MCS 11

ABYSS:

AI MEETS THE DARK SIDE

PAGE
01/16

SPEAKER: STEVEN TIRTADJAJA

TABLE OF CONTENTS

01 ● Introduction & Problem Statement

02 ● Project goals & expected deliverables

03 ● Representations

04 ● Software and Hardware component

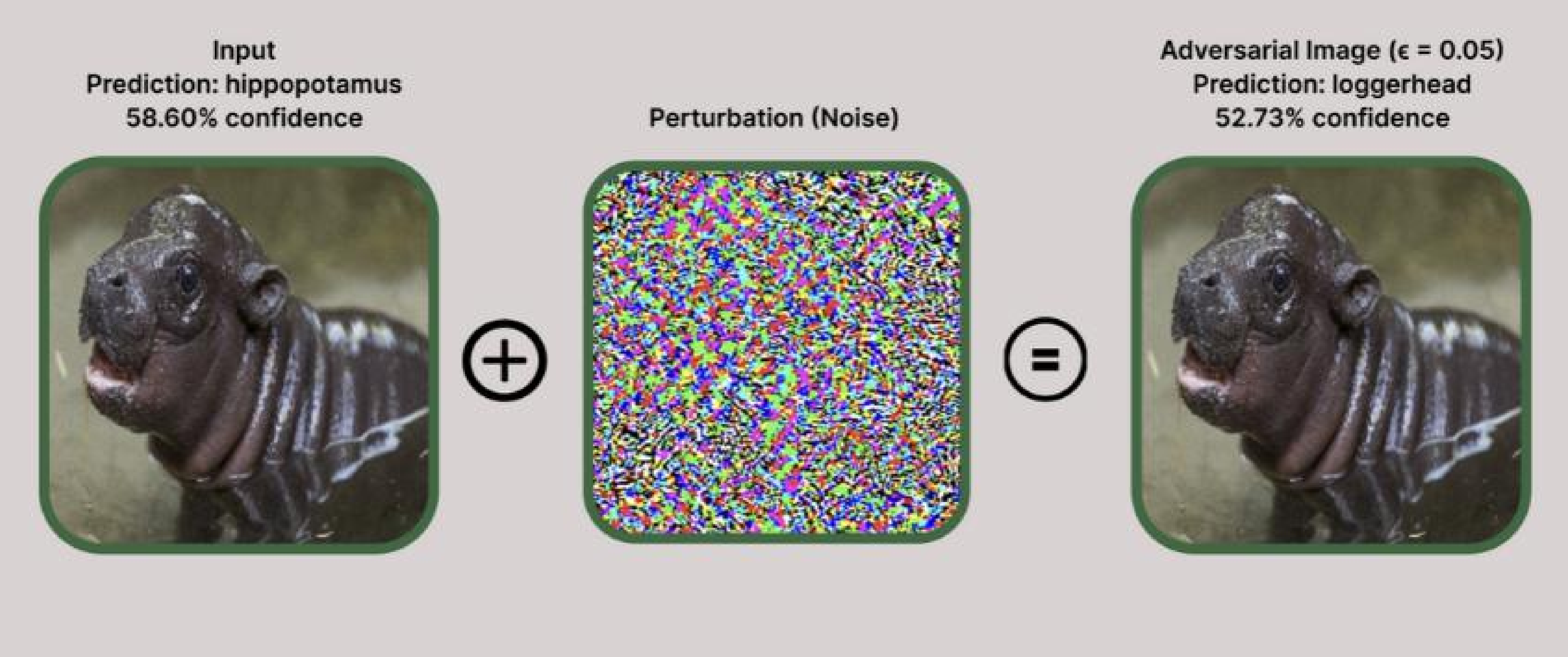
05 ● Project Management

06 ● Progress and Plans

INTRODUCTION

- Rapid AI advancements
- Vulnerabilities in Image Classification AI Models
- Fast Gradient Sign Method

FAST GRADIENT SIGN METHOD



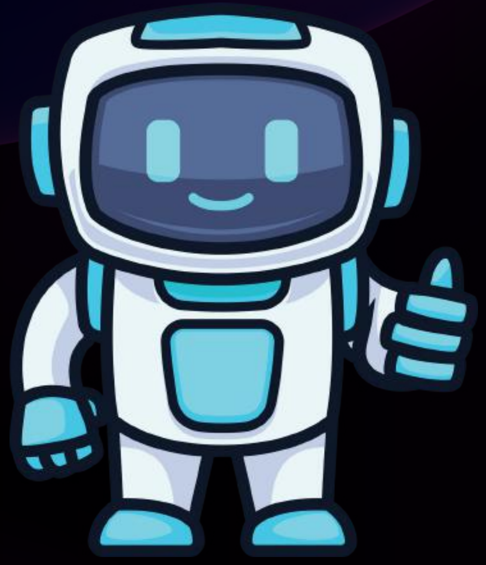
Input
Prediction: hippopotamus
58.60% confidence

Perturbation (Noise)

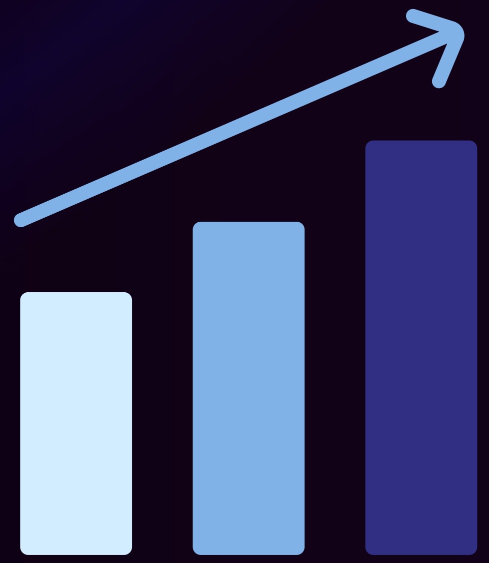
Adversarial Image (ϵ = 0.05)
Prediction: loggerhead
52.73% confidence

SPEAKER: STEVEN TIRTADJAJA

PROBLEM STATEMENT



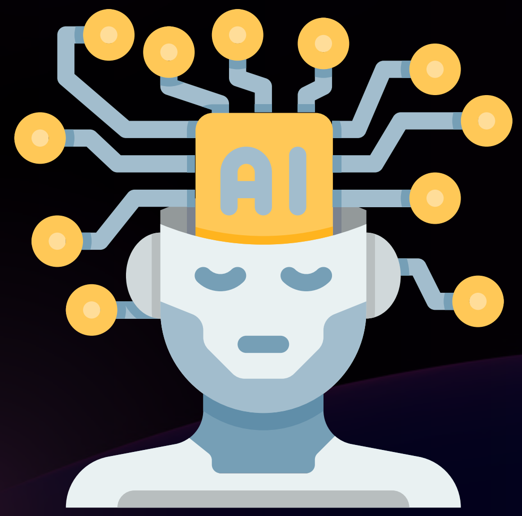
Rapid advancements of AI have uncovered many vulnerabilities especially in image classification AI models



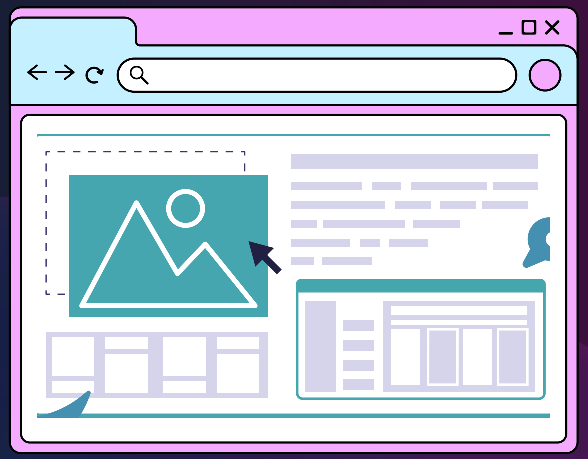
SPEAKER: STEVEN TIRTADJAJA

PROJECT GOALS AND DELIVERABLES

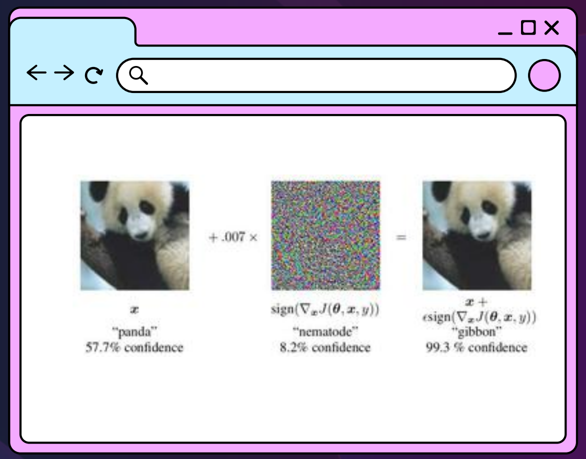
- Examine the vulnerabilities of image classification AI models.
- Raise awareness of AI weaknesses
- Demonstrate FGSM effectiveness



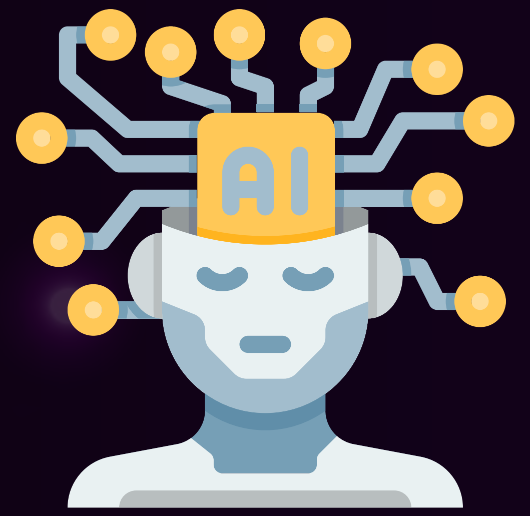
A functional web application to apply FGSM attacks



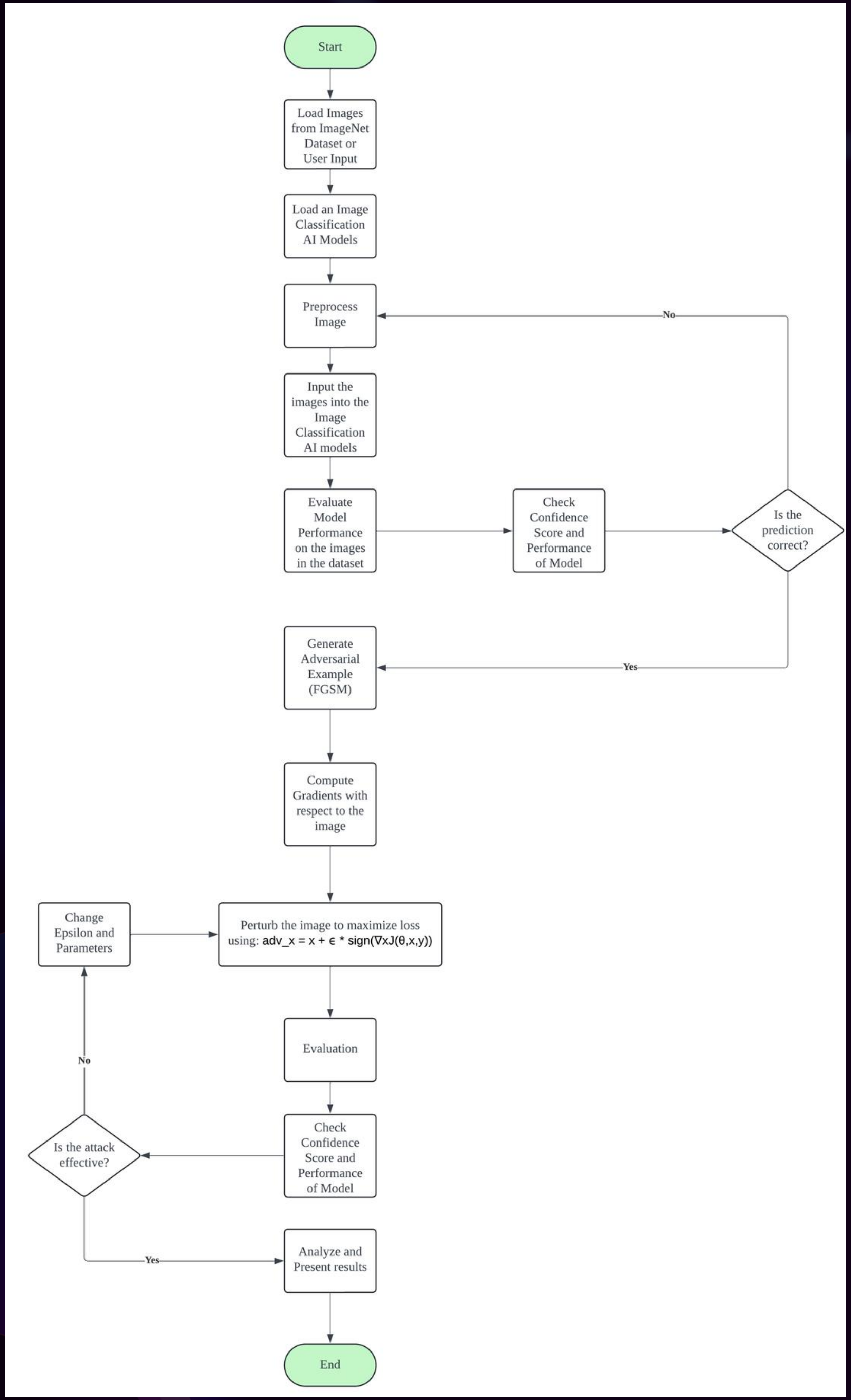
Visual side by side comparisons of original and adversarial images.


+ .007 ×
=
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Evaluate the impact of the adversarial attack



REPRESENTATIONS: FLOW CHART



Start

Load Images from ImageNet Dataset or User Input

Load an Image Classification AI Models

Preprocess Image

Input the images into the Image Classification AI models

Evaluate Model Performance on the images in the dataset

Check Confidence Score and Performance of Model

Is the prediction correct?

No

Yes

Generate Adversarial Example (FGSM)

Compute Gradients with respect to the image

Perturb the image to maximize loss using: $adv\_x = x + \epsilon * sign(\nabla x J(\theta,x,y))$

Change Epsilon and Parameters

Evaluation

Check Confidence Score and Performance of Model

Is the attack effective?

No

Yes

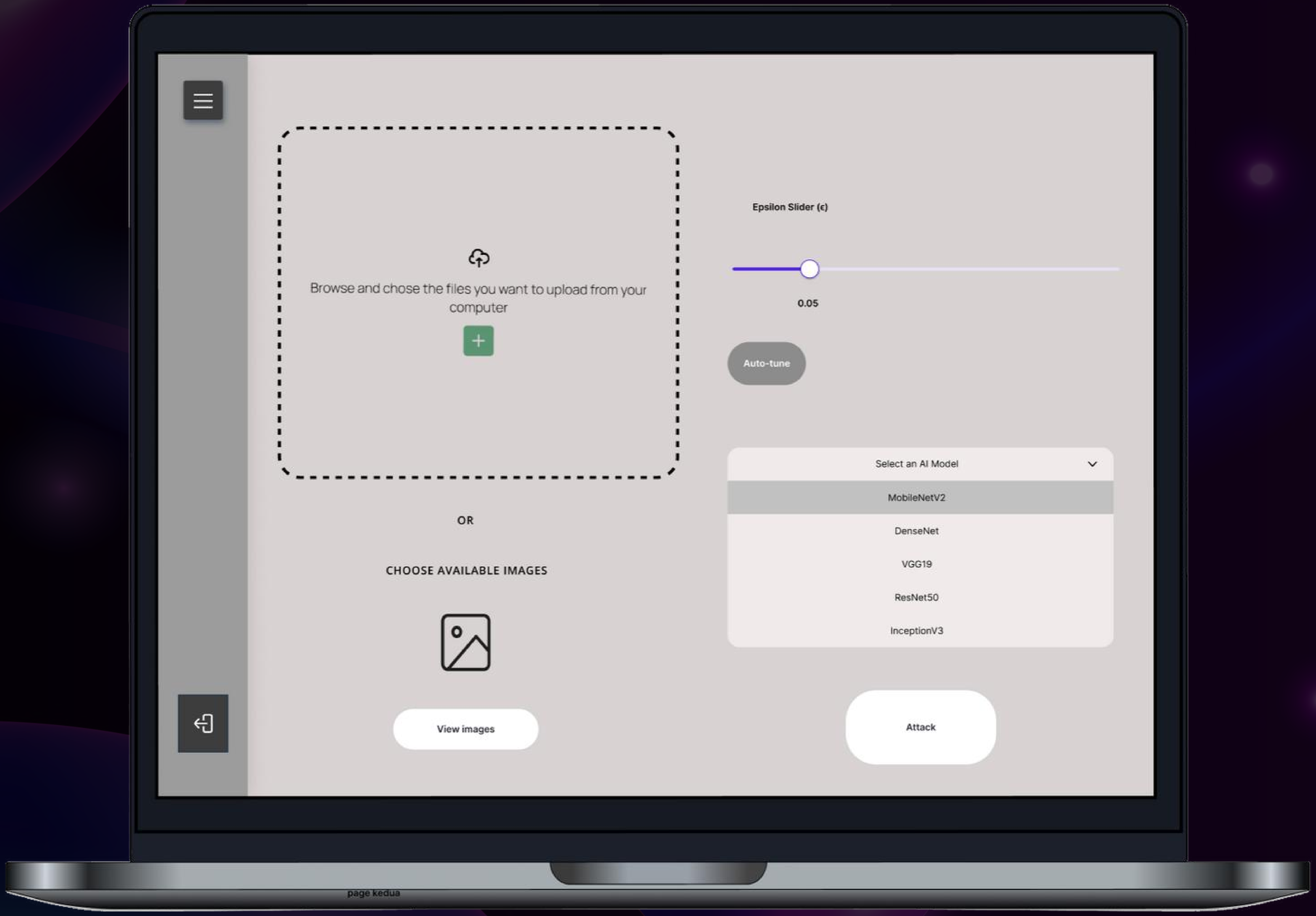
Analyze and Present results

End

SPEAKER: SHENG XIN CHUA

REPRESENTATIONS: USER INTERFACE

SPEAKER: SHENG XIN CHUA



Browse and chose the files you want to upload from your computer

OR

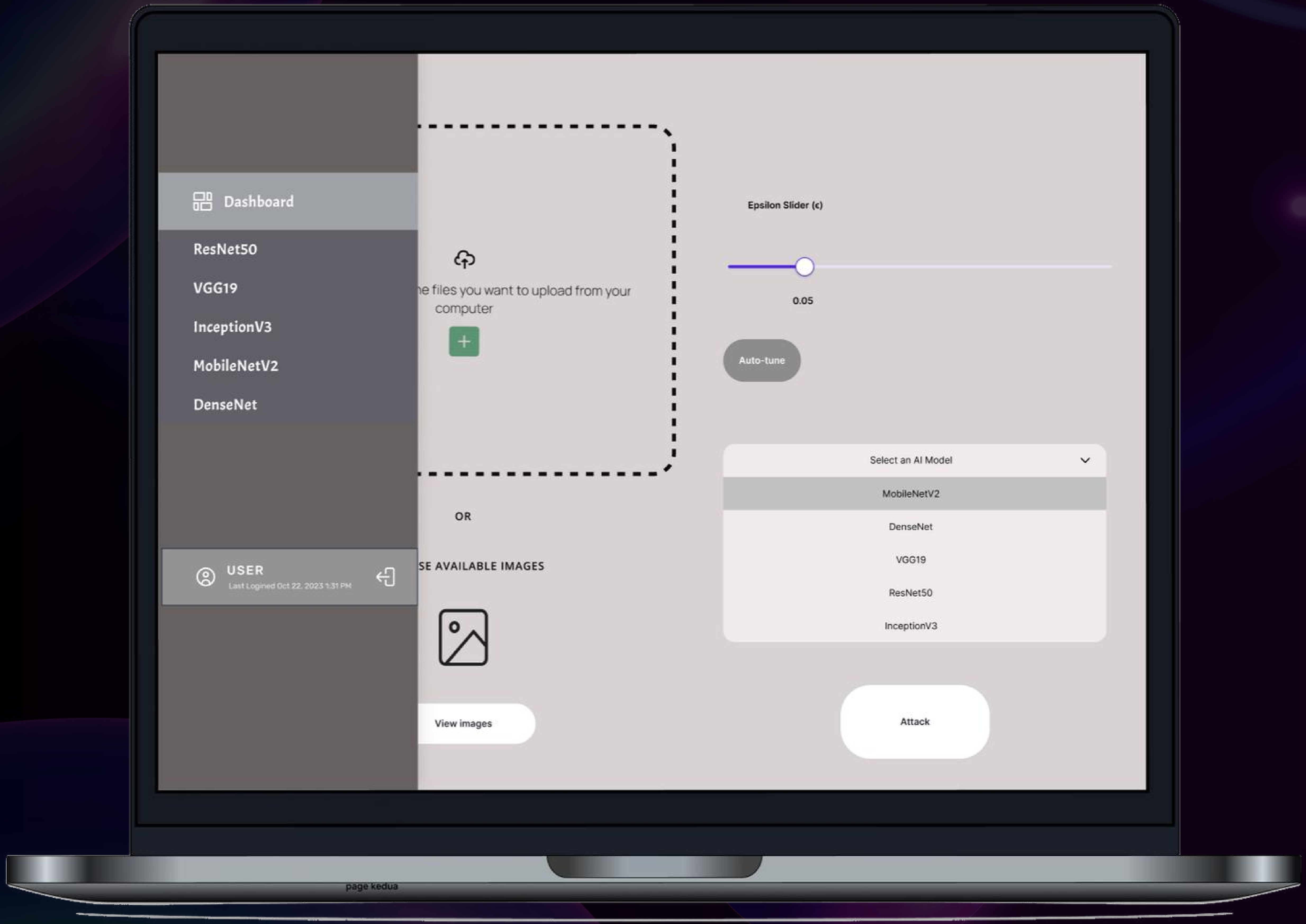
CHOOSE AVAILABLE IMAGES

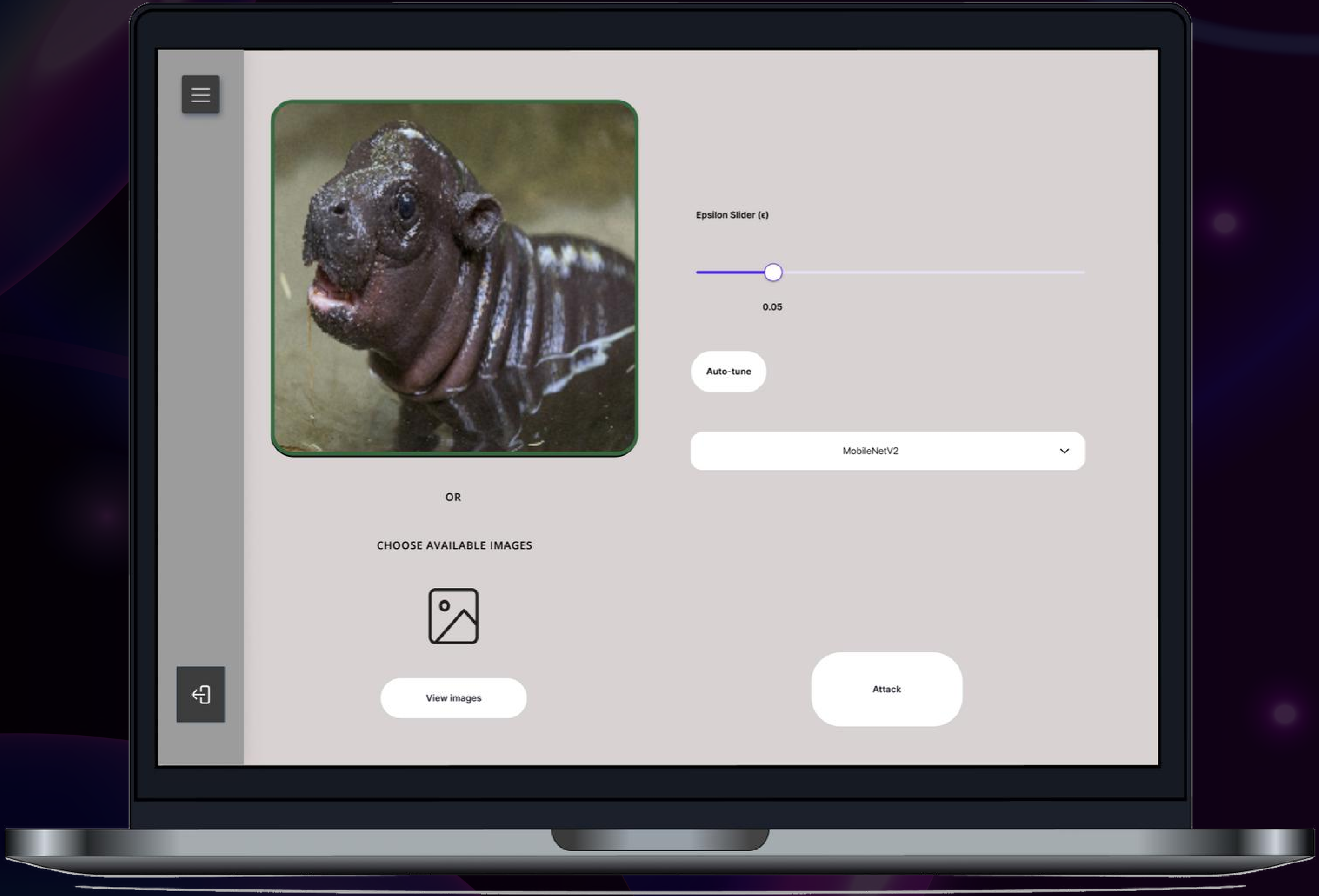
Epsilon Slider (ε)

0.05

Select an AI Model

SPEAKER: SHENG XIN CHUA

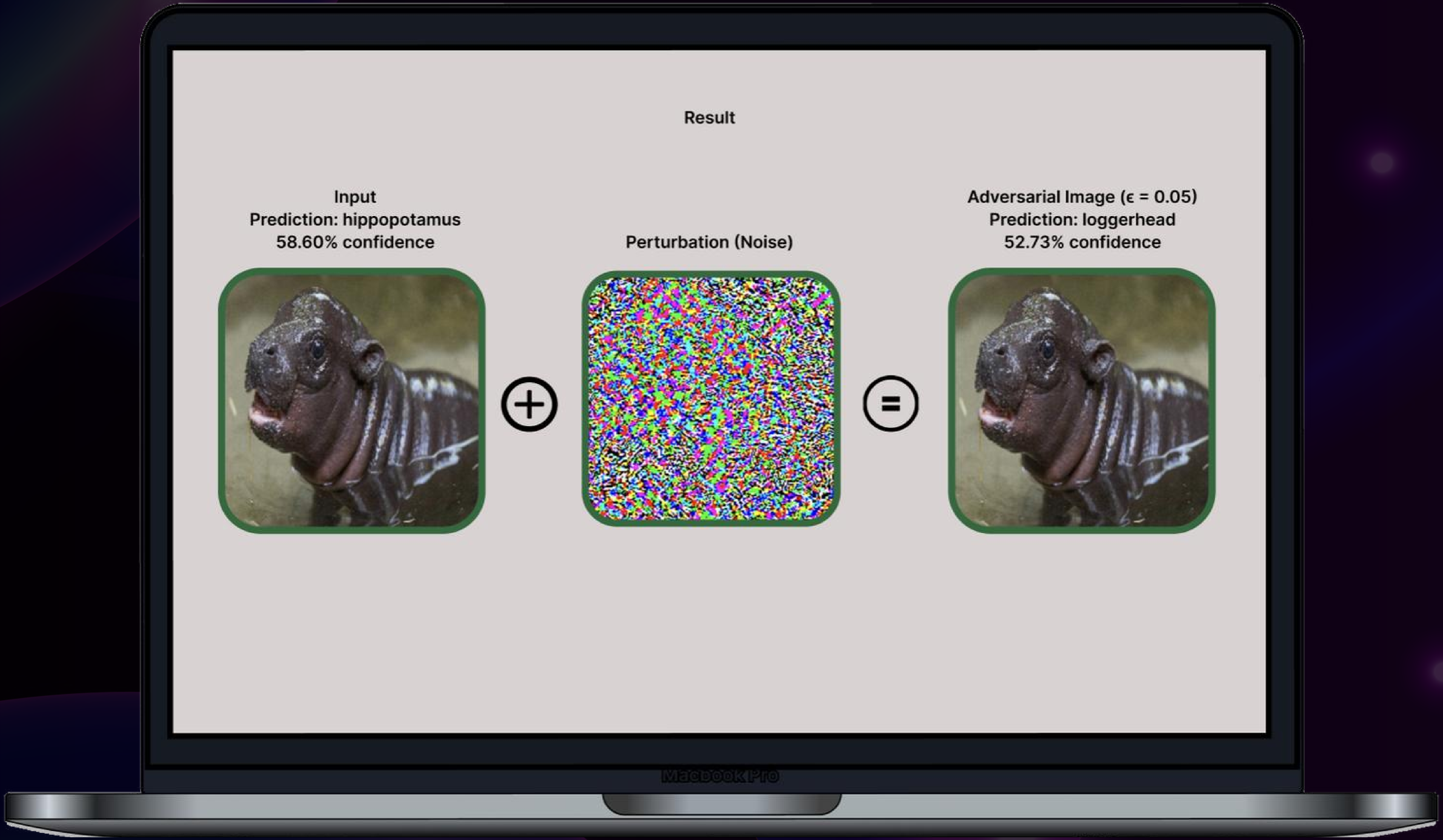

Dashboard
ResNet50
VGG19
InceptionV3
MobileNetV2
DenseNet
USER
Last Logined Oct 22, 2023 1:31 PM
e files you want to upload from your computer
+
OR
SE AVAILABLE IMAGES
View images
Epsilon Slider (ϵ)
0.05
Auto-tune
Select an AI Model
MobileNetV2
DenseNet
VGG19
ResNet50
InceptionV3
Attack
page kedua



Epsilon Slider (ε)

0.05

MobileNetV2



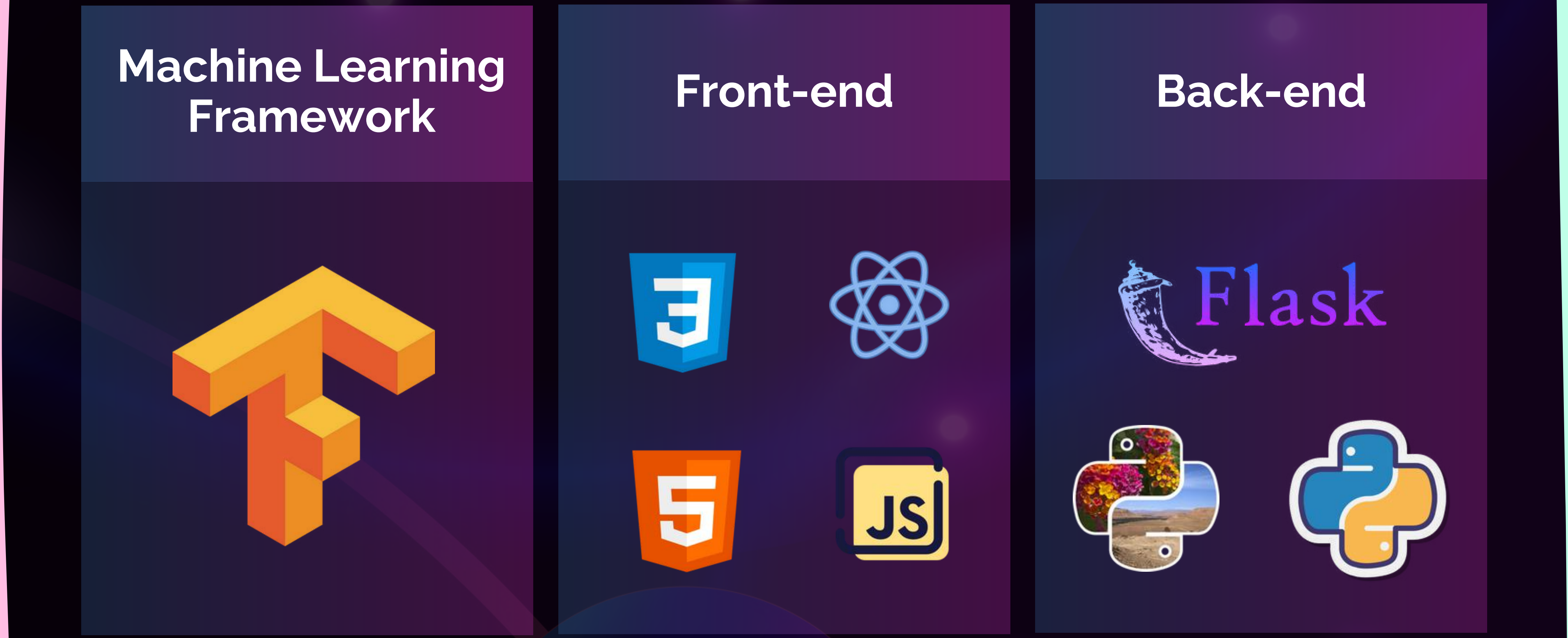
Result

Input
Prediction: hippopotamus
58.60% confidence

Perturbation (Noise)

Adversarial Image (ε = 0.05)
Prediction: loggerhead
52.73% confidence

SOFTWARE COMPONENT



Machine Learning Framework

Front-end

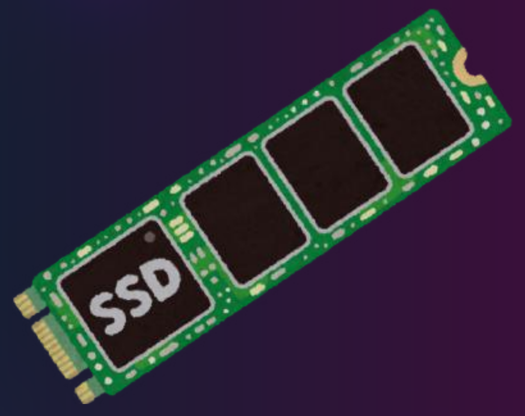
Back-end

SPEAKER: ALVIN ANDREAN

HARDWARE COMPONENT

Storage

512GB SSD or more



External Storage

External SSD

CPU

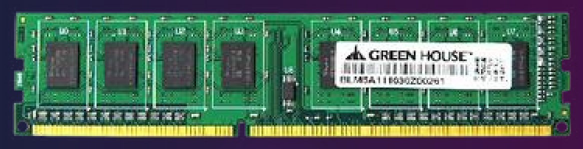
Intel Core i7 or later / AMD Ryzen 7 or later

GPU

NVIDIA® RTX Series GPU

RAM

16GB or more



PROJECT MANAGEMENT

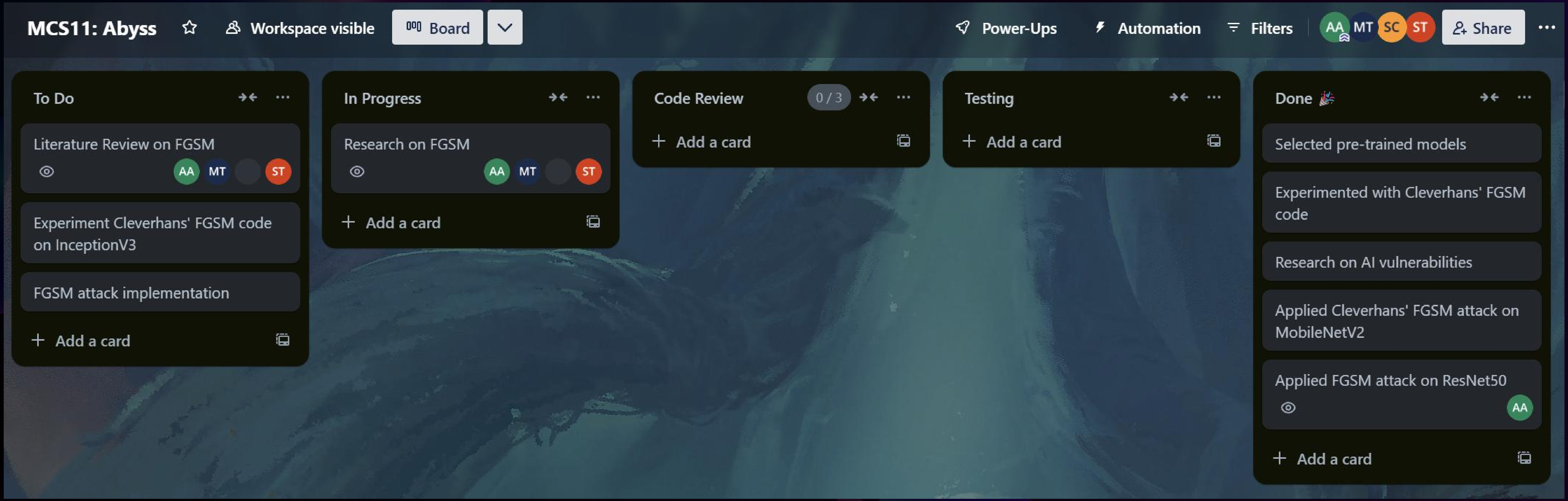
Google Drive

- Cloud storage and accessibility
- Large storage capacity
- Real time collaboration

Github

- Track changes
- Repository management
- Ease of collaboration and integration with Visual Studio Code

HOW?



MCS11: Abyss

To Do
- Literature Review on FGSM (AA, MT, ST)
- Experiment Cleverhans' FGSM code on InceptionV3
- FGSM attack implementation

In Progress
- Research on FGSM (AA, MT, ST)

Code Review

Testing

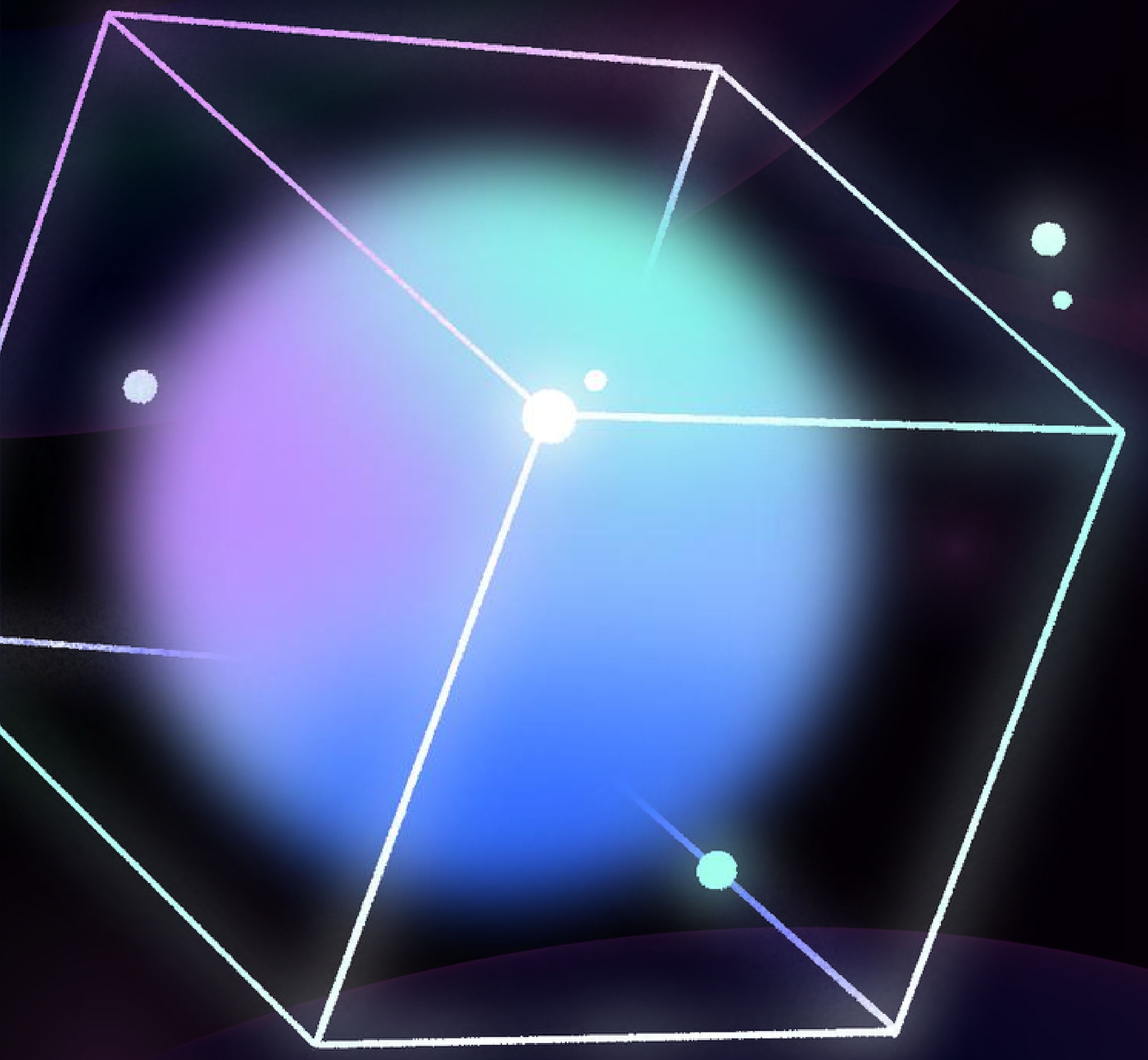
Done
- Selected pre-trained models
- Experimented with Cleverhans' FGSM code
- Research on AI vulnerabilities
- Applied Cleverhans' FGSM attack on MobileNetV2
- Applied FGSM attack on ResNet50 (AA)

KANBAN BOARD

WHY?

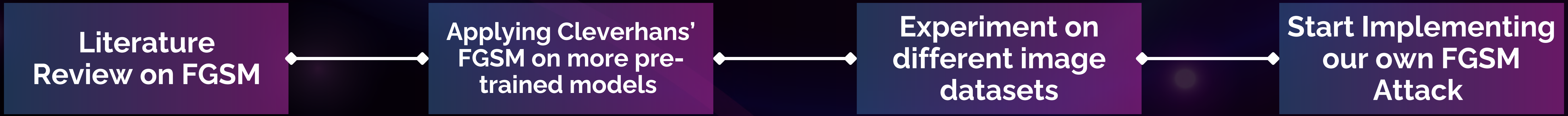
- Improved communication and collaboration
- Clear workflow visualization
- Supports Incremental Progress

CURRENT PROGRESS



Research on AI Adversarial Attack Techniques	Completed initial research on AI vulnerabilities, focusing on image classification models and adversarial attack techniques.
Attack Technique Selection	FGSM was chosen for generating adversarial examples in image classification models
Model Selection	Selected pre-trained models from TensorFlow's library
Applied Cleverhans FGSM Attack	Experimented with Cleverhans' FGSM attack on two pre-trained models (MobileNetV2 and ResNet50)

PLANS



Literature Review on FGSM

Applying Cleverhans' FGSM on more pre-trained models

Experiment on different image datasets

Start Implementing our own FGSM Attack

THANK YOU!

MCS11