



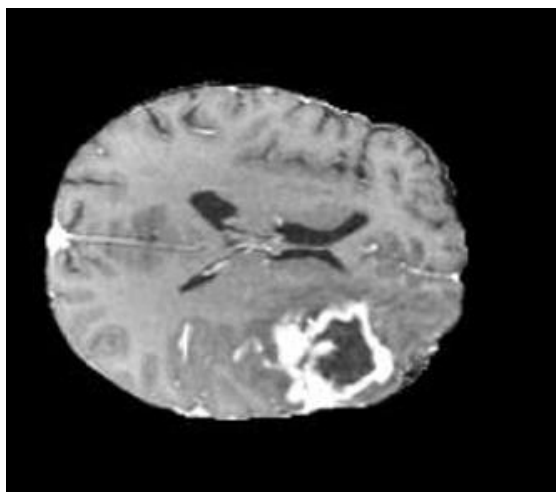
Semmelweis Egyetem



Egészségügyi
Menedzserképző
Központ

Amikor a jó AI nem jelent biztonságos AI-t

Ellenséges támadások hatása orvosi képfeldolgozó
AI-rendszerekben

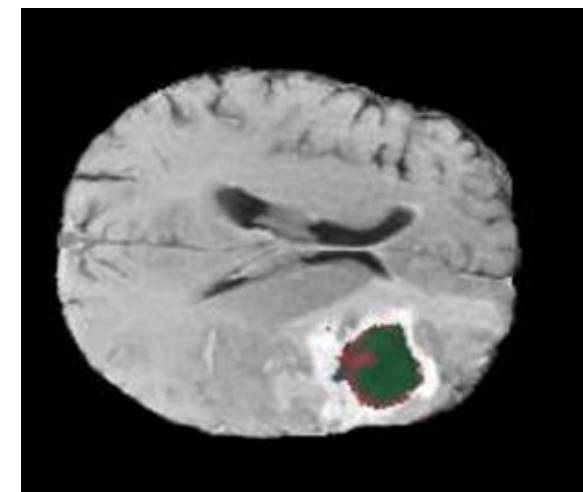


Szabó Barbara

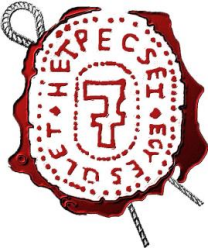
Témavezető: Dr. Palicz Tamás

Semmelweis Egyetem, Egészségügyi Menedzserképző Központ

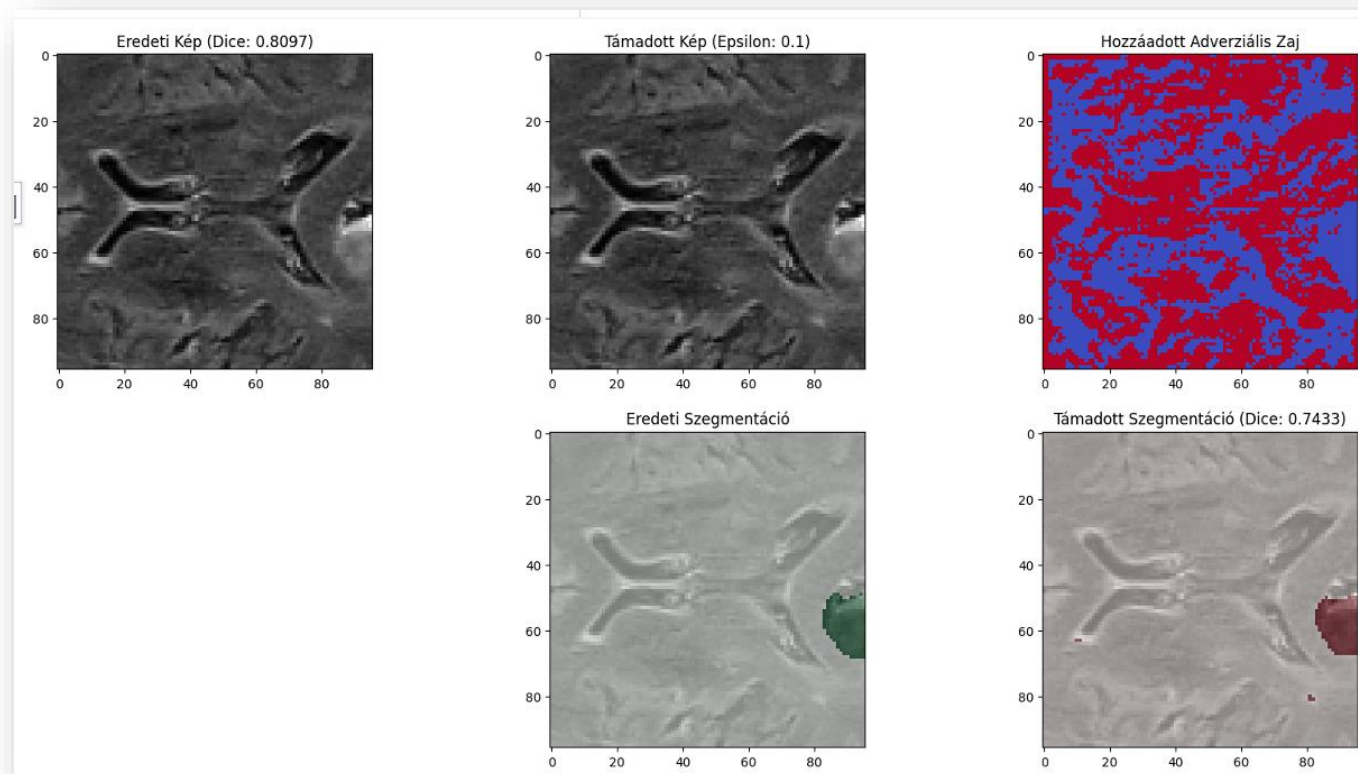
2026. Szeptember 16.



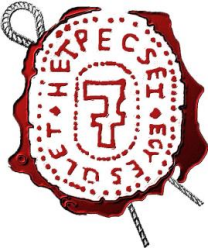
Az év információbiztonsági dolgozata 2026 – Diplomadolgozat kategória



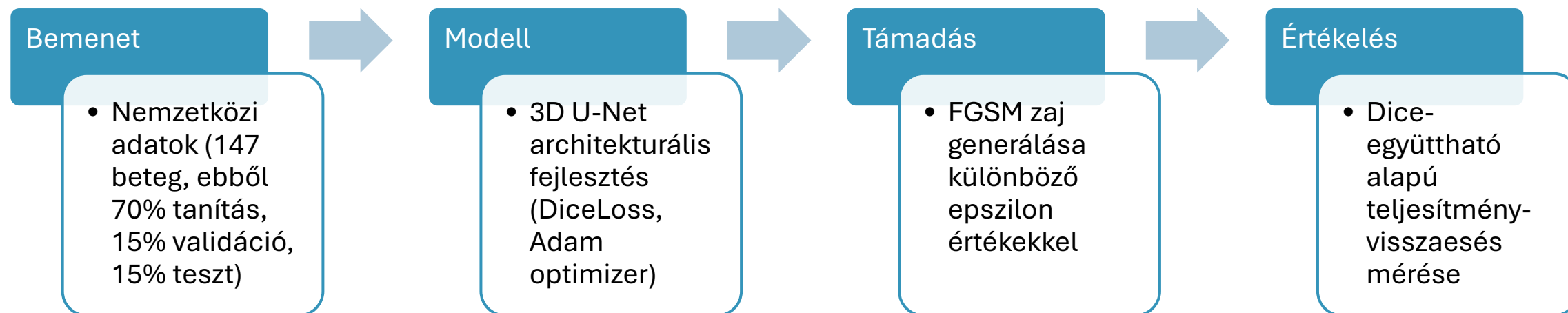
A kép szinte ugyanaz. Az AI döntése már nem.



Eredménykarcolatok a saját kísérletemen: a 3D U-Net modell átlagos Dice-értéke a UPenn-GBM tesztalmazon.



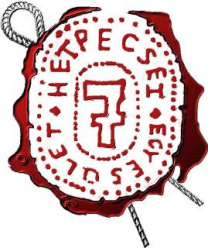
Mit vizsgáltam?



Módszertani korlátok

Tudatos kutatási keretek: A kísérlet szándékosan egyetlen adatbázisra, egyetlen architektúrára és egyetlen támadási típusra korlátozódott.

Indoklás: A cél az általános sebezhetőségi mechanizmus demonstrálása volt egy kontrollált környezetben.



Először: működött a modell?

1. Lépés: Google Drive csatlakozta

=====
Átlagos Teszt Dice Score: 0.8230
Százalékos pontosság: 82.30%
=====

```
from google.colab import drive
import os
import random
import numpy as np
import torch
import nibabel as nib
from torch.utils.data import Dataset, DataLoader
import matplotlib.pyplot as plt
```



```
# Drive csatlakoztatása
drive.mount('/content/drive')
```

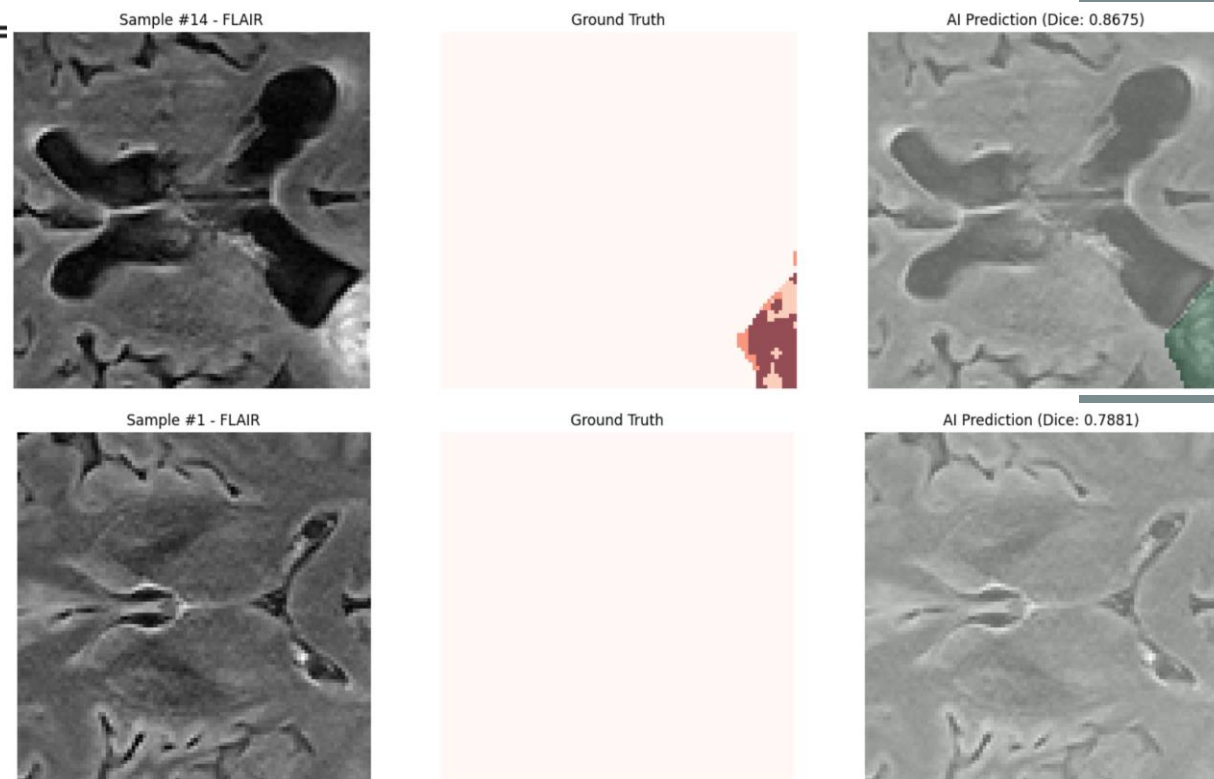
```
# Beállítások - Módosítsd az utat a sajátodra!
```

```
CSV_PATH = '/content/drive/MyDrive/brain_tumor_database/upenn_gbm_pairs.csv'
```

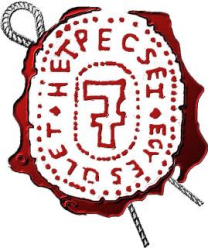
```
DEVICE = torch.device("cuda" if torch.cuda.is_available() else "cpu")
```

```
print(f"Használt eszköz: {DEVICE}")
```

```
Mounted at /content/drive
Használt eszköz: cuda
```

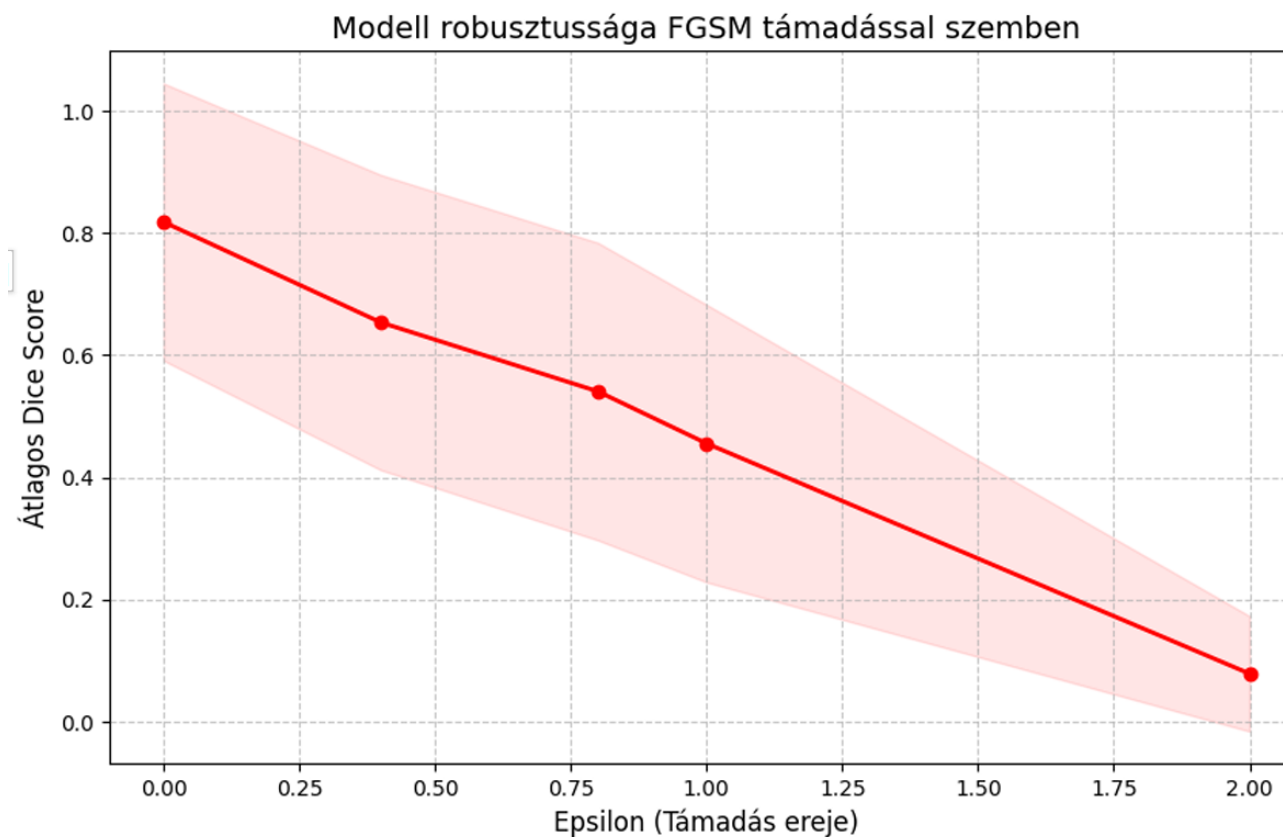


Kódrészlet és eredménykarcolatok a saját kísérletemen: a 3D U-Net modell átlagos Dice-értéke a UPenn-GBM tesztalmon.

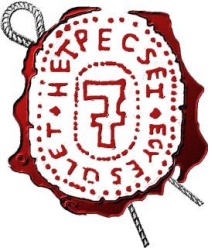


Aztán megtámadtuk.

A támadás hatása: már alacsony perturbáció esetén is érdemi romlás, míg erősebb szinteken a Dice-érték szétesik.



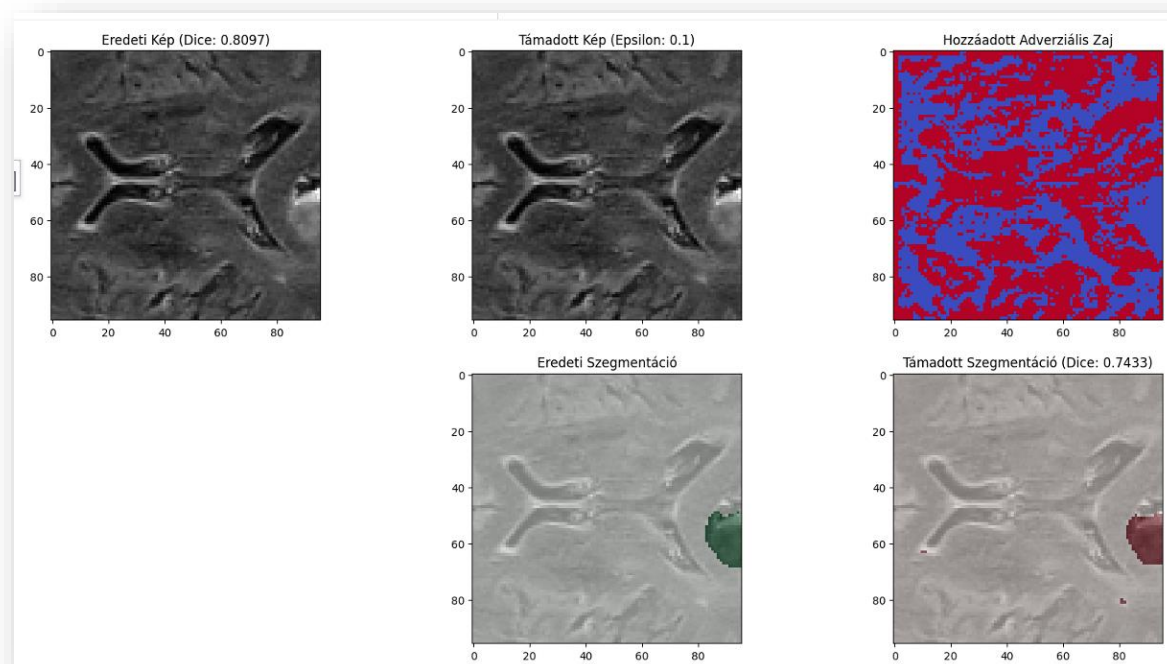
A diagram a saját kísérletem eredménye: a 3D U-Net modell átlagos Dice-értéke a tiszta és az FGSM-mel perturbált UPenn-GBM tesztalmazon.



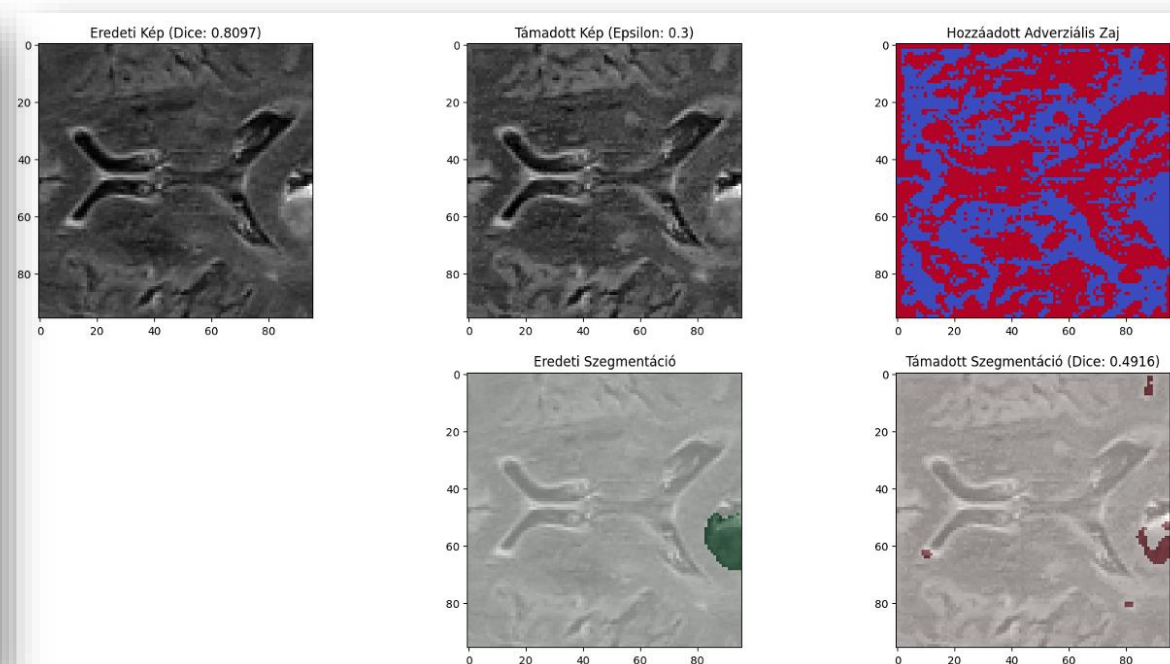
Miért információbiztonsági probléma?

Vizuális, drasztikus bizonyíték:

Manipulált bemenet → Hibás AI-kimenet → Hibás döntéstámogatás → Betegbiztonsági kockázat

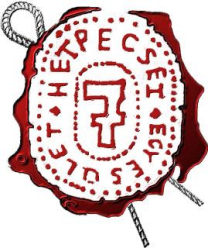


U-PennGBM MRI felvételen tumordetektálás és ellenséges modell támadás (epszilon=0,1)



U-PennGBM MRI felvételen tumordetektálás és ellenséges modell támadás (epszilon=0,3)

A két ábra a saját kísérletem eredménye: a 3D U-Net modell átlagos Dice-értéke a tiszta és az FGSM-mel perturbált UPenn-GBM tesztalmon.



Mit jelent ez és mit tehetünk?

Pontosság $\neq$ Robusztusság $\neq$ Biztonság

Rendszerszintű védelem

Strukturált klinikai kockázatkezelés,
ellenséges támadási tréning beépítése és
bemeneti anomália-szűrés

Techhnikai iteráció

Iteratív támadások (PGD), black-box
és multimodális támadások bevonása
Több architektúra tesztelése

Radiológus szerepe

Esetek priorizálása, ellenőrzése
Ellentmondásos esetek kiszűrése
Adat- és modellminőség
visszacsatolása



A jó AI nem feltétlenül biztonságos AI.

Az orvosi AI-t nemcsak arra kell tesztelnünk, hogy mit tud, hanem arra is, hogy hogyan hibázik.

Köszönöm a figyelmet!

Az orvosi AI sebezhetősége: Láthatatlan támadások az agydaganat-diagnosztikában



Tiszta kép

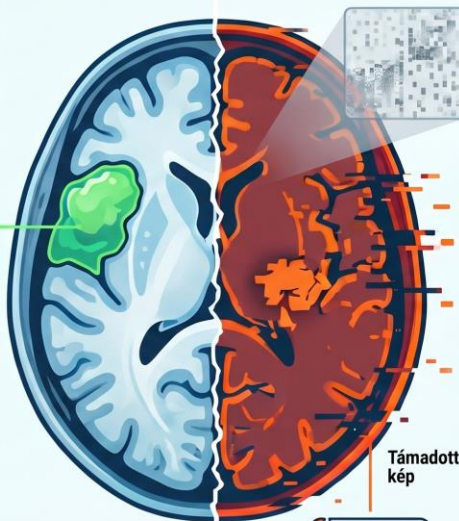
AI Pontosság
(Dice-mutató):
0,818 (81,8%)



Nagy pontosságú
teljesítmény



Kiváló teljesítmény
tisza adatokon



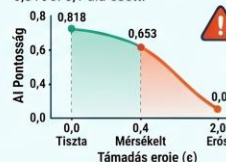
AI Pontosság (Dice-mutató):
< 0,1
(7,8% erős támadásnál)

Pixelszintű manipuláció

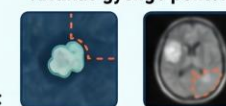
Az emberi szem számára láthatatlan zaj, amely drasztikusan megváltoztatja az AI kimenetét.

80%-os teljesítményromlás

Erős támadás esetén a tumorfelismerés pontossága (Dice-mutató) 0,818-ról 0,1 alá esett.



Kritikus gyenge pontok



Kis méretű
tumorkok
Rosszabb képminőségű
felvételek
A legsebezhetőbbek.



Megoldások és Betegbiztonság



EU AI Act megfelelés

Az orvosi AI magas kockázatú rendszer, ezért kötelező a robusztussági és kiberbiztonsági tesztelés.



Védelmi stratégiák

Szükséges az „ellenséges alapú tanítás” és a modellek folyamatos biztonsági monitorozása.



Kötelező emberi felügyelet

Az AI kimenetét minden esetben szakorvosi kontrollnak kell alávetni a téves diagnózis elkerülésére.

