Faculty of Computers &
Artificial Intelligence

Benha University

AI-Generated Media Detection System

Program Name	
□ Computer Science □ Information Systems	□ Scientific Computing ☑ Artificial Intelligence

English Title	AI-Generated Media Detection System
Arabic Title	نظام الكشف عن الوسائط المولدة بالذكاء الاصطناعي

Project by

1. Romani Nasrat Shawqi Gerges
2. Ahmed Mohamed Ali Abo el-Kassem
3. Zeyad Elsayed Abdel-Azim Ali
4. Sara Reda Moatamed Hassan Eissa
5. Reham Moustafa Ali Abdel-Moati
6. Rawan Abdel-Aziz Ahmed Mahmoud
7. Abd-Allah Mohamed Abdel-monem
8. Mohannad Ayman salah Abdel-fattah
9. Mohamed Abd-Allah Abdel-salam Abdel-Dayem

Supervised by

Dr. Eman Abdel-Latef

Eng. Sahar Mostafa

Abstract

In the dynamic landscape of technological progress, the advent of advanced deep learning techniques, notably Generative Adversarial Networks (GANs) and Variational Auto-encoders (VAEs), has propelled artificial intelligence (AI) content into a realm of heightened credibility and realism. Among these innovations, the emergence of deepfake technology stands out as a powerful tool capable of crafting remarkably authentic synthetic content.

Deepfake, at the crossroads of innovation and deception, opens doors to unprecedented applications in diverse industries like film production, creative arts, and advertising. However, its shadow looms large over Multimedia Information Retrieval Systems (MIPR), challenging facial and speech recognition systems and amplifying the risk of disseminating misleading information.

This project addresses the implications of deepfake by introducing a comprehensive model designed to discern the origin of multimedia content, be it text or images. Operating as a vigilant guardian, the project employs cutting-edge algorithms to unveil the subtle intricacies of manipulated visuals, safeguarding against deceptive imagery. Moreover, it navigates the ambiguous boundaries between human-authored and AI-generated content, offering transparency in an era where the authenticity of digital media is increasingly vital.

With a focus on both image and text, the project stands as a resilient shield against the threats posed by deepfake technology. Beyond protecting facial and speech recognition systems, its scope extends to safeguarding the societal fabric from the perils of misleading information. By championing transparency, authenticity, and ethical use of AI, this project envisions a future where the authenticity of multimedia content prevails, ensuring a robust and trustworthy digital ecosystem.

Table of Content

1. Chapter One :Introduction…………………………………………………
- 1.1 Problem Definition…………………………………………………
- 1.2 Problem Solution…………………………………………………
- 1.3 Project Objective…………………………………………………
- 1.4 Stakeholder List…………………………………………………
- 1.5 Proposed Scope…………………………………………………
- 1.6 Project Constraints…………………………………………………

2. Chapter Two:System Analysis And Design
- 2.1 User and System Requirements……………………………
  - 2.1.1 Functional Requirements……………………………
  - 2.1.2 Non-functional Requirements……………………
- 2.2 System Design
  - 2.2.1 Use Case Diagram…………………………………………………
    - 2.2.1.1 Use Case Definitions…………………………………………
  - 2.2.2 Activity Diagram…………………………………………………
    - 2.2.2.1 Sign-Up Activity…………………………………………
    - 2.2.2.2 Sign-In Activity…………………………………………
    - 2.2.2.3 Select Media Type Activity……………………
    - 2.2.2.4 Upload Media Activity…………………………
    - 2.2.2.5 Detect AI-Generated Media Activity……
    - 2.2.2.6 Feedback Activity…………………..…………
    - 2.2.2.7 Subscription Activity……………………………………
    - 2.2.2.8 Admin Management Activity……………..
    - 2.2.2.9 Show History Activity…………………..
    - 2.2.2.10 Edit Profile Activity………………………………..
  - 2.2.3 Context Diagram…………………………………….
  - 2.2.4 Data Flow Diagram

Table of Content

- 2.2.4.1 Level 0…………………………………….
- 2.2.4.2 Level 1………………………………
- 2.2.5 Sequence Diagram
- 2.2.5.1 User……………………………………………………..
- 2.2.5.2 Admin…………………………………………………..
- 2.2.6 Class Diagram……………………………………………
- 2.2.7 Database Design…………………………………………………
- 2.2.8 ERD Diagram…………………………………………………

Table of Content

LIST OF FIGURES

Figure 1: Use Case ……………………………………………..
Figure 2: Activity Diagram – Sign-Up…………………………
Figure 3: Activity Diagram – Sign In…………………………
Figure 4: Activity Diagram – Select Media Type………
Figure 5: Activity Diagram – Upload Media……………………
Figure 6: Activity Diagram – Detect AI-Generated Media
Figure 7: Activity Diagram – Feedback………………
Figure 8: Activity Diagram – Subscription………………….
Figure 9: Activity Diagram – Admin Management……
Figure 10: Activity Diagram – Show History…………
Figure 11: Activity Diagram – Edit Profile…………
Figure 12: Context Diagram…………….
Figure 13: Data Flow Diagram – Level 0…………………
Figure 14: Data Flow Diagram – Level 1………………
Figure 15: Sequence Diagram-User…………………………….
Figure 16: Sequence Diagram-Admin…………………………….
Figure 17: Class Diagram…………………………………….
Figure 18: Database Design…………………………..
Figure 19: ERD Diagram………………………………….

ChapterOne

1. Introduction

In an era dominated by advanced artificial intelligence, the line between authentic and AI-generated content has blurred, raising concerns about misinformation and deepfake threats. Our project, the AI-Generated Media Detection System, addresses these challenges by introducing a cutting-edge model to discern between content created by humans and that produced by AI. Focused on images and text, our web-based platform aims to empower users to verify the authenticity of media content in an increasingly complex digital landscape. Join us in building a tool that navigates the evolving realm of AI-generated content, ensuring trust and reliability in digital interactions

===

1.1 Problem Definition

In the rapidly evolving landscape of artificial intelligence, tech giants like OpenAI, Meta, Microsoft, and Google engage in a competitive race, focusing on groundbreaking technologies such as transformers, Generative Adversarial Networks (GANs), and Large Language Models (LLMs). This fervent competition stems from the realization that AI is a pivotal force shaping the future.

With the emergence of advanced AI, particularly GANs and Transformers, there is a remarkable stride in creating lifelike images. However, this progress brings forth a pressing challenge – the increasing difficulty in distinguishing AI-generated images from real ones. This predicament gives rise to issues like fake profiles, scams, and the dissemination of deceptive information, reminiscent of challenges posed by fake news.

Large Language Models (LLMs) contribute to the predicament by generating text that closely mirrors human language. This poses a unique challenge, especially in educational settings where discerning between authentic student work and AI-generated content becomes intricate.

As open-source technologies advance, the line between machine-generated and human-created text and images blurs, paving the way for malicious actors to produce convincing fake content. Hence, the need for robust systems becomes paramount to discern whether content originates from a machine or a human, preserving trust in online interactions

1.2. Problem Solution

In response to the challenges posed by advanced AI technologies, particularly in the generation of deceptive content such as deepfakes, our solution is a comprehensive and accessible online platform. We envision constructing a sophisticated website housing specialized detector models designed to address distinct facets of the issue. Our array of detector models includes:

1. AI-Generated Image Detection Model:
   - This model is finely tuned to discern images that have been generated using advanced AI techniques, with a particular focus on identifying content produced by Generative Adversarial Networks (GANs).

2. Manipulation Detection Model:
   - Specifically engineered to identify images whose features have been manipulated by AI, this model aims to expose any alterations or distortions introduced to deceive the viewer.

3. AI-Generated Text Detection Model:
   - In addition to visual content, our platform incorporates a cutting-edge model adept at identifying text generated by AI. This is crucial for distinguishing between human-crafted narratives and those produced by large language models (LLMs).

--Web Accessibility for All:
To ensure widespread utility and user-friendliness, we have chosen to implement these detector models within the framework of an intuitive website. This strategic decision is driven by our commitment to making advanced AI detection technology accessible to users across diverse backgrounds and skill levels.

- Key Features of the Website:

- User-Friendly Interface: The website is designed with simplicity in mind, allowing users to easily navigate and utilize the detection tools without the need for specialized technical knowledge.

- Multi-Model Integration:Our platform seamlessly integrates multiple detector models, offering a holistic solution that addresses the nuanced challenges presented by AI-generated images and text.

- Upload and Analyze Functionality:Users can effortlessly upload images and text for analysis, receiving detailed insights generated by our detector models.

- Transparent Reporting:The analysis results provided by the models are presented in a clear and understandable format, promoting transparency and fostering user trust in the authenticity of the content.

-Empowering Users Against AI Deception:
By consolidating these detector models into a user-centric website, we aspire to empower individuals, content creators, and organizations with a robust defense mechanism against the threats posed by AI-induced deception. Our solution not only detects and differentiates AI-generated content but also contributes to the broader narrative of fostering transparency and trust in the digital landscape.

1.3 Project Objective

To develop a website utilizing machine learning and data processing techniques to detect images and texts generated by AI, aiming to

1. **Identify AI-Generated Content**: Distinguish images and texts generated by AI models, such as those produced by neural networks.

2. **Differentiate Human-Created Content from AI**: Differentiate natural human-generated content from AI-generated content to verify credibility and authenticity.

3. **Provide a User-Friendly Interface**: Offer an intuitive user interface for users to upload images and texts for analysis, presenting analysis results in an easily understandable format.

4. **Develop Machine Learning Models**: Construct accurate and efficient machine learning models capable of identifying and categorizing suspicious images and texts.

1.4 Stakeholder List

- User: Utilizes the website to detect AI-generated images and texts.
- Admin: Manages user accounts and profiles

1.5 Proposed Scope

1. Key Features:

- AI-Generated Image Detection: Identifying images created by AI and distinguishing them from human-created images.
- AI-Generated Text Detection: Identifying text generated by AI and discerning it from human-generated text.

2.Core Requirements:

- ✓ **Machine Learning Model Development**: Building ML models capable of differentiating AI-generated images and texts from human-created ones.
- ✓ **Data Collection** and Categorization: Gathering a substantial dataset with examples of AI-generated and human-generated images and texts for model training.
- ✓ **User Interface** Development: Creating a user-friendly interface for users to upload images or texts for analysis.

3. Scope Exclusions:

- **Privacy and Security Measures**: Ensuring user data remains secure and doesn’t violate privacy standards.

- **Technical Challenges**: Addressing difficulties in detecting advanced, hidden AI techniques.

1.6 Project Constraints

- **Time Limitations**: Fixed deadlines or timeframes for project completion that might affect the depth of development or testing phases.

- **Resource Constraints**: Limited budget for acquiring necessary tools or technologies, or constraints on available workforce.

- **Technological Limitations**: Mandated use of specific programming languages, frameworks, or restrictions on employing certain AI models due to compatibility issues.

- **Regulatory Compliance**: Adherence to data protection laws, privacy regulations, or ethical guidelines governing the use of AI-generated content.

- **Scope Creep Management**: Ensuring the project remains focused on its defined objectives without expanding beyond the established scope.

Chapter Two
2.System analysis And Design

2.1 User and System Requirements

2.1.1 Functional Requirements

1. User Authentication:
- Users can sign up with a valid email address, name, and password.
- Users can log in using their credentials.

2. Media Submission:
- Users can choose the type of media (text, image, or deep fake).
- Users can submit the media (upload text or image).

3. Subscription:
- Users can subscribe to a plan.
- Subscription plans are presented to the user.
- Users can select a plan.
- Users must pay for the selected plan.

4. Profile Management:
- Users, after logging in, can view their profile page.
- Users can edit their profile information.

5. History:
- Users can view a history page that displays their past interactions or submissions.

6. Admin Functions:
- Admins can log in.
- Admins can observe system data and analysis.
- Admins have access to an admin page.

7. Detection System:
- The system can detect AI-generated content.
- The system records the date and result of the content analysis.

2.1.2 Non-functional Requirements

1. Security:

- User passwords are securely stored (hashed and salted).
- Media submissions are securely handled to prevent unauthorized access.

2. Usability:

- The website has an intuitive and user-friendly interface.
- Responsive design for various devices and screen sizes.

3. Performance:

- The detection system should provide timely results.
- The website should handle simultaneous user requests efficiently.

4. Scalability:

- The system should be able to handle an increasing number of users and media submissions.

5. Reliability:

- The system should be available and reliable for users and admins.

6. Payment Processing:

- Secure and reliable payment processing for subscription plans.

7. Logging and Auditing:

- The system logs user actions, especially those related to media submissions and payments.
- Admins have access to detailed logs for analysis.

8. Data Backup:

- Regular backups of user data and system logs.

9. User Notifications:

- Users receive notifications for successful subscription, payment, and other important events.

10. Compliance:

- The system complies with relevant data protection and privacy regulations.

11. User Support:

- Provide a mechanism for users to seek help or support.

12. Cancellation of Subscription: - Users can cancel their subscription, and the system should handle this process appropriately.

2.2 System Design

2.2.1 Use Case Diagram

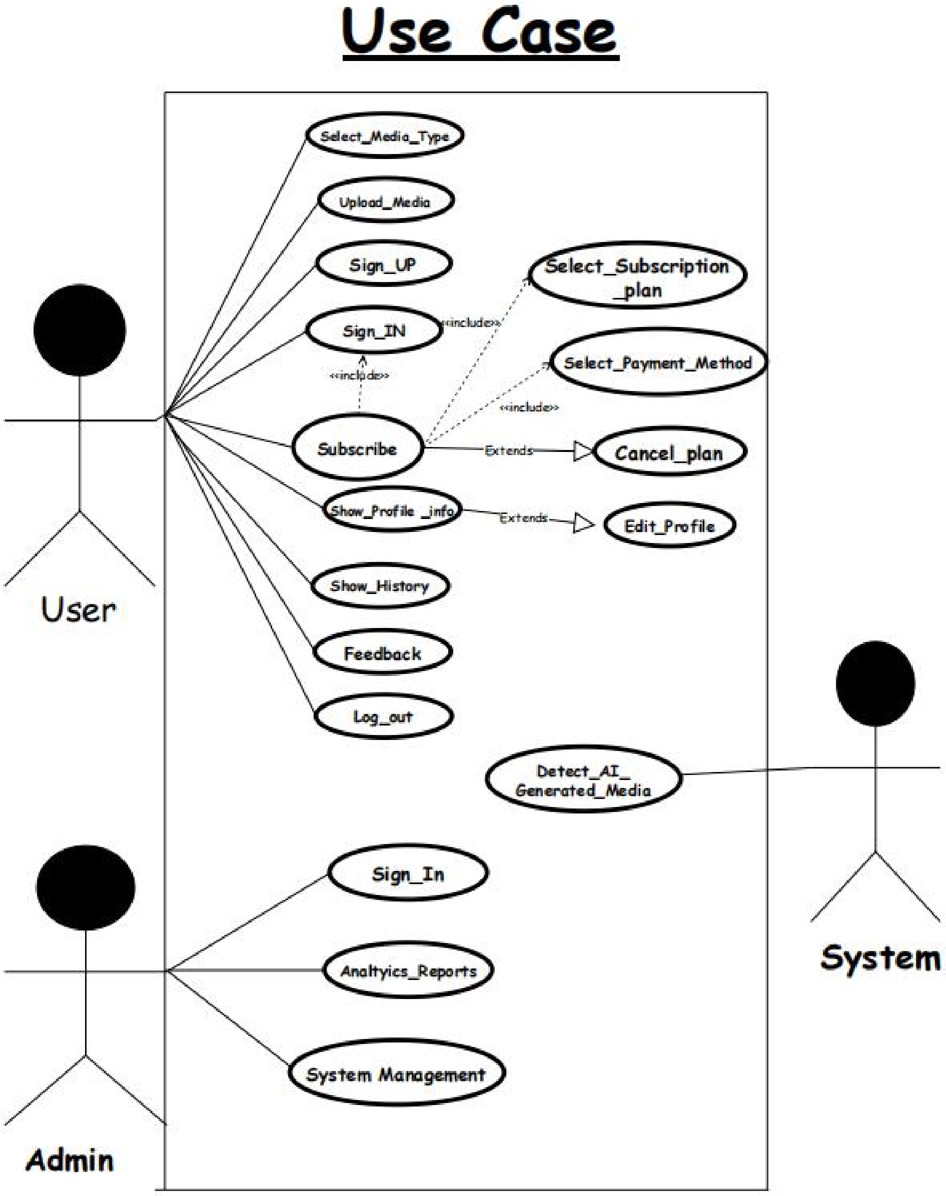


Figure 1

2.2.1.1 Use case definitions

-- Select Media Type

Actor	User
Pre-conditions	User is on the main page.
Post-conditions	Media type is selected
Basic flow	1-User navigates to the main page. 2-User selects the media type.
Relationship	User - Main Page: The user interacts with the main page to navigate to different sections, including selecting the media type.

--Enter the Media

Actor	User
Pre-conditions	User has selected a media type.
Post-conditions	Media is entered.
Basic flow	-User selects a media type. -User enters the media.
Relationship	User - Selected Media Type: The user's action of selecting a media type is a prerequisite for entering the media. User - Media Entry: The user interacts with the system to input or provide the media content.

-Sign In or Sign Up

Actor	User
Pre-conditions	User is on the main page.
Post-conditions	User is signed in or signed up
Basic flow	-User navigates to the main page. -User chooses to sign in or sign up.
Relationship	User - Main Page: The user's action of navigating to the main page is a prerequisite for signing in or signing up. User - Authentication status: The user interacts with the authentication system during the sign-in or sign-up process

-Select Subscription Plan

Actor	User
Pre-conditions	User is signed in or signed up.
Post-conditions	User is subscribed to a plan.
Basic flow	-User signs in or signs up. -The system displays available subscription plans. -User selects a subscription plan based on their preferences. - User makes a payment.
Relationship	User - Authentication Status: The user's sign-in or sign-up status is a prerequisite for subscribing to a plan. User - Subscription Page: The user interacts with the subscription page to choose a plan User - Payment System: The user interacts with the payment system during the subscription process.

-Logout

Actor	User
Pre-conditions	User is signed in
Post-conditions	User is logged out
Basic flow	-The user, already signed in, clicks on the "Log Out" option. -The system receives the log-out request from the user. -The system updates the user's session status to indicate a logged-out state. -The user is successfully logged out. -The system provides feedback to the user, confirming the successful log-out.
Relationship	User - Authentication Status: The user's sign-in status is a prerequisite for logging out. User - Session Management: The user's session is managed during the logout process.

-Show History (Page)

Actor	User
Pre-conditions	User is signed in.
Post-conditions	User views the history page.
Basic flow	-The user, who is already signed in, navigates to the section or option that allows them to view the history. This could be a button or a link labeled "Show History." -The system recognizes the user's request to view the history and checks for an active session. -If the user has an active session, the system proceeds to the next step. -If the user doesn't have an active session, they may be prompted to sign in again. -The system grants access to the history component since the user has a valid session. -The user interacts with the history component, which could involve navigating through a list of historical data or specifying parameters for the history they want to see. -The system retrieves and displays the relevant history information based on the user's request. -The user views the history page, containing the requested historical data.
Relationship	User - Session: The user needs an active session to access the history page. User - History Component: The user interacts with the history component during the process.

-Show Profile Info

Actor	User
Pre-conditions	User is signed in
Post-conditions	The user views their profile information.
Basic flow	-The user, already signed in, navigates to the profile section. -The system displays the user's profile information.
Relationship	User - Profile Component: The user interacts with the profile component during the process, indicating a relationship with the profile feature.

-Edit profile

Actor	User
Pre-conditions	User is signed in.
Post-conditions	The user successfully updates their profile information.
Basic flow	-The user, already signed in, navigates to the profile section. -The system displays the user's current profile information. -The user selects the option to edit their profile. -The system presents a form with the user's current information for editing. -The user makes desired edits and submits the form. -The system verifies and updates the user's profile information. -If the update is successful, the system notifies the user. -If there are errors, the system displays error messages and prompts the user to correct them
Relationship	User - Profile Component: The user interacts with the profile component to make changes, establishing a relationship with the profile editing functionality

-Select Payment Method

Actor	User
Pre-conditions	User has selected a subscription plan.
Post-conditions	User has chosen a payment method, and the subscription is activated.
Basic flow	-After selecting a subscription plan, the user proceeds to the payment step. -The system presents various payment methods (credit card, PayPal, etc.). -User selects their preferred payment method. -The system processes the payment and activates the chosen subscription plan.
Relationship	User - Payment Process: The user interacts with the system to choose a payment method and complete the subscription process

-Cancel Plan:

Actor	-User
Pre-conditions	User has an active subscription.
Post-conditions	User's subscription is canceled.
Basic flow	-User navigates to the subscription management section. -The system displays the user's active subscription details. -User selects the option to cancel the subscription. -The system processes the cancellation request. -The user's subscription is canceled, and they no longer have access to the subscribed features
Relationship	User - Subscription Management: The user interacts with the system to manage their subscription, including canceling the plan.

-Sign In (Admin)

Actor	Admin
Pre-conditions	The admin is registered in the system with valid credentials.
Post-conditions	-Admin is signed in. -The admin gains access to the admin dashboard.
Basic flow	-The admin navigates to the admin sign-in page. -The system presents a form for the admin to input their credentials (username and password) -The admin enters their valid credentials. -The system verifies the admin's credentials. -If the credentials are valid, the system grants access to the admin dashboard. If the credentials are invalid, an error message is displayed, and the admin is prompted to re-enter their credentials
Relationships	Admin - Authentication status: The admin interacts with the authentication module during sign-in. This relationship signifies the involvement of the authentication system. //Admin - Session: A session is established upon successful sign-in, allowing the admin to interact with the system, indicating a dependency on the session management component.

-Analytics Report

Actor	Admin
Pre-conditions	Admin is signed in.
Post-conditions	Admin views analysis reports
Basic flow	-The admin, already signed in, navigates to the analytics report section. -The system retrieves and displays analytics reports.
Relationships	Admin - Analytics Report Interaction: The admin interacts with the analytics module to retrieve and view analysis reports, establishing a relationship with the analytics component. Admin - Authorization for Analytics Report: This relationship indicates the need for proper authorization to access and view analytics reports. The authorization component is involved in ensuring the admin has the necessary permissions to access the analytics report feature.

-System Management:

Actor	Admin
Pre-conditions	Admin is signed in.
Post-conditions	The admin has access to system management tools.
Basic flow	-The admin, already signed in, navigates to the system management section. -The system provides access to various system management tools.
Relationships	-System management tools are available only to authenticated admins.Actor -Actor -Admin: Accesses and utilizes system management tools.

Detect AI-Generated Media

Actor	Admin
Pre-conditions	Media is present in the system.
Post-conditions	**The process concludes with the successful identification of AI-generated media.**
Basic flow	**-The system initiates the process to analyze media content.** **-The system utilizes the Media Analysis Module to conduct a thorough analysis of the provided media content.** **-The analysis involves various checks and assessments to determine the characteristics and patterns within the media.** **-The system employs the AI Detection component to specifically identify any instances of AI-generated media based on predefined criteria and patterns.** **-If AI-generated media is detected, the system proceeds to the next step.** **-If no AI-generated media is detected, the process may end, and the system can provide feedback or log the result.** **-The system marks or flags the identified media as AI-generated.**
Relationships	**System - Media Analysis Module**: The system interacts with a media analysis module for the analysis of media content, indicating a relationship with the analysis component. **System - AI Detection:** There is a relationship with the AI detection component, which is responsible for identifying AI-generated media

2.2.2 Activity Diagram

2.2.2.1 Sign up activity

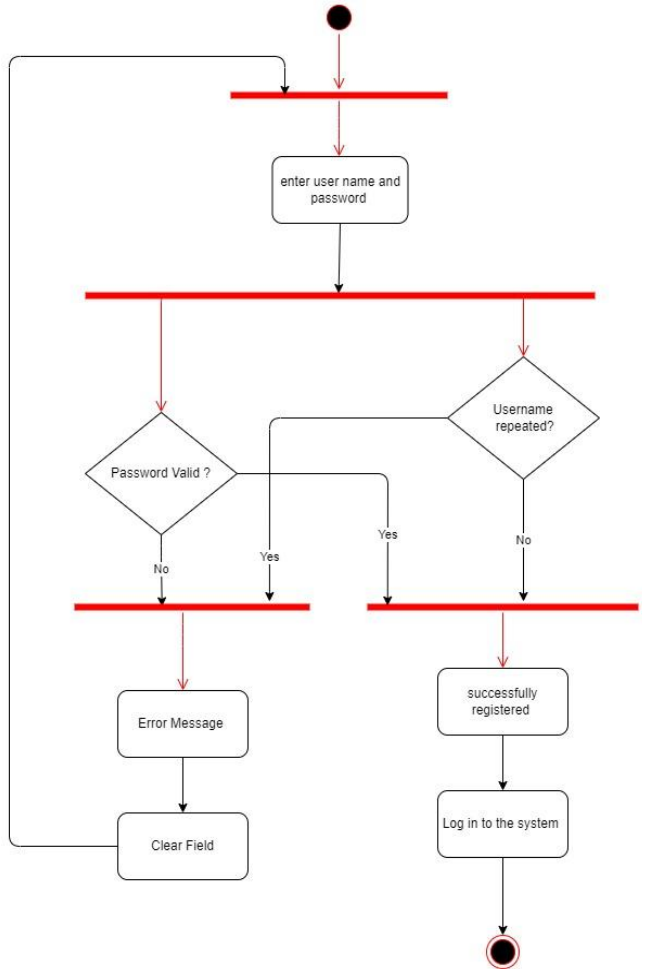
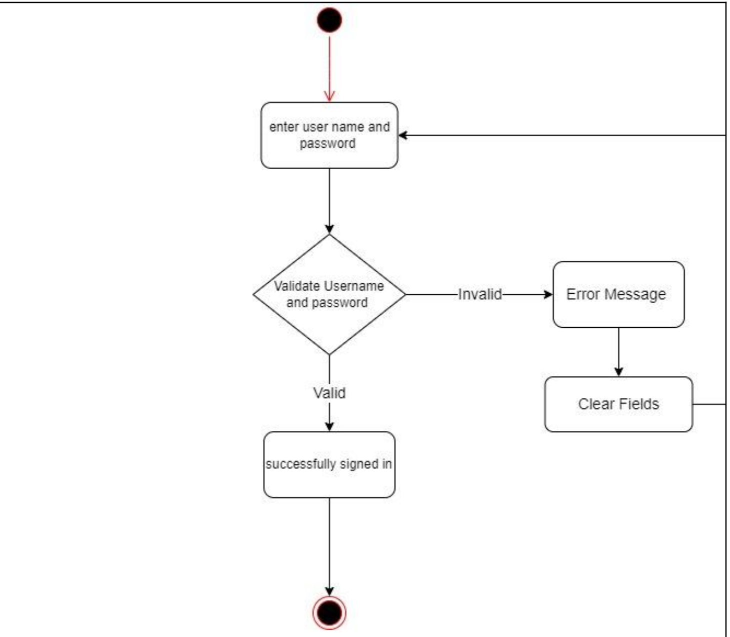
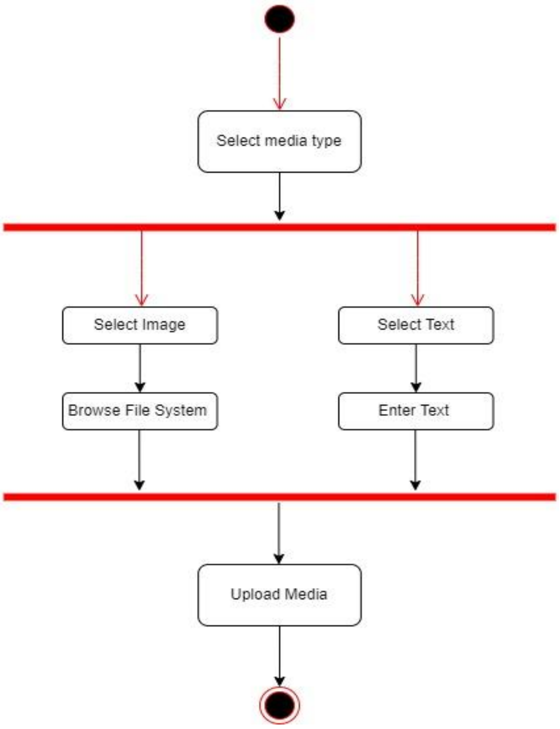


Figure 2

2.2.2.2 Sign In Activity	 enter user name and password Validate Username and password Invalid Error Message Clear Fields Valid successfully signed in Figure 3
2.2.2.3 Select media type activity.	 Select media type Select Image Select Text Browse File System Enter Text Upload Media Figure 4

2.2.2.4 Upload media activity

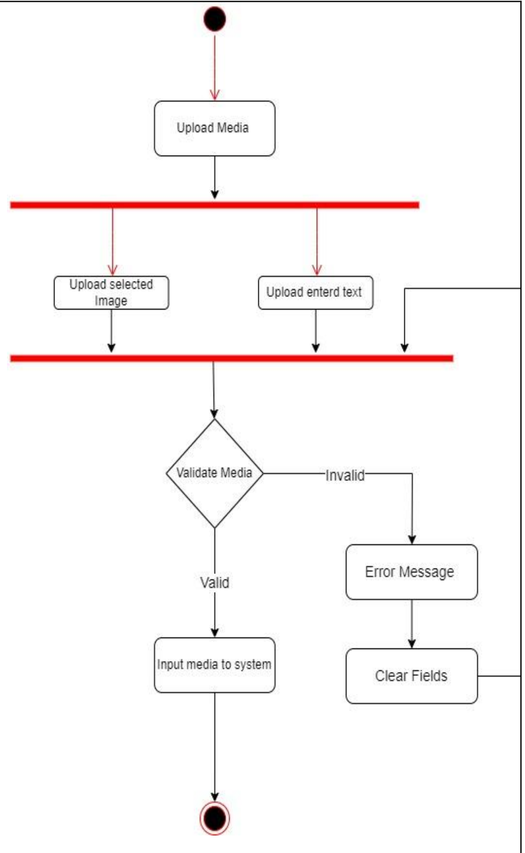
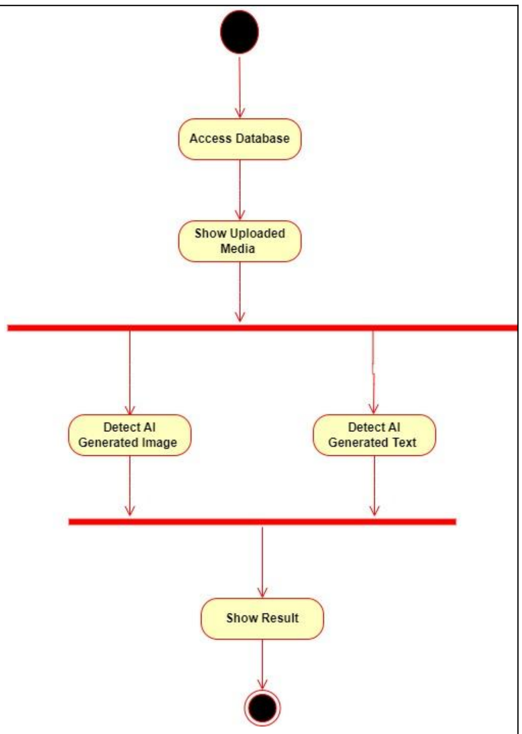


Figure 5

2.2.2.5 Detect Ai-generated Media activity.	 Access Database Show Uploaded Media Detect AI Generated Image Detect AI Generated Text Show Result Figure 6

2.2.2.6 Feedback activity

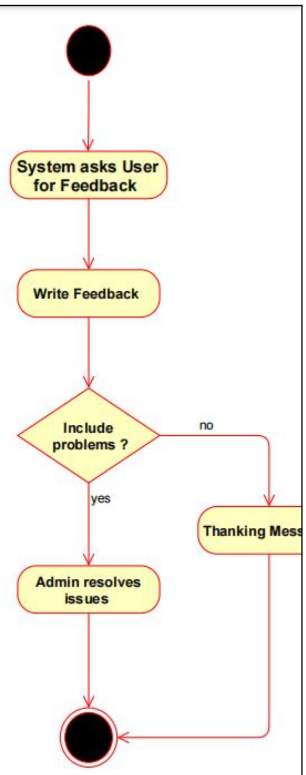


Figure 7

2.2.2.7 Subscription activity

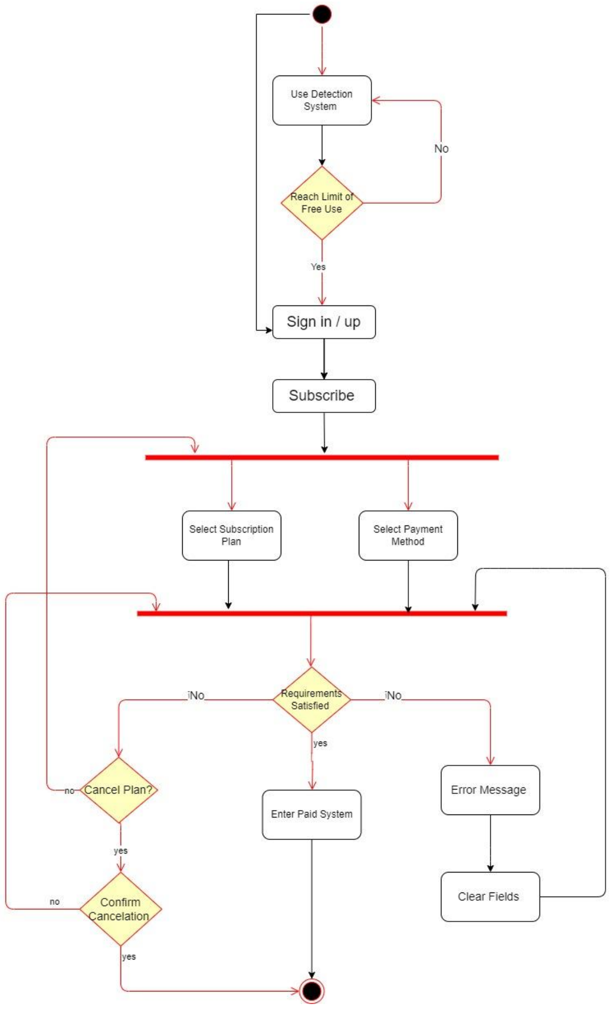
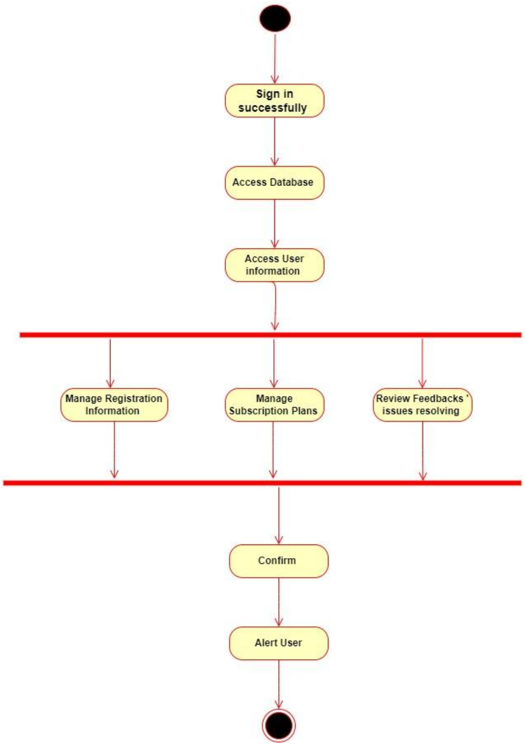


	Figure 8
2.2.2.8 Admin Management activity.	 Sign in successfully Access Database Access User information Manage Registration Information Manage Subscription Plans Review Feedbacks ' issues resolving Confirm Alert User Figure 9

2.2.2.9 Show History activity.

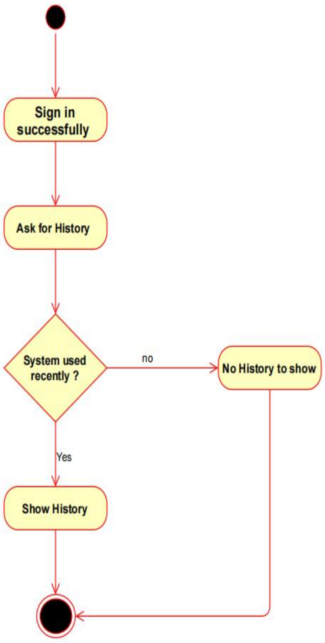


Figure 10

2.2.2.10 Edit Profile activity.

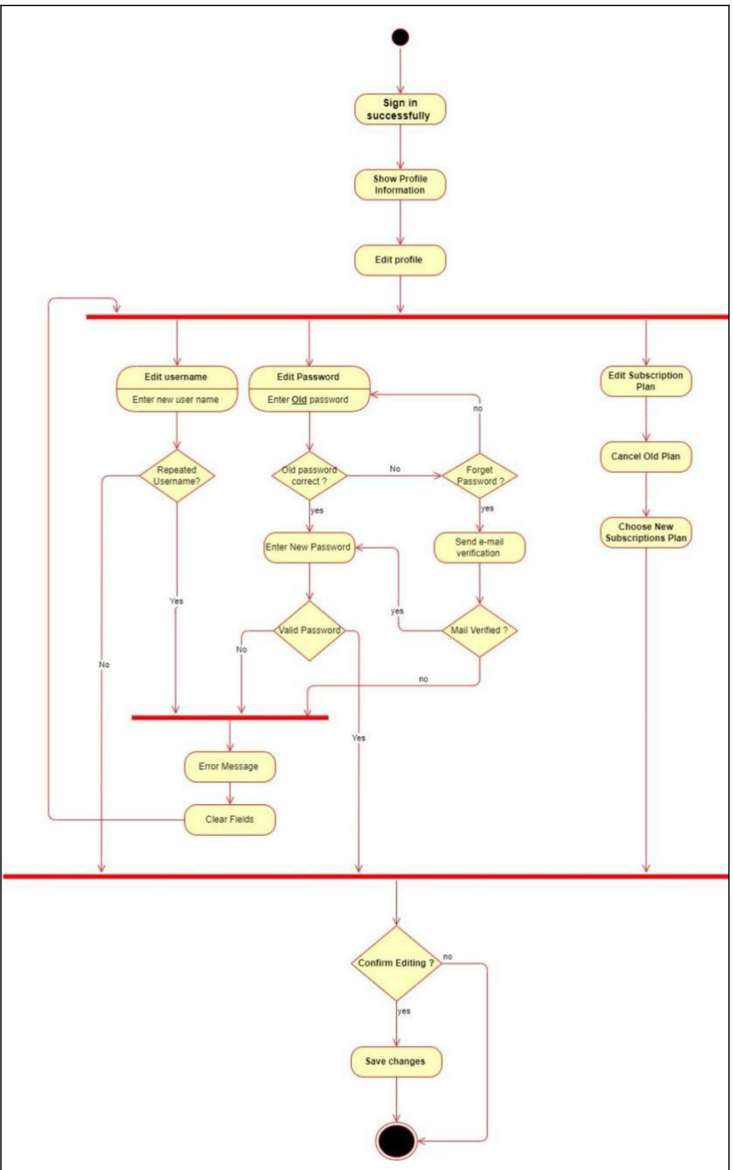


Figure 11

2.2.3 Context Diagram

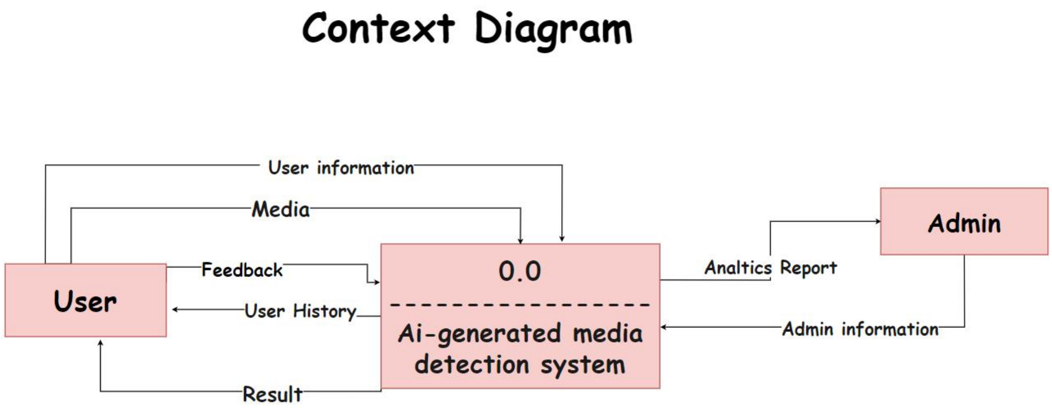


Figure 12

2.2.4 Data Flow Diagram

2.2.4.1 Level 0

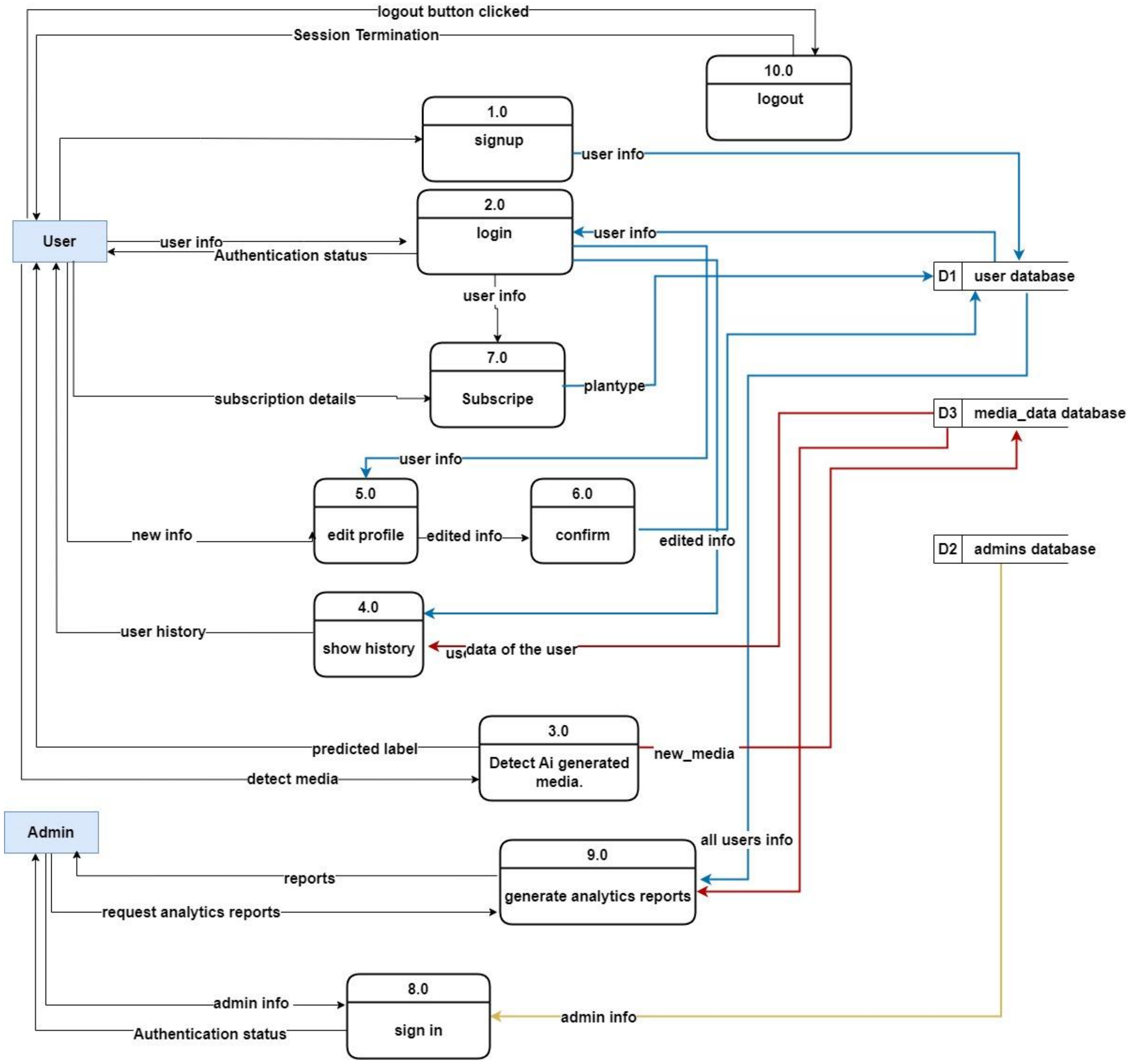


Figure 13

2.2.4.2 level 1

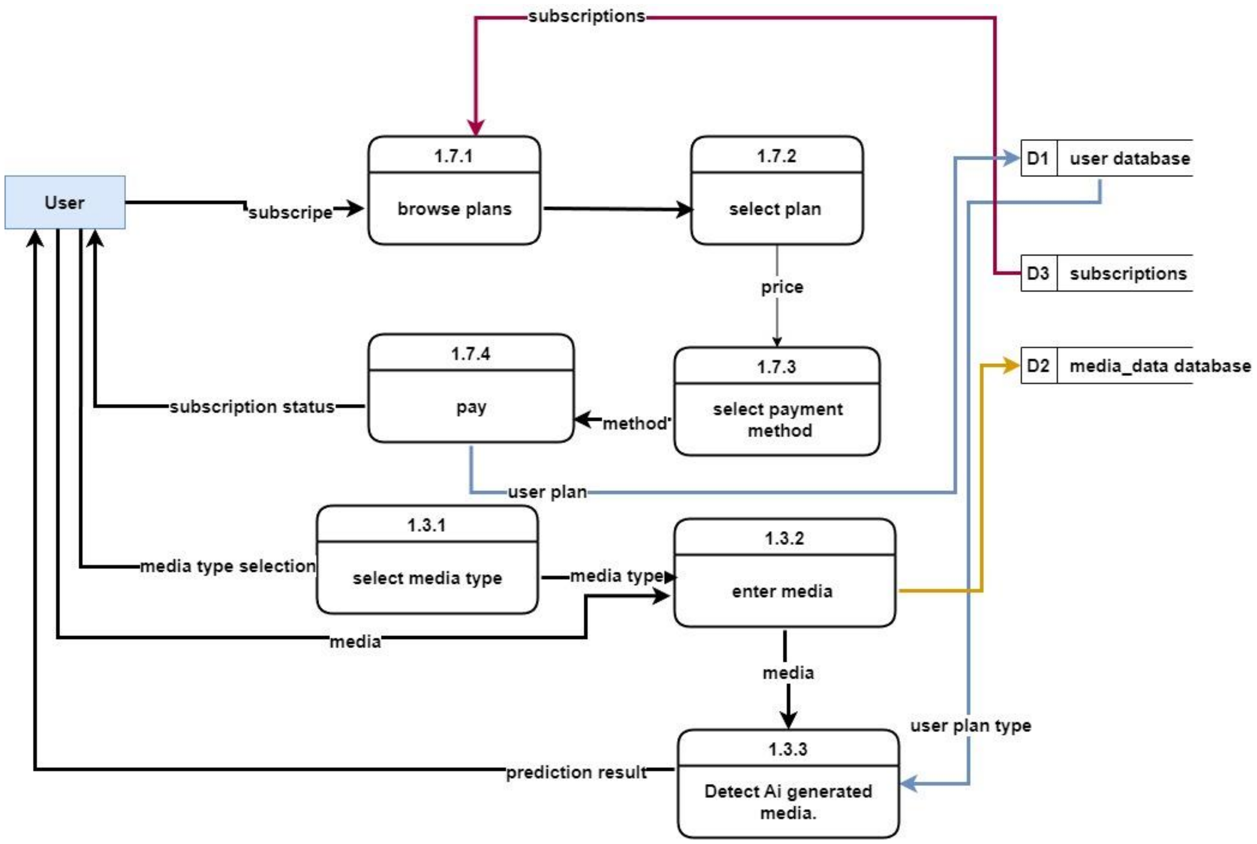


Figure 14

2.2.5 Sequence diagram

2.2.5.1-User

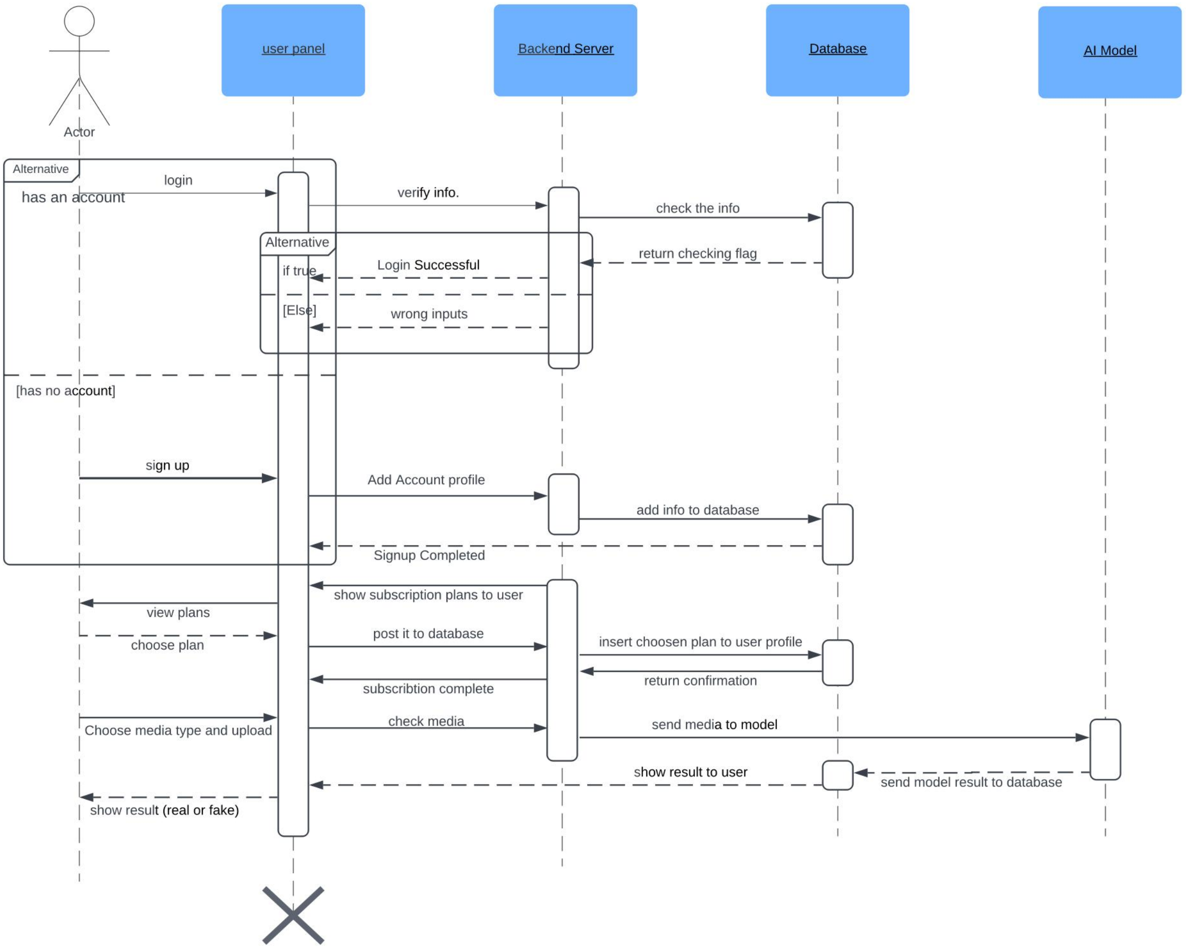


Figure 15

2.2.5.2 Admin

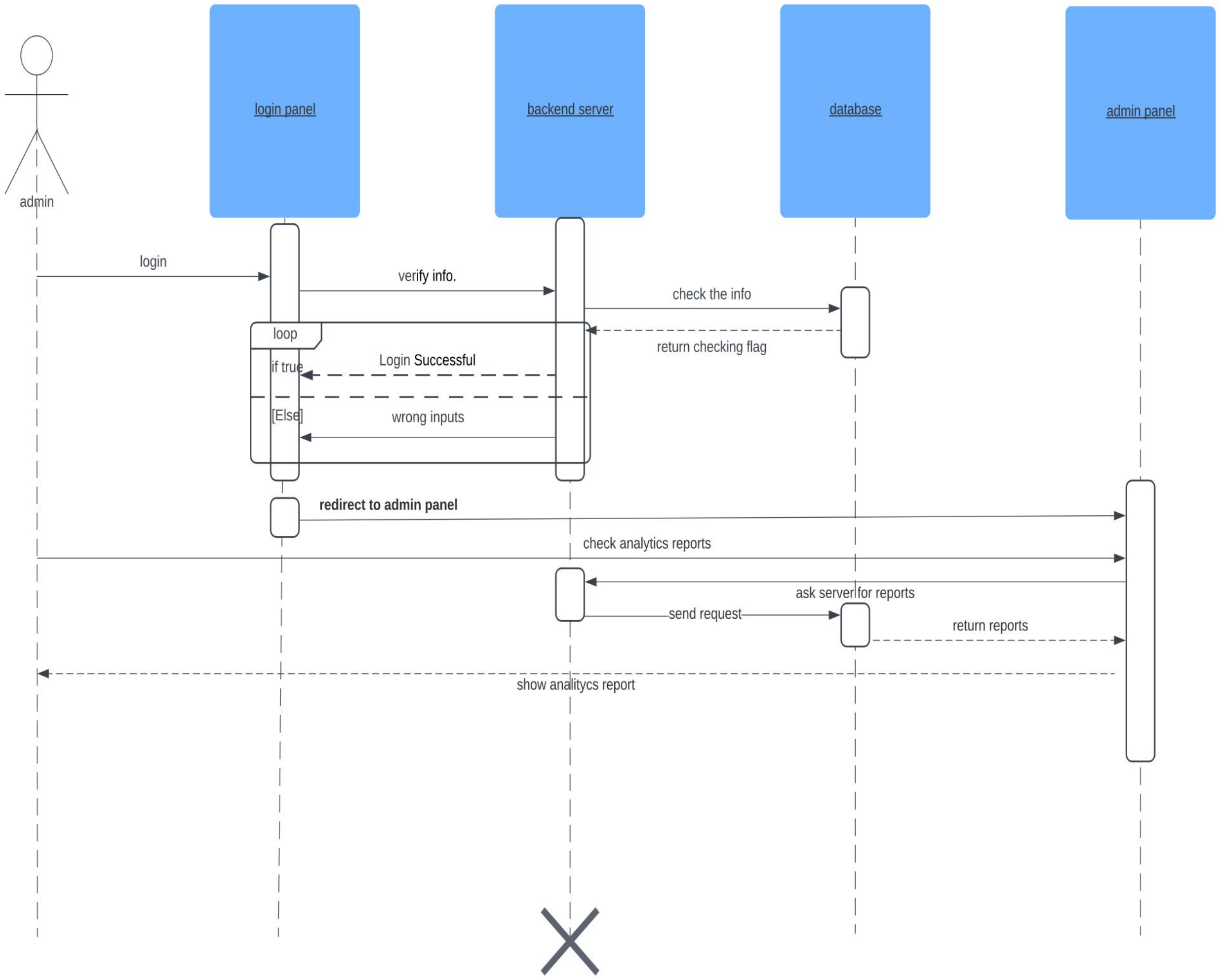


Figure 16

2.2.6 Class Diagram

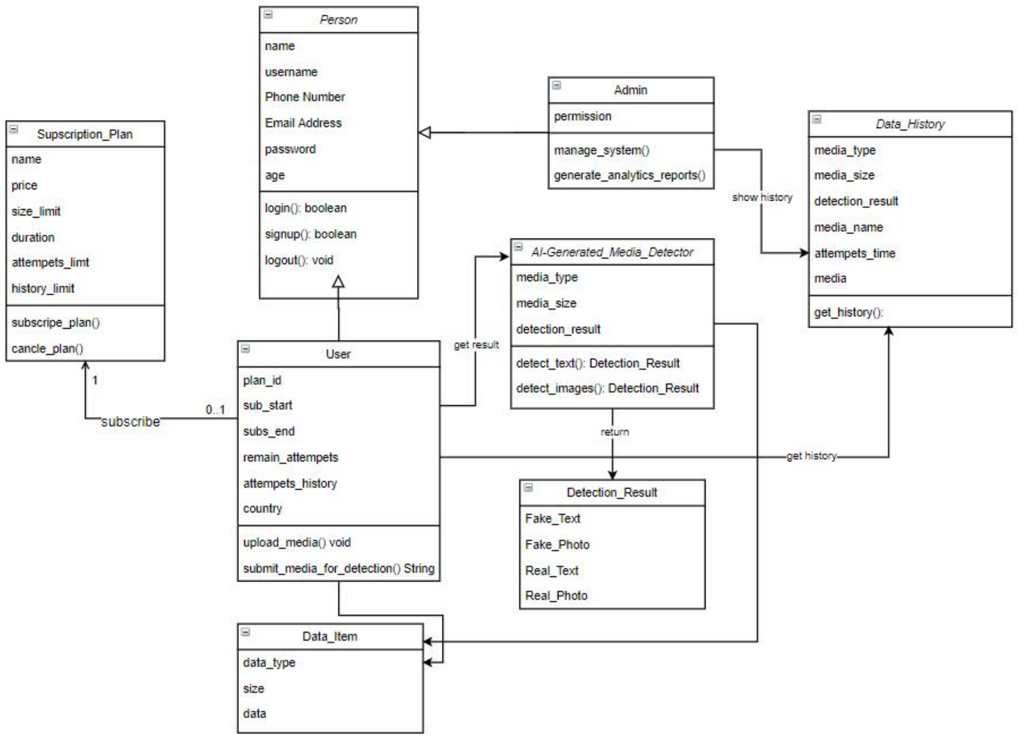
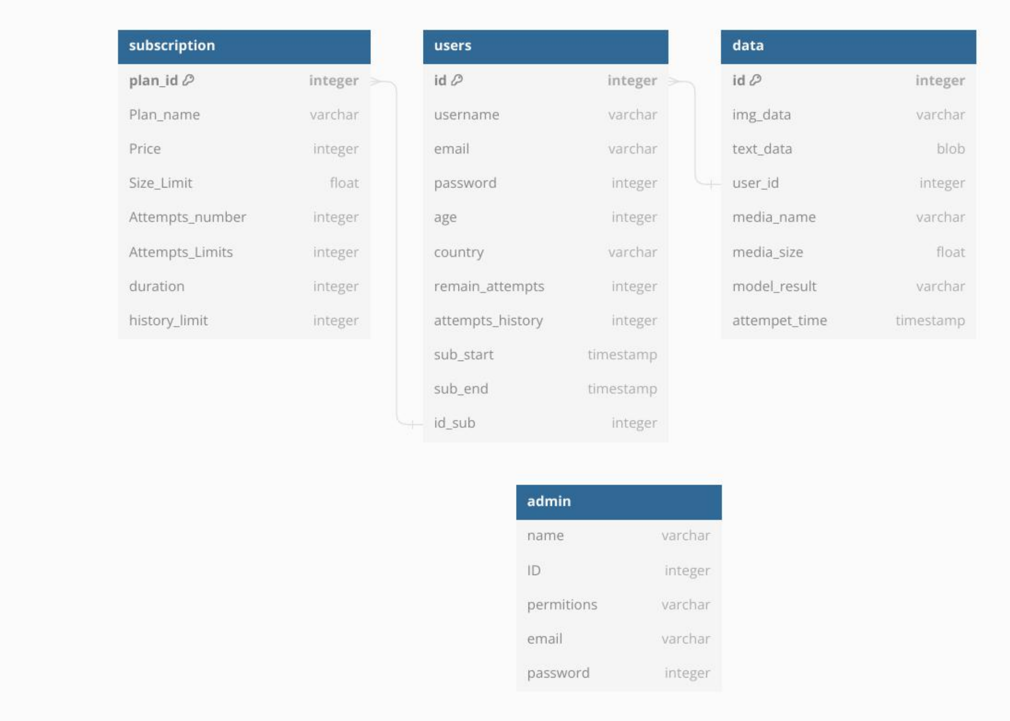


Figure 17

2.2.7 Database Design



subscription	
plan_id	integer
Plan_name	varchar
Price	integer
Size_Limit	float
Attempts_number	integer
Attempts_Limits	integer
duration	integer
history_limit	integer

users	
id	integer
username	varchar
email	varchar
password	integer
age	integer
country	varchar
remain_attempts	integer
attempts_history	integer
sub_start	timestamp
sub_end	timestamp
id_sub	integer

data	
id	integer
img_data	varchar
text_data	blob
user_id	integer
media_name	varchar
media_size	float
model_result	varchar
attempet_time	timestamp

admin	
name	varchar
ID	integer
permitions	varchar
email	varchar
password	integer

Figure 18

2.1 2.2.8 ERD Diagram

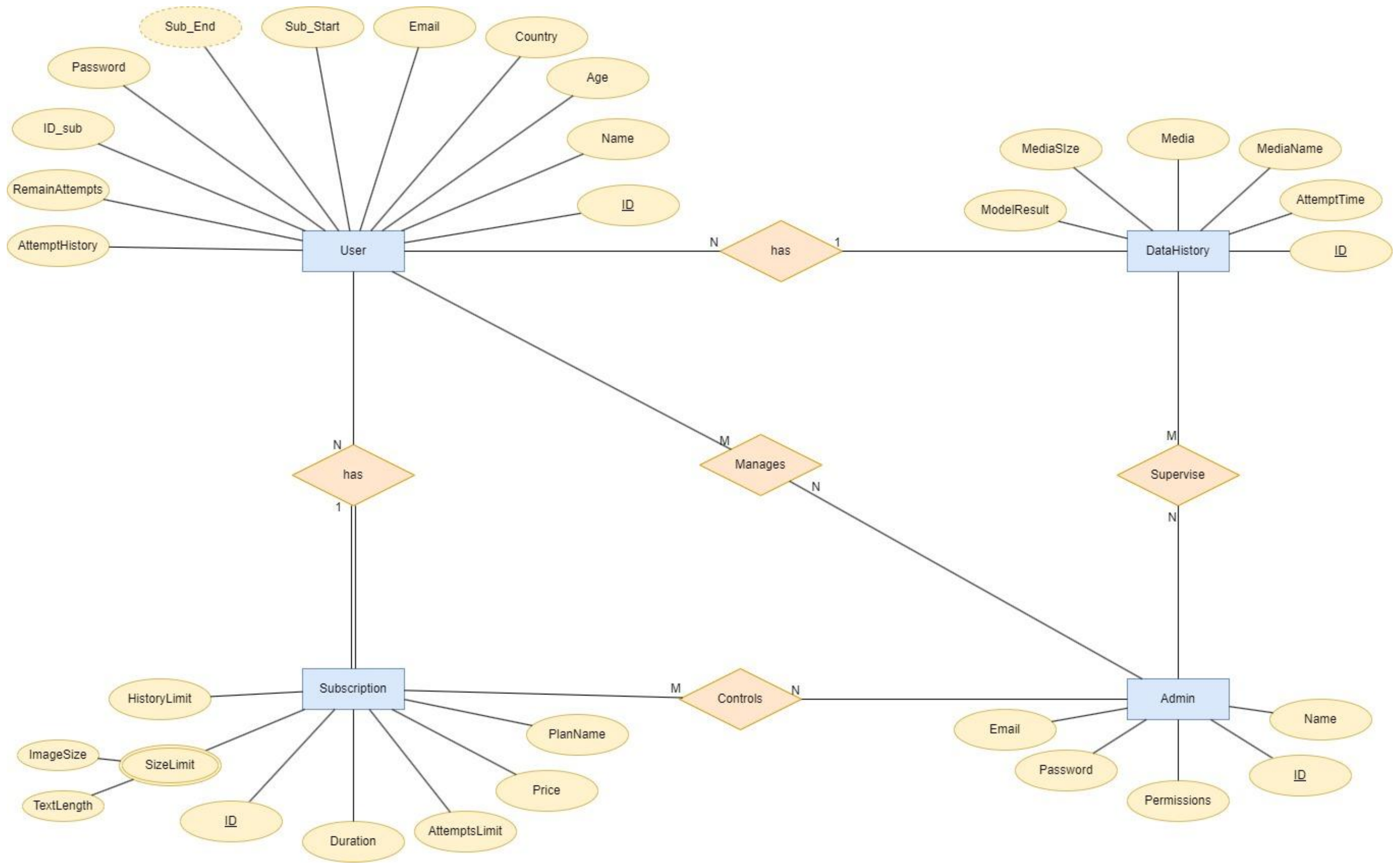


Figure 19

2.3 Used Technologies and tools.

2.3.1.Technologies

2.3.2.tools