatial Pyramid Matching (SPM) method for scene classification, providing a more detailed representation of the spatial layout of features.

3. "Learning Spatial Context: Using Stuff to Find Things"
- Authors: Abhinav Gupta, Alexei A. Efros, and Martial Hebert
- Published in: Computer Vision and Pattern Recognition (CVPR), 2010
- Discusses the importance of incorporating spatial context in scene recognition and proposes methods to learn contextual information from images.

4. "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Scarification"
- Authors: Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P. Xing
- Published in: Advances in Neural Information Processing Systems (NeurIPS), 2010.
- Introduces the concept of an "Object Bank" for scene classification, providing a high-level representation of scenes based on the presence of object detectors.

5. "Places: A 10 Million Image Database for Scene Recognition"
- Authors: B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba
- Published in: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017
- Describes the Places database, a large-scale dataset for scene recognition, and provides insights into training deep neural networks for this task.

6. "Large-Scale Scene Understanding Challenge 2016: Scene Classification and Detection from High-Resolution Remote Sensing Imagery"
- Authors: Zhen Zhu, Shuhui Wang, and Baocai Yin
- Published in: IEEE Transactions on Geoscience and Remote Sensing, 2017.
- Discusses challenges and advancements in scene understanding, particularly in the context of high-resolution remote sensing imagery.

7. "Scene Graph Generation from Objects, Phrases and Region Captions"
- Authors: Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei
- Published in: Computer Vision and Pattern Recognition (CVPR), 2017
- Explores the generation of scene graphs from images, capturing relationships between objects and providing a more structured representation of scenes.

8. "Neural Scene Graphs for Interpretable Scene Generation"
- Authors: Justin Johnson, Agrim Gupta, Li Fei-Fei
- Published in: Advances in Neural Information Processing Systems (NeurIPS), 2018 .
- Introduces the concept of Neural Scene Graphs, providing a way to represent scenes hierarchically, capturing objects and their relationships.

b)Plagiarism Report

DetectWave - Scenario recognition from image

ORIGINALITY REPORT

6%	5%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.asianssr.org Internet Source	3%
2	www.mdpi.com Internet Source	1%
3	dokumen.pub Internet Source	1%
4	www.ijert.org Internet Source	<1%
5	Submitted to University of Northumbria at Newcastle Student Paper	<1%
6	summit.sfu.ca Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

c) Project review sheet

Project review sheet 1 -

Inhouse/ Industry _Innovation/Research:

Class: D17 A/~~B/C~~

Sustainable Goal: 7, 9, 10

Project Evaluation Sheet 2023 - 24

Group No.: 34

Title of Project: Detect Wave - A Scenario recognition from Image.

Group Members: Harsh Loya (36 | D17A), Viraj Joshi (26 | D17A), Gautam Dinga (14/D17A), Harsh Patil (53 | D17A).

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
3	3	3	2	4	1	1	1	1	1	2	2	2	2	3	31

Comments: Object should be detected by hovering on Image, Self ~~Des~~ Model to be developed, Object Detection.

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
3	3	3	2	3	1	1	1	1	1	2	2	2	2	3	30

Comments:

Date: 10th february, 2024

Name & Signature Reviewer 2

Project review sheet 2 -

34

Project Evaluation Sheet 2023 - 24

Title of Project: Scenario recognition from image - detect wave

Group Members: Harsh Loya (36 | D17A), Gautam Dinga (14/D17A), Viraj Joshi (26/D17A), Harsh Patil (53 /D17A)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	4	1	1	1	1	1	2	2	2	2	3	34

Comments:

Name & Signature Reviewer1

Inhouse/ Industry _Innovation/Research:

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	4	1	1	1	1	1	2	2	2	2	3	34

Comments: Do not use pregenerated model.

Date: 9th March, 2024

Name & Signature Reviewer 2

2) Competition Certificates -

ICUS 2024 Conference Presentation Certificate -

3rd International Conference of UG Students

ICUS 2024

Organized By
Soft Computing Research Society
March 01, 2024

Certificate of Presentation

This is to certify that **Harsh Loya** has presented the paper titled **Detectwave - Scenario Recognition from Image** authored by **Mannat Doultani, Gautam Dinga, Harsh Loya, Viraj Joshi, Harsh Patil** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma
(Organizing Chair, ICUS 2024)

3rd International Conference of UG Students

ICUS 2024

Organized By
Soft Computing Research Society
March 01, 2024

Certificate of Presentation

This is to certify that **Viraj Joshi** has presented the paper titled **Detectwave - Scenario Recognition from Image** authored by **Mannat Doultani, Gautam Dinga, Harsh Loya, Viraj Joshi, Harsh Patil** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma
(Organizing Chair, ICUS 2024)

3rd International Conference of UG Students

ICUS 2024

Organized By
Soft Computing Research Society
March 01, 2024

Certificate of Presentation

This is to certify that **Gautam Dinga** has presented the paper titled **Detectwave - Scenario Recognition from Image** authored by **Mannat Doultani, Gautam Dinga, Harsh Loya, Viraj Joshi, Harsh Patil** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma
(Organizing Chair, ICUS 2024)

3rd International Conference of UG Students

ICUS 2024

Organized By
Soft Computing Research Society
March 01, 2024

Certificate of Presentation

This is to certify that **Harsh Patil** has presented the paper titled **Detectwave - Scenario Recognition from Image** authored by **Mannat Doultani, Gautam Dinga, Harsh Loya, Viraj Joshi, Harsh Patil** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma
(Organizing Chair, ICUS 2024)