AI-Generated Media Detection System

Supervised by Dr. **Eman Abdel-Latef**

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 Team

1. Romani Nasrat Shawqi Gerges	روماني نصرت شوقي جرجس .1
2. Ahmed Mohamed Ali Abo el-Kassem	احمد محمد علي ابو القاسم .2
3. Zeyad Elsayed Abdel-Azim Ali	زياد السيد عبدالعظيم علي يوسف .3
4. Sara Reda Moatamed Hassan	ساره رضا معتمد حسن .4
5. Reham Moustafa Ali Abdel-Moati	ريهام مصطفى علي عبد المعطى .5
6. Rawan Abdel-Aziz Ahmed Mahmoud	روان عبدالعزيز أحمد محمود .6
7. Abd-Allah Mohamed Abdel-monem	عبدالله محمد عبدالمنعم عبدالفتاح .7
8. Mohannad Ayman salah Abdel-fattah	مهند ايمن صلاح عبدالفتاح .8
9. Mohamed Abd-Allah Abdel-salam Abdel-Dayem	محمد عبدالله عبدالسلام عبدالدايم .9

Project Supervisor

Faculty member	## Dr. Eman Abdel-Latef

Assistant

Teaching Assistant	Eng. Sahar

TABLE OF CONTENTS

Contents

LIST OF FIGURES

ABSTRACT

years have made the difficulty of differentiating between the media (images or text) generated by generative adversarial networks (GANs) , Transformers or large language models (LLMs) that closely resemble the real thing has increased, particularly with the remarkable development of   GANs or LLMs and the rapid advancement of artificial intelligence. And this started performing this duty better than expected. In the upcoming years, a major development is anticipated. This raises many questions: What if these models are used in a negative and harmful way? What would happen if people used it to create fake images for social media scams? Or how will this affect the level of the educational process for many people by making these models perform academic tasks, making it difficult for the teacher to evaluate students correctly? More and more of these questions and problems can arise as that continue to develop.

So, in this project, we are working hard to help figuring out whether this content (images or text) is generated or processed by powerful AI models (LLMS or GANs) or by human.

Chapter One

1. project initiation

1.1.   problem DEFINITION

Companies like OpenAI, Meta, Microsoft, and Google are racing ahead in AI, particularly in things like transformers, GANs, and LLMs. They're all competing fiercely because they want to lead this tech revolution. Each one knows that AI will shape the future, so they're striving to create the best models that can generate amazing content. For instance, OpenAI has ChatGPT, Meta developed Llama, Microsoft has Bing, and Google created Gemini. It's a race to be at the forefront of this game-changing technology.

New types of AI, like Generative Adversarial Networks (GANs) and Transformers, have made huge improvements in creating realistic images. But because it's getting harder to tell fake AI-generated made images from real ones, there's a rise in problems like fake profiles, scams, and spreading lies. Just like with fake news, we need smart tools to spot these fake images. It's super important to have ways to check if a picture was made by a computer or a person. This helps us keep trust in what we see online.

Large language Models (LLMs) can write things that are exactly like what people write. But it is difficult to know whether words come from a person or a machine. This can be a problem for teachers who are checking whether students have done their own work or used these forms to finish assignments.

These large language models create text that looks almost exactly like what people write. This is great for a lot of things! But it also brings new problems. Therefore, we will make tools to find out whether text was generated by artificial intelligence (LLMs) or by a person.

Open-source technology that generates text and images has become so advanced that it's hard to tell what's made by machines or people. Bad actors use this to create convincing fake content, which makes it super important to

have strong systems that can spot the difference between what's made by a machine and what's made by a human.

1.2. problem SOLUTION

We will build a website containing our detector models. It contains detector models, which are: A model for detecting images generated using ai, and a special model for detecting images whose features have been manipulated using ai and a text detector model that was generated by ai. We made it a website to make it easier for everyone to use.

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**: who uses this website to detect images and texts generated by AI
➢ **Admin**: who manage user account and profile

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
System analysis And Design

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.
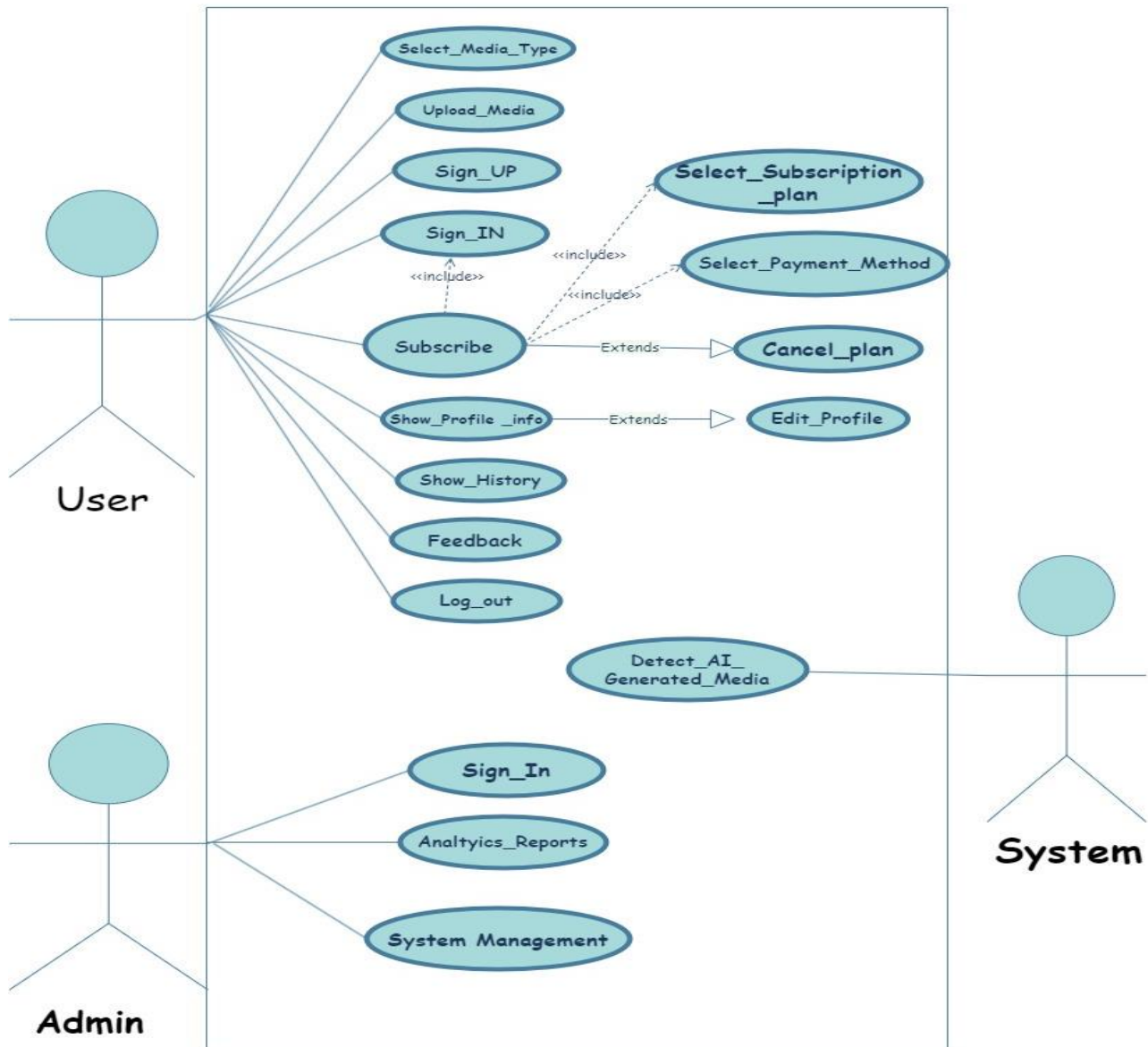
2.2 System Design

2.2.1 use case diagram.

Figure 1

2.2.2 use-case definitions.

1. 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.

2-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.

3-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

4-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.

5-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.

6-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.

7-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.

8-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

9-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

10-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.

Admin :Sign In

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.

System :
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

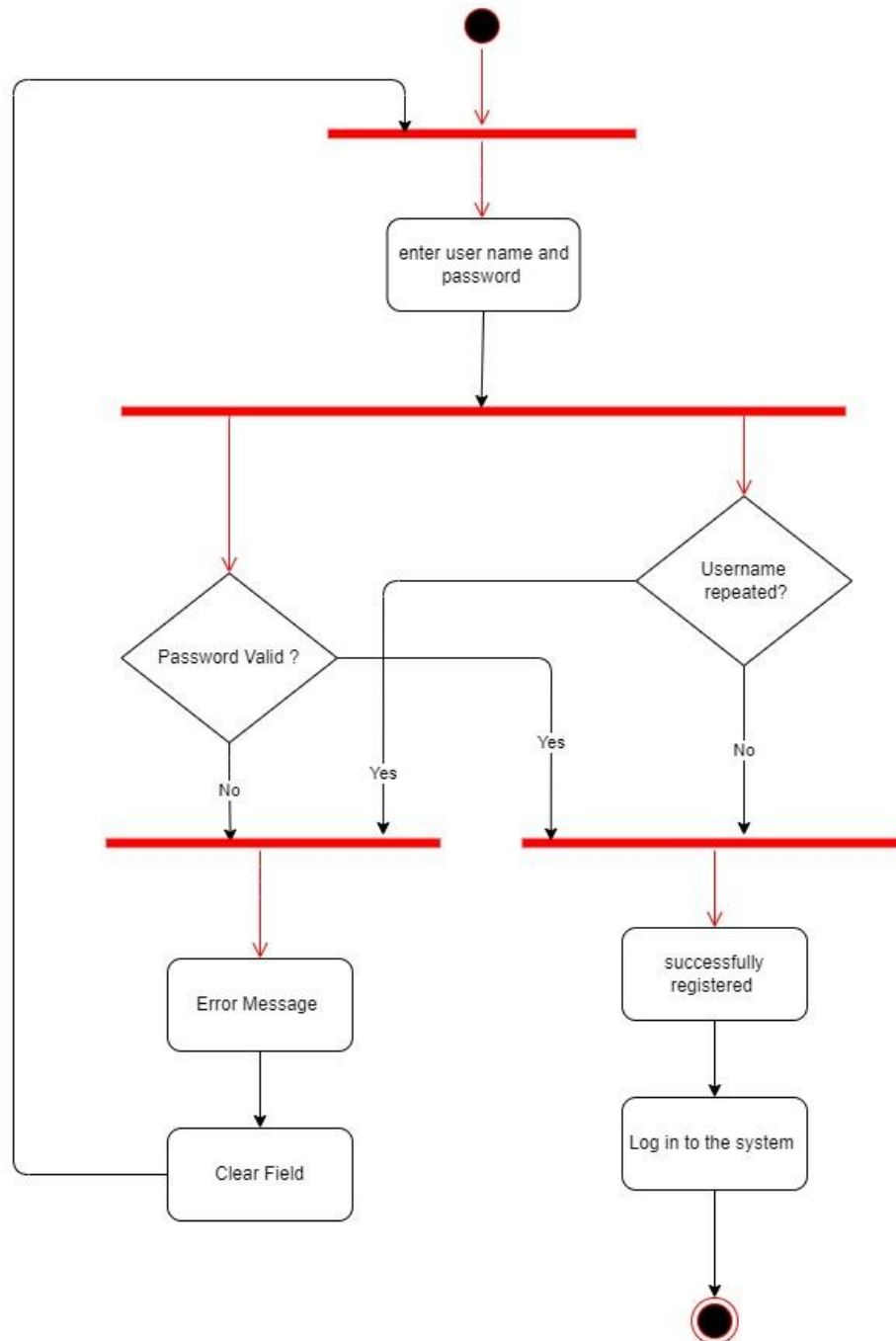### 2.2.3   ACTIVITY diagram

1. Sign up activity

Figure 2

2. Sign in activity

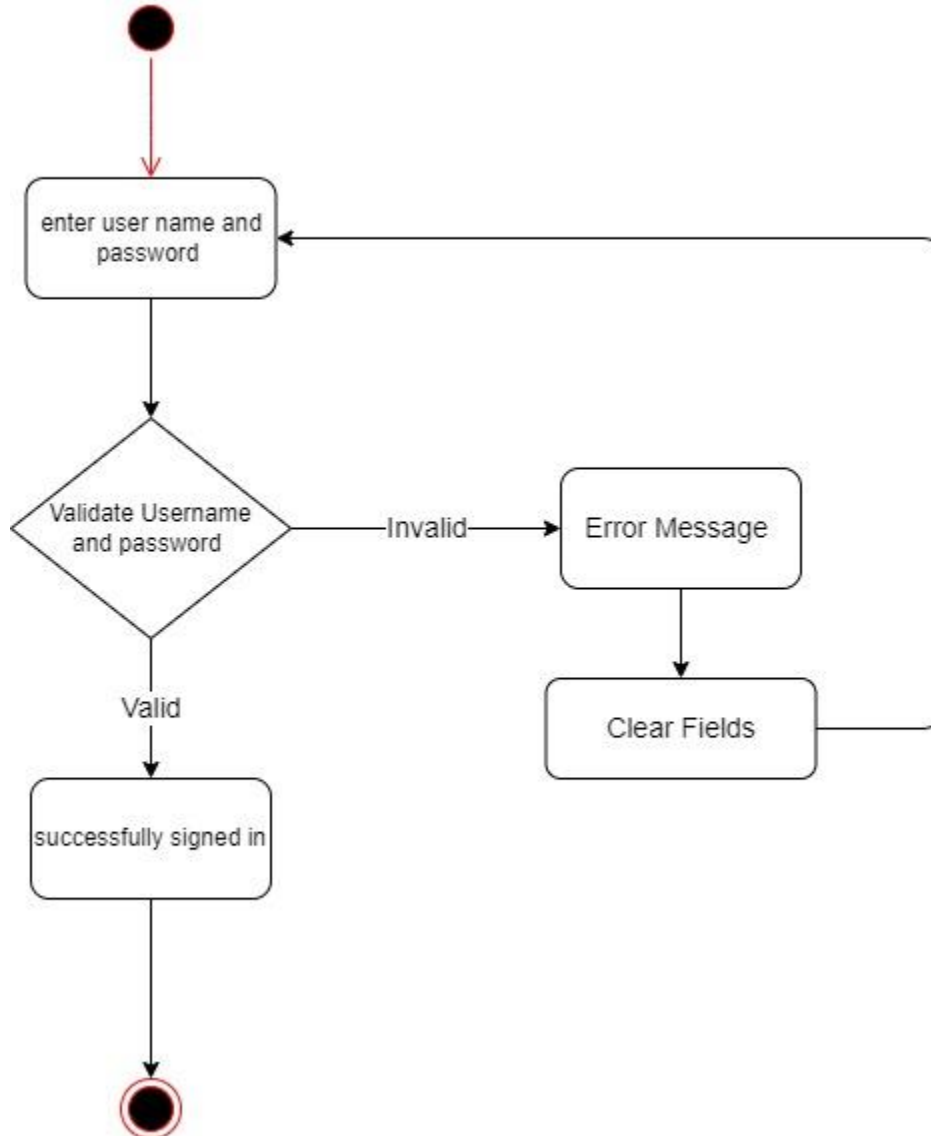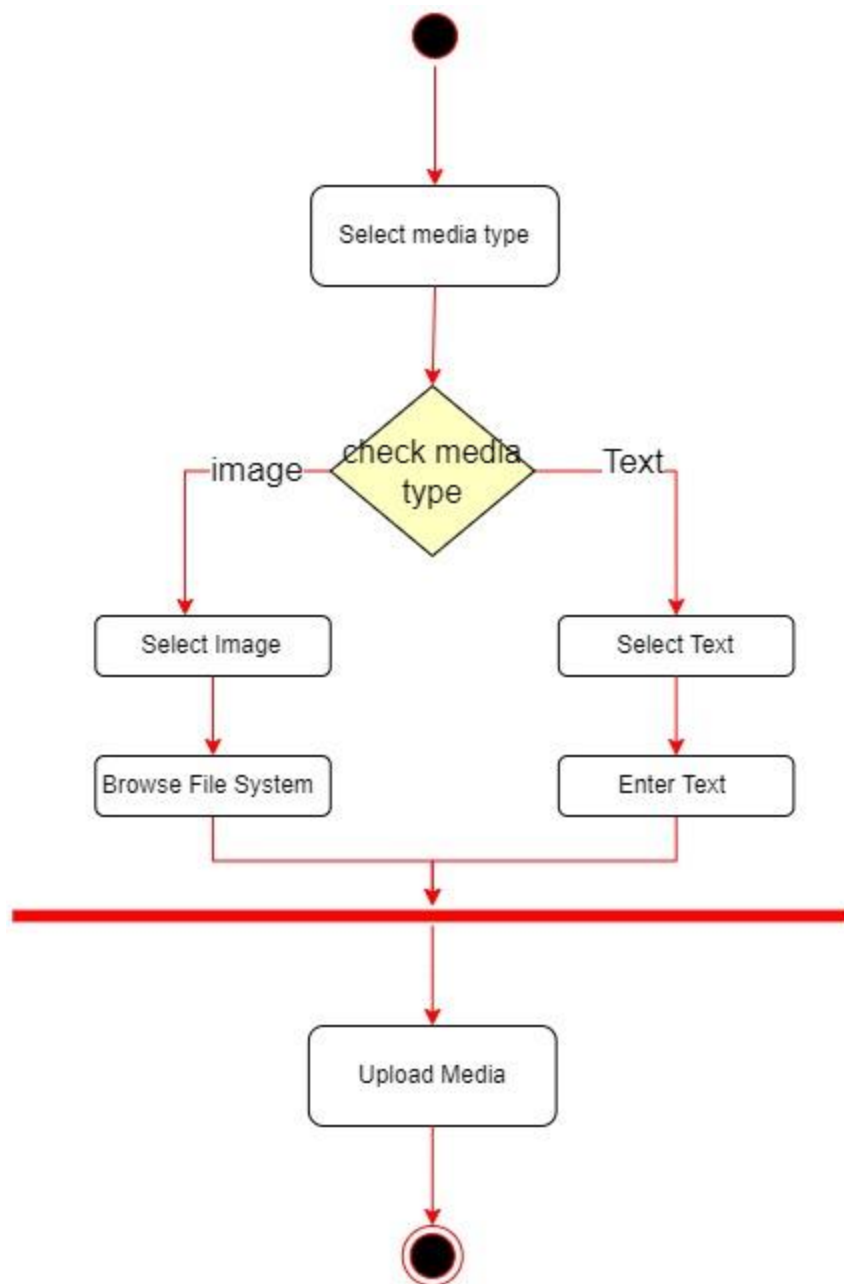
Figure 3

3. Select media type activity.

Figure 4

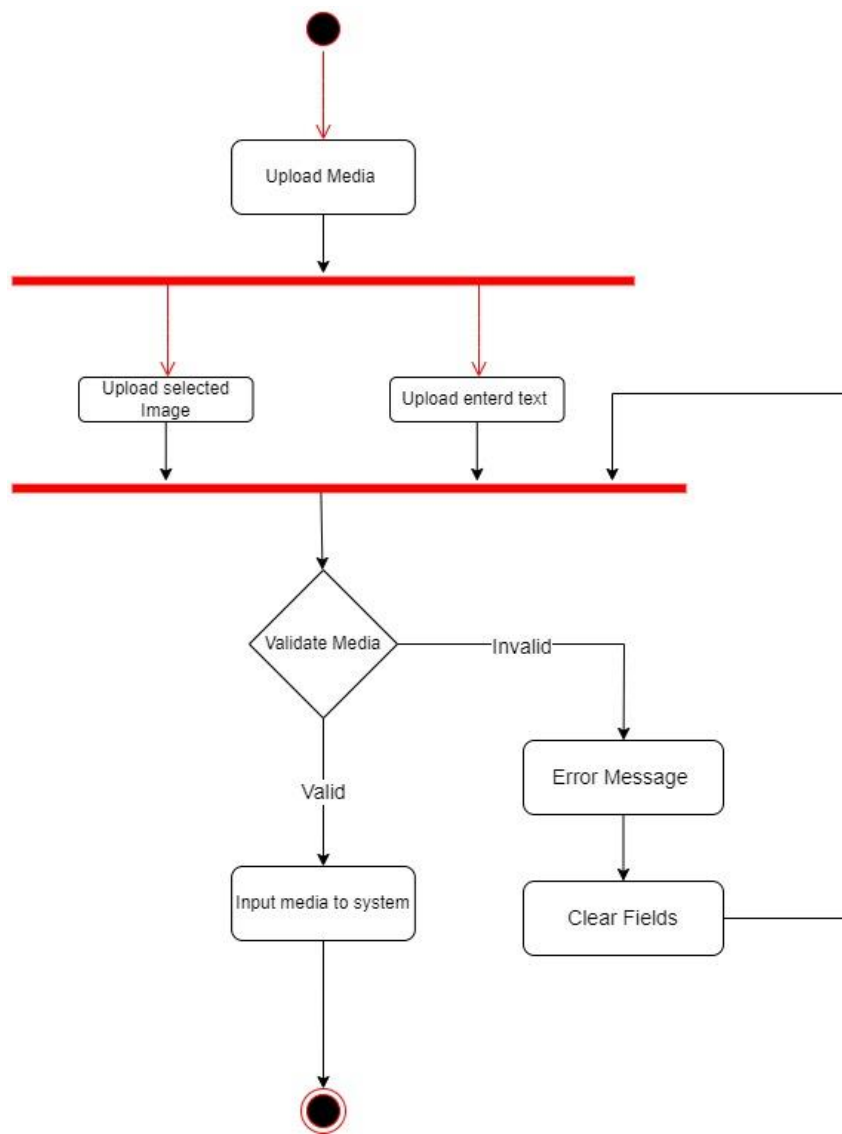## 4. Upload media activity

Figure 5

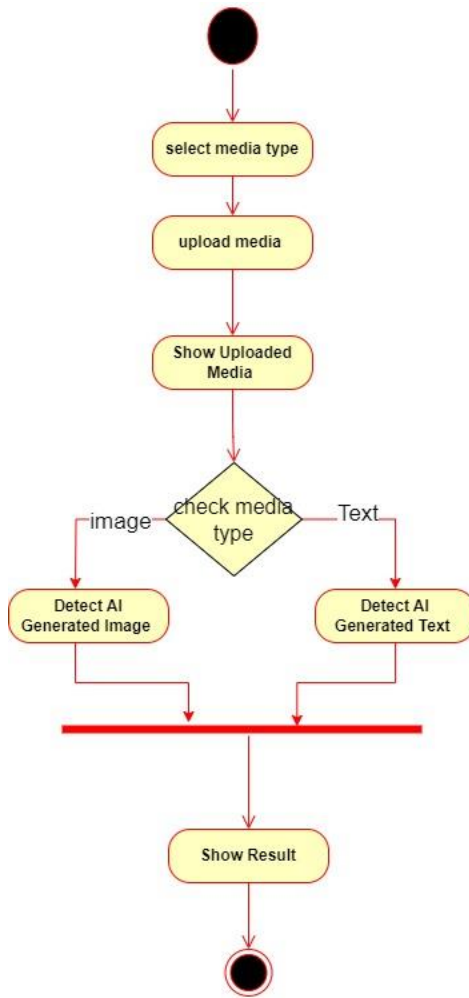## 5. Detect Ai-generated Media activity.
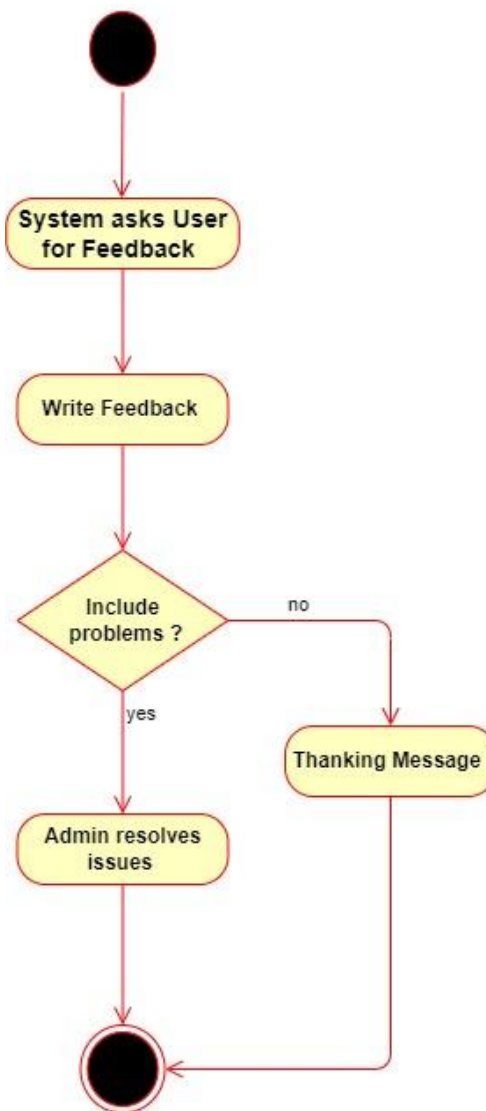

Figure 6

6. Feedback activity

Figure 7

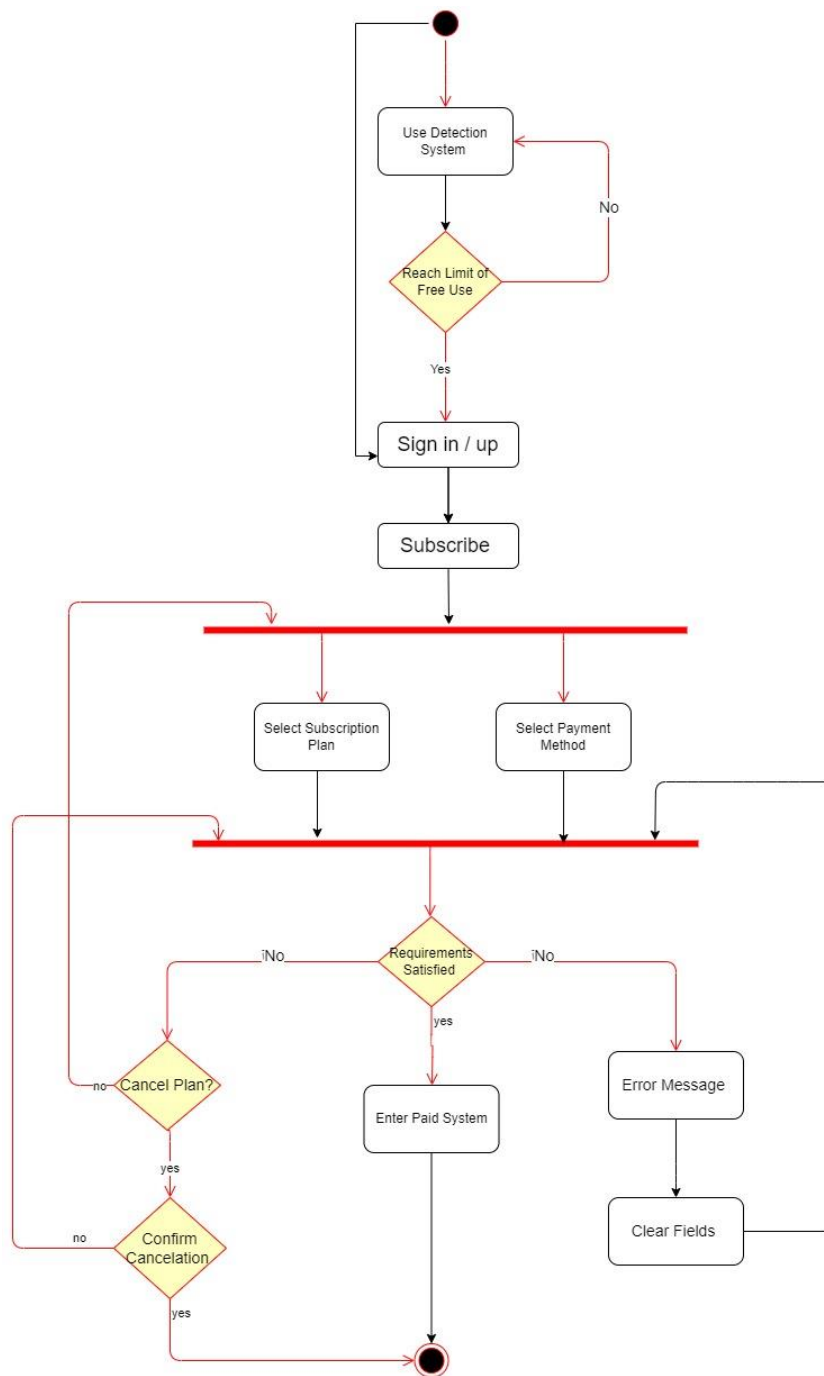## 7. Subscription activity

Figure 8

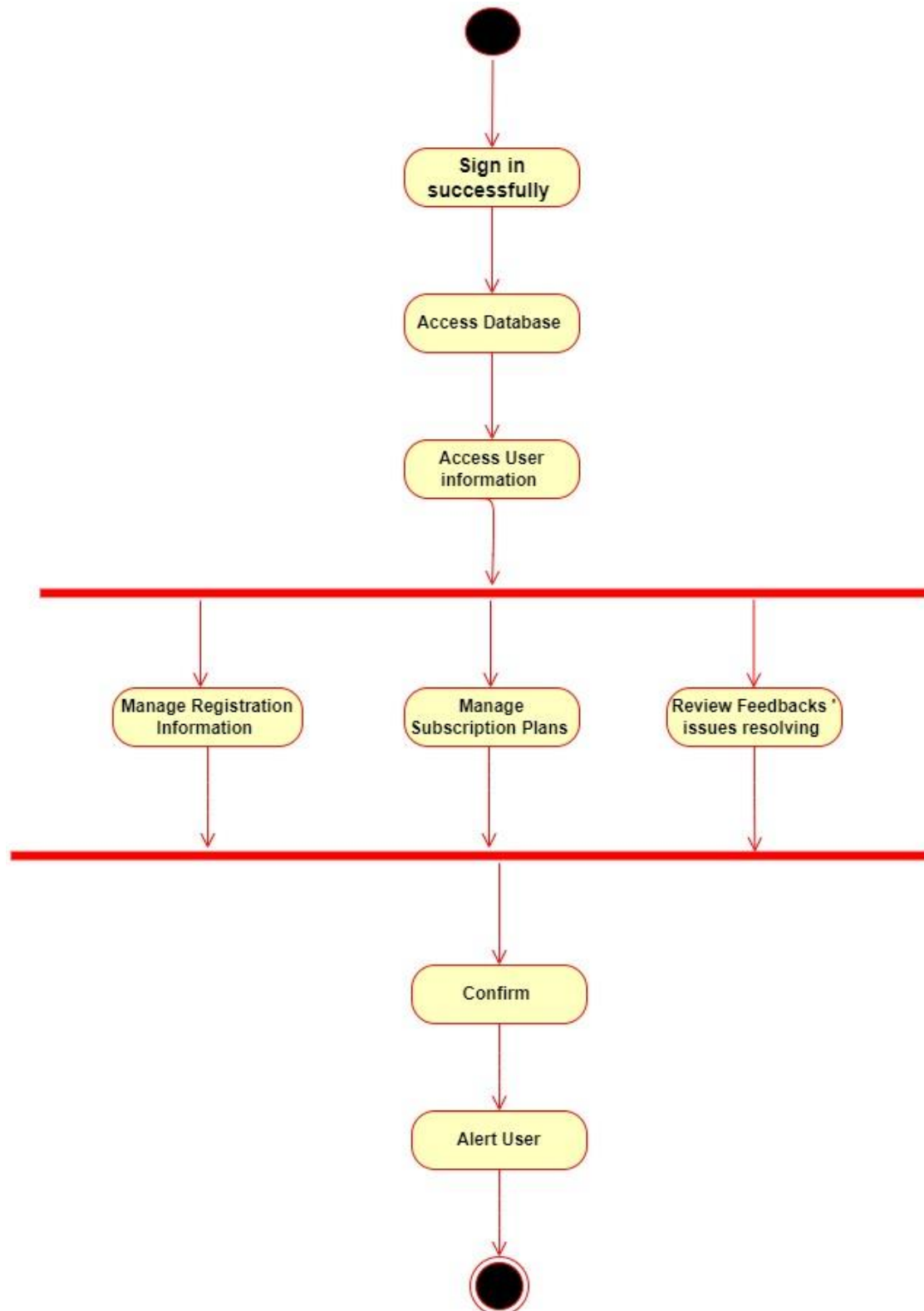## 8. Admin Management activity.

Figure 9

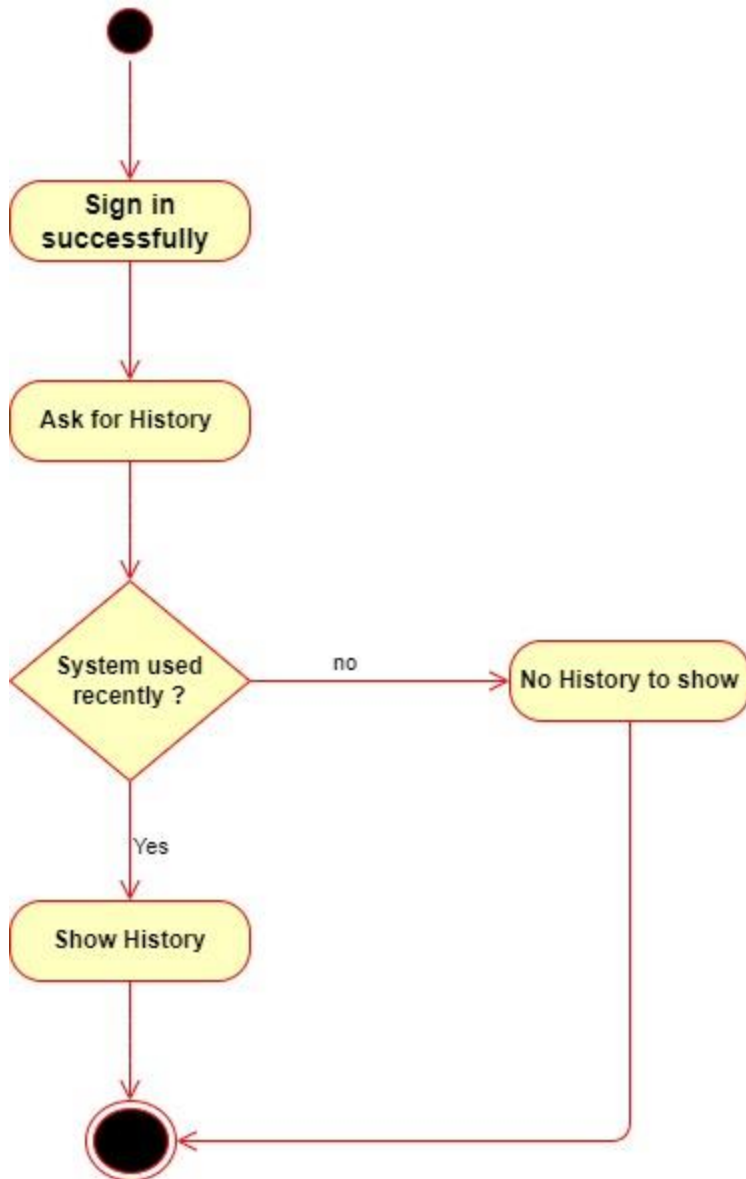**9. Show History activity.**

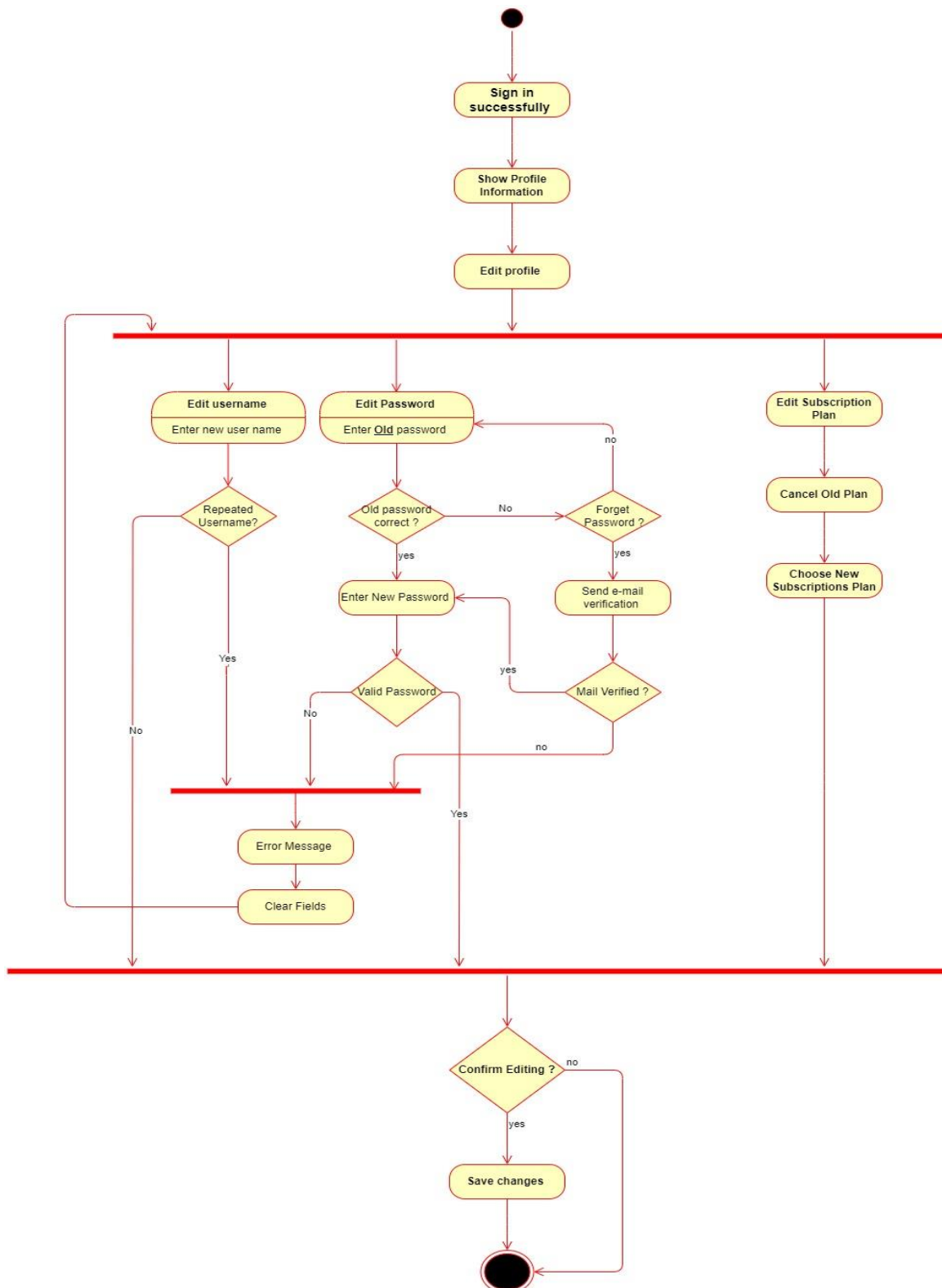Figure 10

10.   Edit Profile activity.
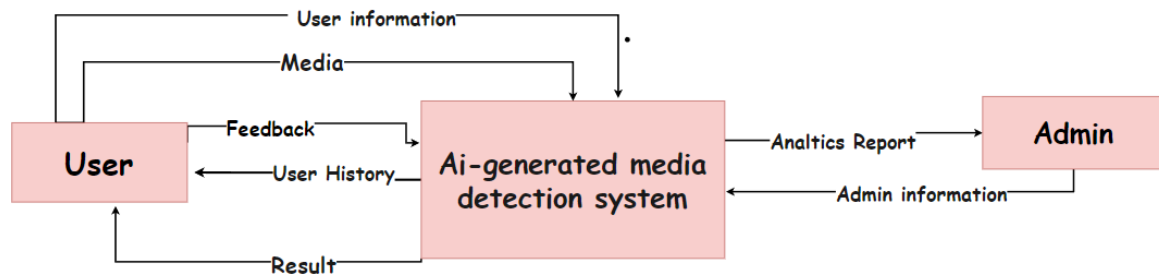

Figure 11

2.2.4 Context diagram

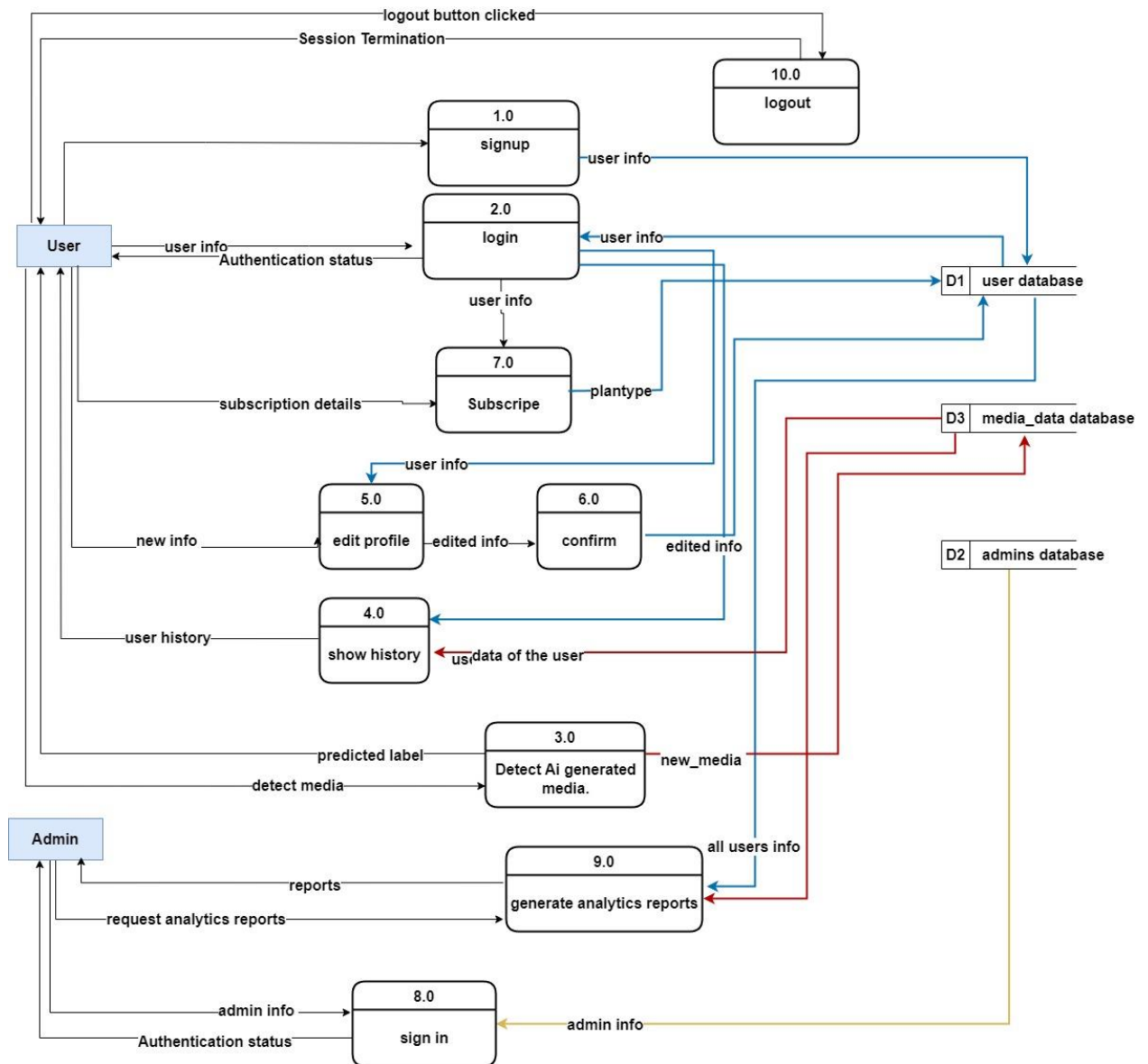Figure 12

2.2.5   Data flow diagram
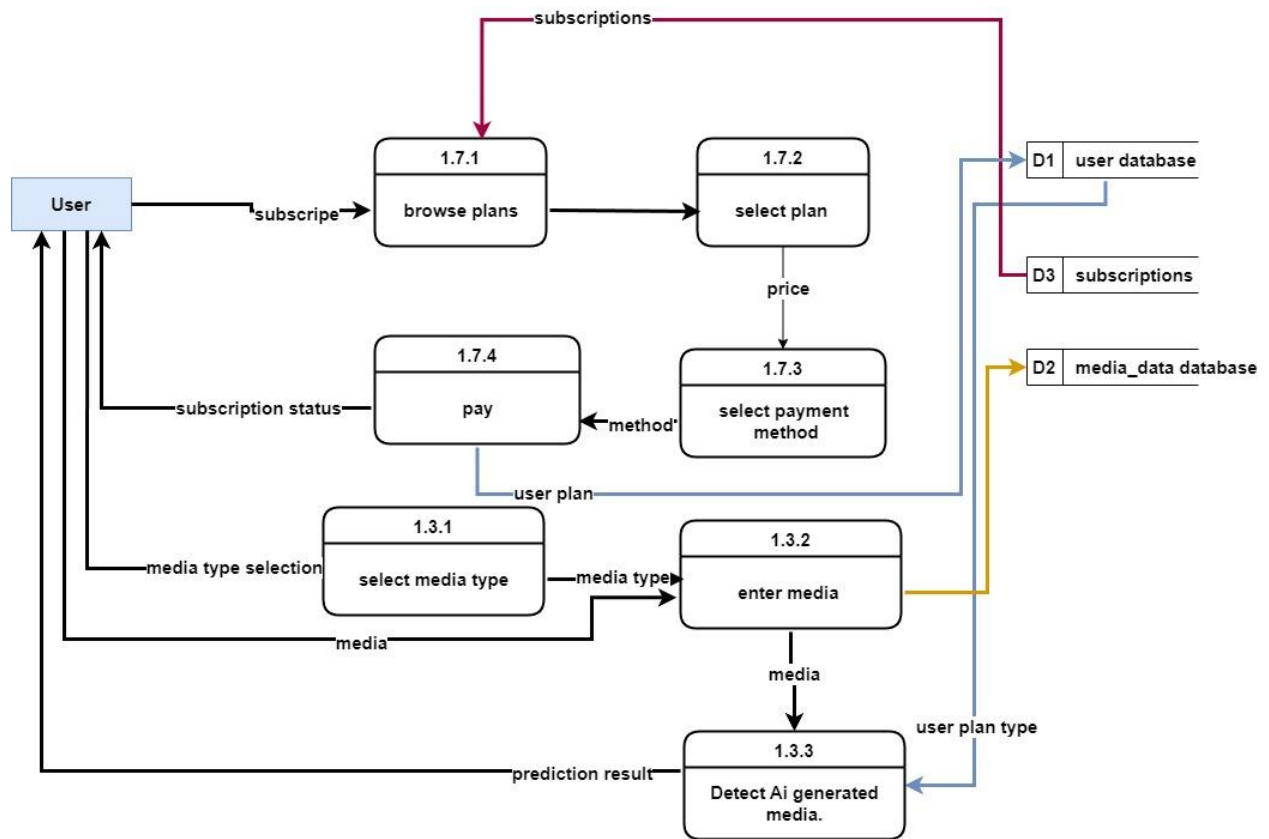
Level 0

Figure 13

Level 1

Figure 14

2.2.6 Sequence diagram
1. User
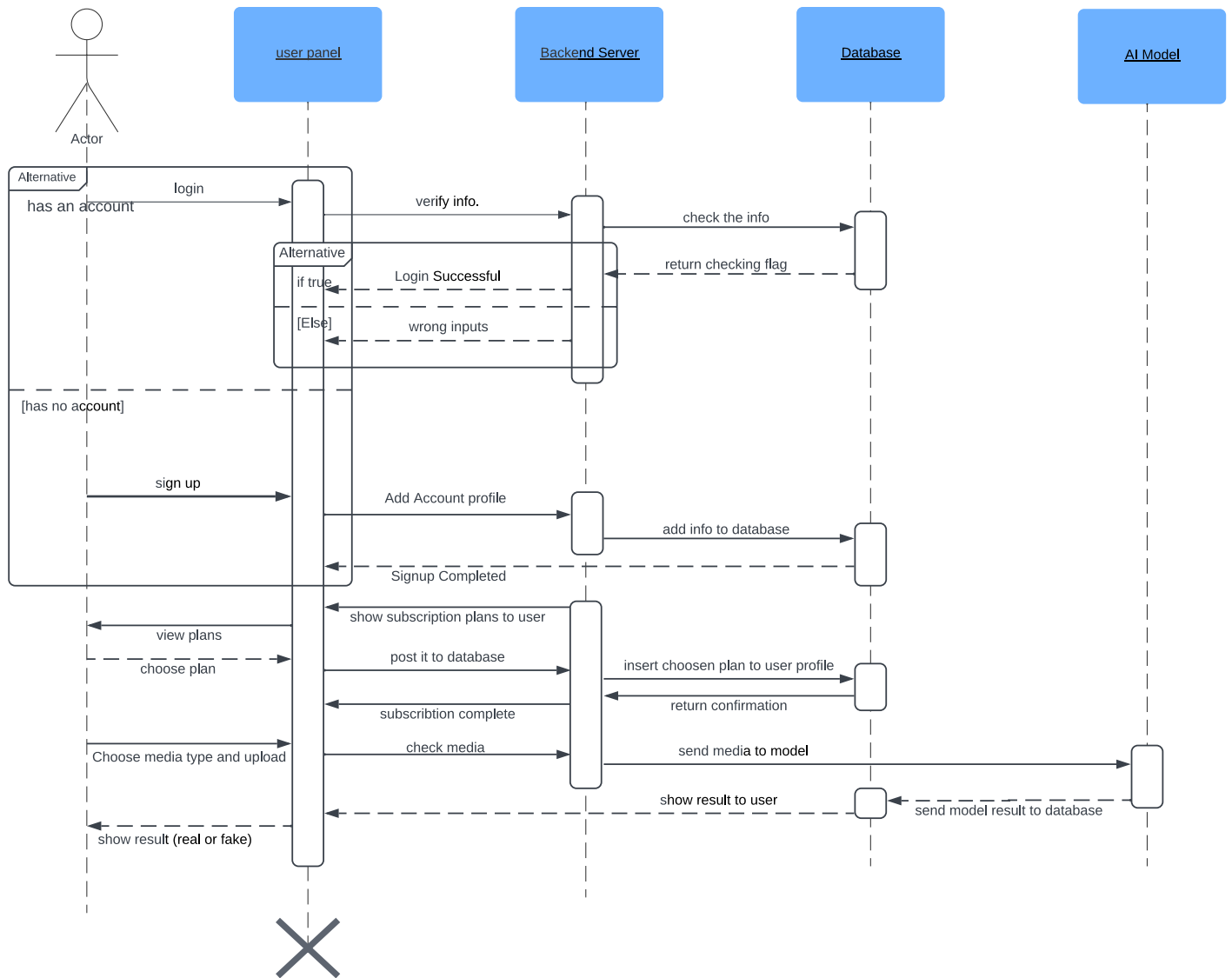
Figure 15

2. Admin
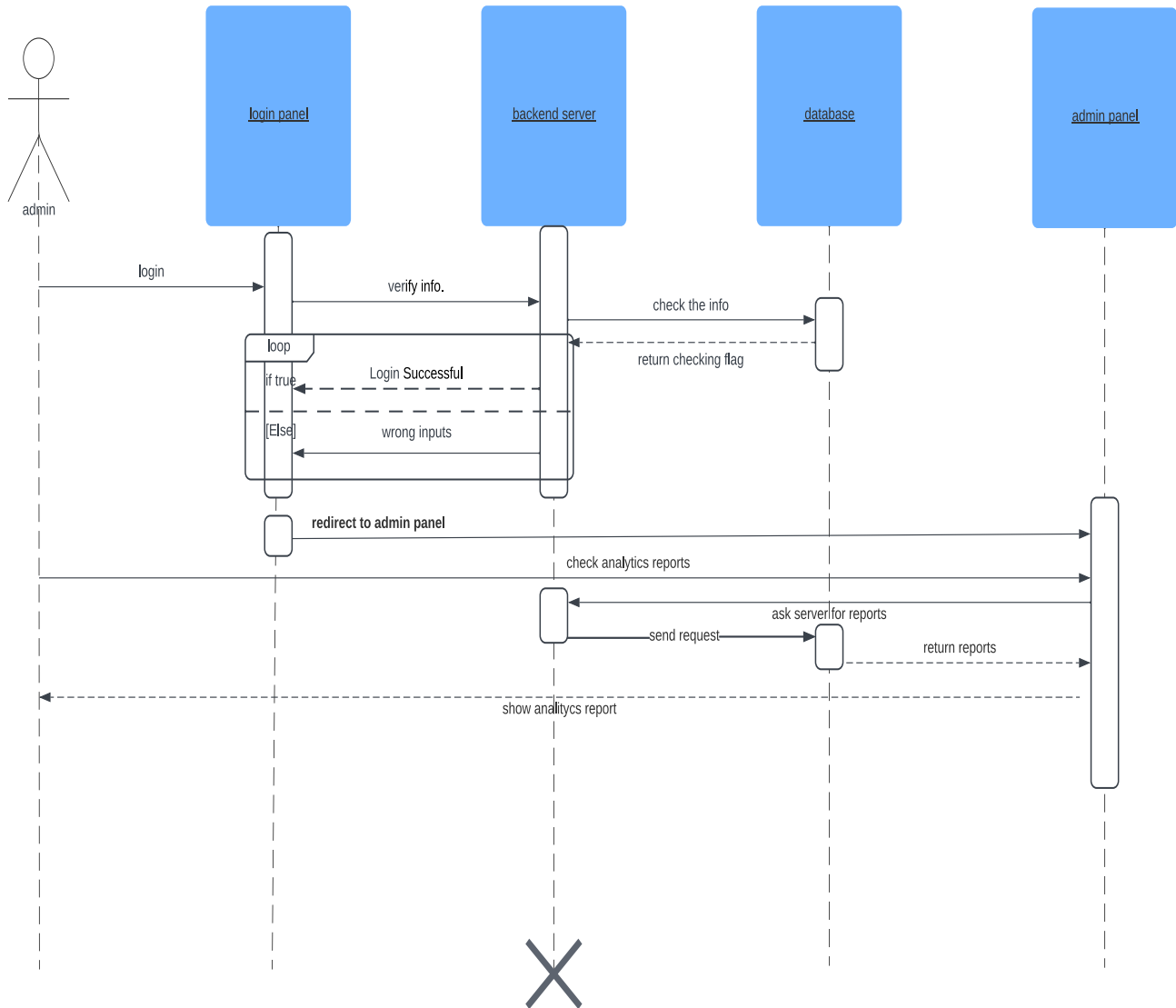
Figure 16

2.2.7   class DIAGRAM

Person
- name
- username
- Phone Number
- Email Address
- password
- age

- login(): boolean
- signup(): boolean
- logout(): void

Admin
- permission

- manage_system()
- generate_analytics_reports()

Data_History
- media_type
- media_size
- detection_result
- media_name
- attempets_time
- media

- get_history():

Supscription_Plan
- name
- price
- size_limit
- duration
- attempets_limt
- history_limit

- subscripe_plan()
- cancle_plan()

User
- plan_id
- sub_start
- subs_end
- remain_attempets
- attempets_history
- country

- upload_media() void
- submit_media_for_detection() String

AI-Generated_Media_Detector
- media_type
- media_size
- detection_result

- detect_text(): Detection_Result
- detect_images(): Detection_Result

Detection_Result
- Fake_Text
- Fake_Photo
- Real_Text
- Real_Photo

Data_Item
- data_type
- size
- data

subscribe   1   0..1

get result

show history

return

get history

Figure 16

2.2.8   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

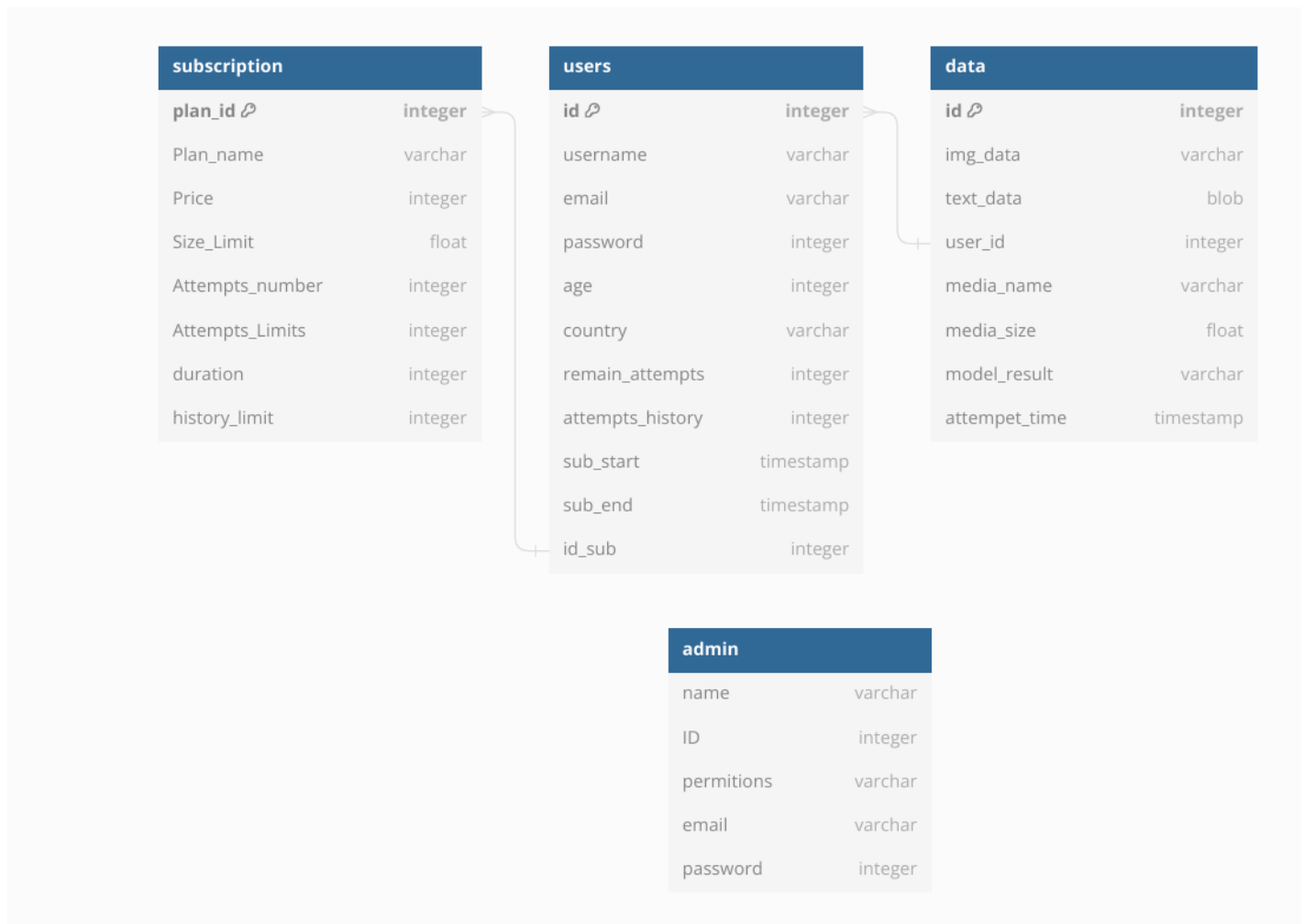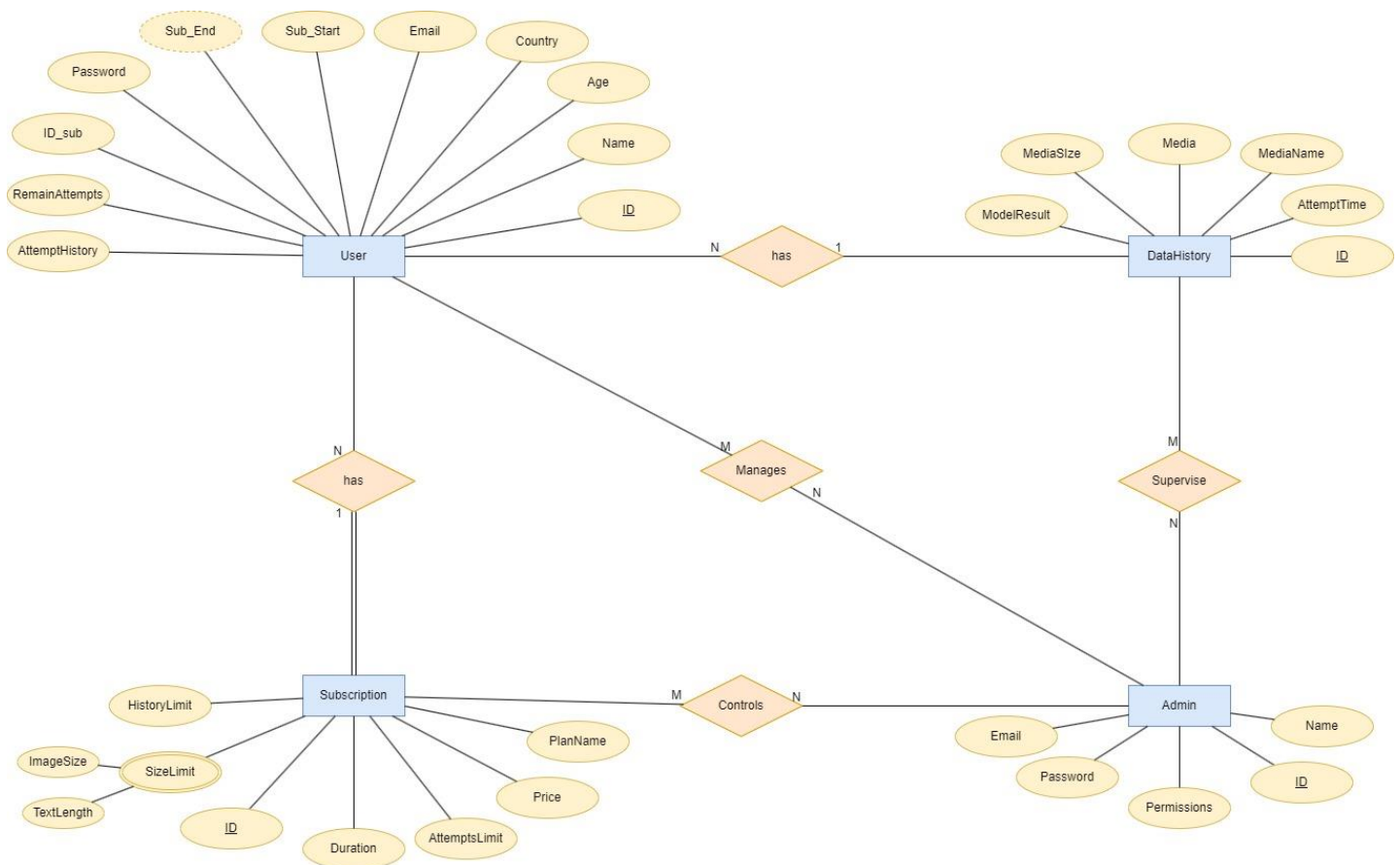|	
name	varchar
ID	integer
permitions	varchar
email	varchar
password	integer

Figure 17

2.2.9   ERD DIAGRAM

Figure 18

2.3 Used Technologies and tools.

2.3.1 Technologies

2.3.2 tools