פרויקט סיום · קורס מולטימדיה ולמידת מכונה

זיהוי תמונות מזויפות שנוצרו בבינה מלאכותית

מערכת RealOrFake

98.28%	0.9986
דיוק (Accuracy) על סט הוולידציה	שטח מתחת לעקומת ROC (AUC)

מגישים: אדיר שלומוב, פלורה

מאי 2026

1. רקע ובעיה

בשנים האחרונות חלה עלייה דרמטית ביכולת של כלים מבוססי בינה מלאכותית — בהם Midjourney, DALL·E 3 ו-Stable Diffusion — לייצר תמונות פוטוריאליסטיות שכמעט בלתי אפשר להבחין בהן מתמונות אמיתיות בעין האנושית. מגמה זו מהווה איום ממשי על אמינות המידע החזותי ברשת: הפצת חדשות כזב, זיוף ראיות, יצירת זהויות מזויפות ברשתות חברתיות ועוד.

מטרת הפרויקט. פיתוח מערכת המסוגלת לסווג תמונה כ„אמיתית" (צולמה במצלמה) או כ„מזויפת" (נוצרה בבינה מלאכותית), ולספק הסבר חזותי להחלטת המודל באמצעות מפת חום (Grad-CAM).

ערך מוסף. הפתרון מיועד לעיתונאים, לפלטפורמות מדיה חברתית ולמשתמשים פרטיים המבקשים לאמת תמונות לפני שיתופן.

2. תיאור הדאטה

מקור הנתונים — מאגר Defactify

השתמשנו במאגר Defactify שפורסם במסגרת תחרות AAAI 2025. להלן מאפייני המאגר כפי שהורד ועובד אצלנו:

מאפיין	ערך
סך התמונות שהורדו	56,000
תמונות אמיתיות	16,000 (מקור: MS COCO)
תמונות שנוצרו ב-AI	40,000
גנרטורים	Midjourney v6, DALL·E 3, SDXL, SD3
חלוקת אימון / ולידציה	85% / 15%
מקור התמונות האמיתיות	MS COCO — צילומי יומיום מגוונים

חוסר איזון בדאטה. המאגר אינו מאוזן — יחס של כ-1:2.5 בין תמונות אמיתיות למזויפות. טיפלנו בכך באמצעות משקלי מחלקה (class weights) בפונקציית ה-loss: משקל 1.75 למחלקה האמיתית ו-0.70 למחלקת ה-AI.

עיבוד מקדים

- כיווץ כל התמונות לרזולוציה אחידה של 256×256 פיקסלים.
- נירמול לפי הממוצע והשונות של ImageNet (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

- הגדלת נתונים (Augmentations): היפוך אופקי אקראי, שינויי בהירות וניגוד, ודחיסת JPEG אקראית (הסתברות 30%, איכות 75) כדי לחזק עמידות לארטיפקטים.

3. המודל והאימון

ארכיטקטורה — EfficientNet-B0 עם Transfer Learning

בחרנו ב-EfficientNet-B0 על בסיס משקלים שאומנו מראש על ImageNet (1.28 מיליון תמונות, 1000 קטגוריות). הרשת כבר למדה לזהות מאפיינים בסיסיים — קצוות, טקסטורות וצורות — ואנו כיוונו אותה (fine-tuning) לזהות חתימות אופייניות לתמונות שנוצרו בבינה מלאכותית.

שינוי בארכיטקטורה. שכבת הסיווג הסופית הוחלפה ברצף Dropout(0.3) → Linear(1280→256) → ReLU → Dropout(0.3) → Linear(256→2). כל הפרמטרים אומנו, לרבות שכבות ה-backbone.

פרמטר אימון	ערך
Optimizer	AdamW
Learning Rate	5×10^{-5}
Weight Decay	1×10^{-4}
Batch Size	32
Scheduler	Cosine Annealing
Early Stopping	patience = 5, min_delta = 5×10^{-4}
חומרה	Apple Silicon — MPS

מהלך האימון. האימון נעצר אוטומטית לאחר 8 epochs בזכות ה-early stopping, כשהמודל הטוב ביותר התקבל כבר ב-epoch 2 — עדות להתכנסות מהירה ולחיסכון משמעותי במשאבי חישוב.

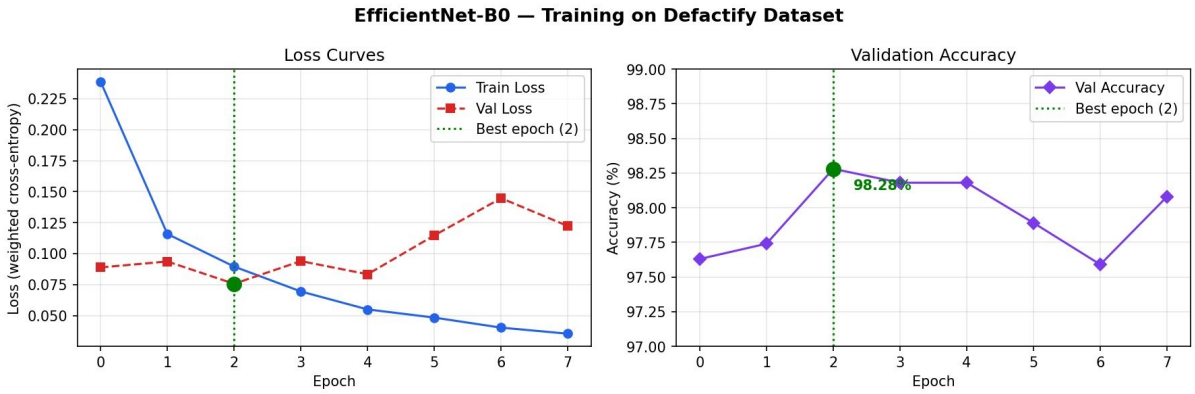
4. תוצאות

מדדי ביצועים על סט הוולידציה (כ-8,000 תמונות)

מדד	ערך
Accuracy	98.28%
AUC (ROC)	0.9986

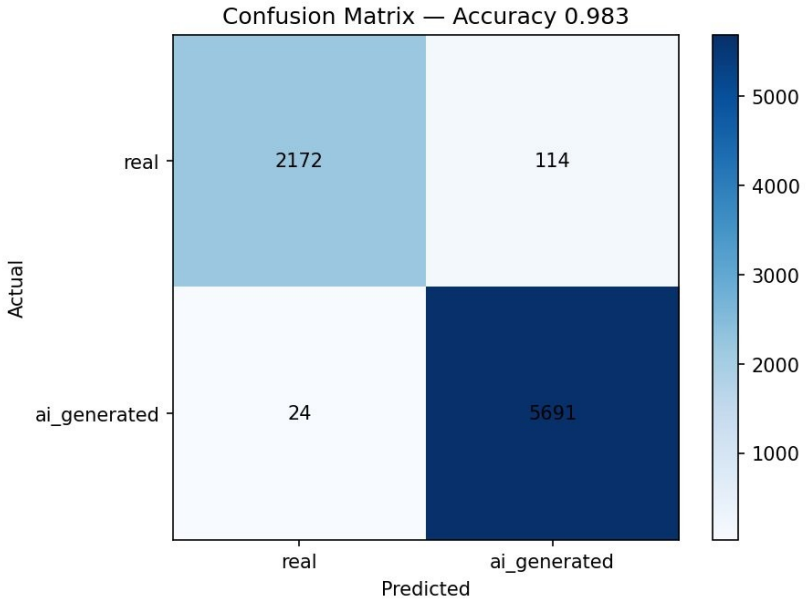
מדד	ערך
Best Epoch	2
Epochs עד עצירה	8 (early stopping)

עקומות האימון



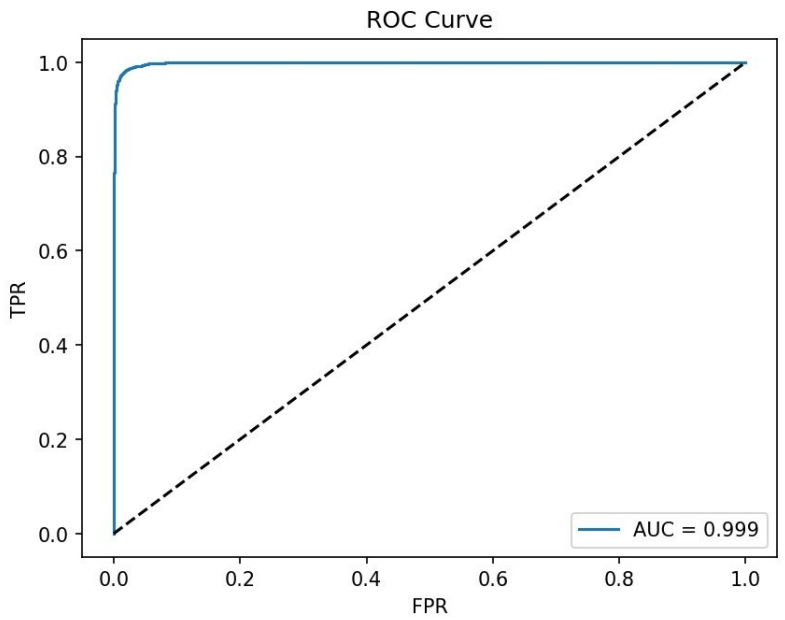
איור 1: עקומות ה-Loss וה-Accuracy לאורך האימון. ההתכנסות מתרחשת כבר ב-epoch 2, ולאחריו ניכרת מגמת overfitting קלה ב-Validation Loss.

מטריצת הבלבול (Confusion Matrix)



איור 2: מטריצת הבלבול על סט הוולידציה. רוב הטעויות הן 114 תמונות אמיתיות שסווגו בטעות כ-AI, לעומת 24 תמונות AI בלבד שסווגו כאמיתיות.

עקומת ROC



איור 3: עקומת ROC. הערך הגבוה (AUC = 0.9986) מעיד על הפרדה כמעט מושלמת בין שתי המחלקות.

Grad-CAM — הסבר חזותי להחלטה

המערכת מפיקה מפת חום (Grad-CAM) המדגישה את האזורים שהשפיעו ביותר על החלטת המודל. בבחינת הדוגמאות זיהינו כמה דפוסים חוזרים:

- פנים ועור — המודל מתמקד בטקסטורה חלקה מדי, האופיינית לתמונות AI.
- שיער ופרטים עדינים — אזורים שבהם גנרטורים נוטים לייצר חוסר קוהרנטיות.
- קצוות שבין אובייקט לרקע — המעבר נראה שונה מצילום אמיתי.

5. קשיים ומסקנות

קשיים שנתקלנו בהם

1. **חוסר איזון במאגר (1:2.5).** המאגר מכיל פי 2.5 יותר תמונות AI מאשר אמיתיות; נפתר באמצעות משקלי מחלקה.
2. **תאימות ל-Python 3.9.** שגיאות תחביר מסוג str | None בגרסה הישנה; הוחלף ב-Optional[str] מהמודול typing.
3. **צריכת זיכרון בהורדת הדאטה.** ניסיון להחזיק 80 אלף תמונות בזיכרון גרם לקריסה; נפתר בשמירה ישירה לדיסק תוך כדי streaming.
4. **MPS ו-pin_memory.** שבבי Apple Silicon אינם תומכים ב-pin_memory=True ב-DataLoader; הוגדר אוטומטית לפי סוג ההתקן.

5. **Grad-CAM מול no_grad.** ה-decorator @torch.no_grad() מנע את הגרדיאנטים הנדרשים ל-Grad-CAM; תוקן לשימוש ב-context manager על ה-forward pass בלבד.

מסקנות

- EfficientNet-B0 עם Transfer Learning מספק ביצועים מצוינים (98.28% דיוק) בזיהוי תמונות AI, גם על מאגר בגודל בינוני.
  - זמן עיבוד של כ-66 מילישניות לתמונה על CPU — מהיר דיו לשימוש בזמן אמת.
  - Grad-CAM מספק שקיפות לגבי ההחלטה ומחזק את אמון המשתמש במערכת.

כיוונים עתידיים

- אימון על מאגר גדול יותר (מעל מיליון תמונות) ועם גנרטורים נוספים.
- הרחבת זיהוי ה-Deepfake לסרטוני וידאו (כבר ממומש חלקית).
- פרסום המודל כ-API ציבורי לשימוש עיתונאים וחוקרים.
- כוונון נוסף על גנרטורים חדשים כגון Sora ו-Gemini Imagen 3.

הפרויקט פותח במסגרת קורס מולטימדיה ולמידת מכונה, 2025–2026.
קוד מקור: github.com/IconiqMotion/real-or-fake