CROSS-XAI : מסגרת להערכה חוצת-מודלים של שיטות הסבר בבינה מלאכותית

מגישים : מאור נוי ויהונתן ירט
מנחה : פרופסור אבי רוזנפלד

AUGUST 3, 2026

מוסד: המרכז האקדמי לב(JCT)
שנת הגשה: תשפ"ו / 2026

תקציר

בשנים האחרונות, השילוב של מודלי למידה עמוקה במערכות מבוססות ראייה ממוחשבת חווה פריצה משמעותית. עם זאת, רשתות אלו פועלות לרוב כ"קופסה שחורה", תכונה המעלה חששות כבדים בנוגע לאמינותן ובטיחותן בתחומים קריטיים. כדי לגשר על פער זה, פותחו מגוון שיטות להסבר החלטות המודל. אולם, מרבית כלי ההערכה הקיימים כיום סובלים מכשל מתודולוגי של "אימות מעגלי", מצב שבו המודל שיצר את מפת החום הוא גם המודל המשמש להערכתה. תופעה זו מובילה למדידת קיצורי דרך מתמטיים של הרשת במקום להעריך נאמנות סמנטית אמיתית.

פרויקט זה מציג את CROSS-XAI – מסגרת הערכה חדשנית ומקצה לקצה, המיועדת לבחון את האמינות והעמידות של שיטות הסבר באמצעות פרוטוקול הערכה בלתי-תלוי וחוצה-מודלים. ארכיטקטורת המערכת מבוססת על תהליך בן ארבעה שלבים: הפקת מפות חום על גבי מודל מקור, הפעלת מנגנון חסימה הדרגתי המעלים פיקסלים בהתאם לחשיבותם, העברת התמונות החסומות לבחינת שופטים חיצוניים דוגמת מודלי שפה-חזותיים, וכימות קצב דעיכת הדיוק של השופטים באמצעות מדדים סטטיסטיים. היבט ייחודי שנחקר בפרויקט הוא הטמעתן של שיטות סגמנטציה חזקות ובלתי-תלויות בגרדיאנטים, בפרט המודלים DINOv2 ו-U^2Net, כבסיס להשוואה סמנטית אובייקטיבית.

ממצאי הניסויים הניבו תובנות מרכזיות בתחום ה-XAI. ראשית, נמצא כי שיטות סגמנטציה בלתי-תלויות מספקות מסכות הסבר המשתוות באיכותן לשיטות מבוססות הגרדיאנטים הטובות ביותר. ממצא זה מאשש כי מודלי ה"קופסה השחורה" נוטים להסתמך על האובייקט הסמנטי הבולט ביותר בתמונה לצורך סיווג. שנית, נחשף פער פרדיגמטי עמוק בין סוגי השופטים: בעוד שופטי CNN זקוקים להסבר המבוסס על אזורים רציפים ומרוכזים, שופטי VLM מפגינים חוסן ומצליחים לפענח את התמונה גם תחת הסברים מקוטעים או מפוזרים, הודות ליכולתם להישען על הקשר סביבתי רחב.

הפיתוח בוצע בשפת Python תוך שימוש בספריות מתקדמות ללמידת מכונה וכלים מודרניים להאצת חומרה וניהול זיכרון. פלטפורמת CROSS-XAI מהווה תשתית הנדסית ומחקרית מן המעלה הראשונה, ומספקת כלי מעשי לפיתוח בינה מלאכותית בטוחה, לפענוח מנגנוני למידה ממוחשבת ולבניית אמון מוצק בין בני האדם למכונות.

Abstract

In recent years, the integration of deep learning models into computer vision systems has experienced a significant breakthrough. However, these networks often operate as "black boxes", raising serious concerns regarding their reliability and safety in critical domains. To bridge this gap, various explanation methods have been developed. Nevertheless, most current evaluation tools suffer from an inherent methodological flaw known as "Circular Validation", where the model that generated the heatmap is also used to evaluate it. This phenomenon leads to the measurement of mathematical shortcuts of the network rather than evaluating true semantic faithfulness.

This project introduces CROSS-XAI, an innovative end-to-end framework designed to assess the faithfulness and robustness of explanation methods through an independent, cross-model evaluation protocol. The system's architecture is based on a four-phase modular process: generating heatmaps on a source model, applying a progressive occlusion mechanism based on importance levels, passing the occluded images to external judges like advanced Vision-Language Models, and quantifying the accuracy decay rate using statistical metrics. A unique aspect explored within the project is the integration of strong, gradient-free segmentation methods, particularly DINOv2 and U²Net, as a baseline for objective semantic comparison.

The experimental findings yielded key insights in the field of XAI. First, independent segmentation methods provide explanation masks that match or surpass the best gradient-based attribution methods. This significantly corroborates the assumption that "black box" models tend to rely on the most prominent semantic object for classification. Second, a deep paradigmatic gap was revealed between the judges: while CNN judges require explanations based on continuous and spatially concentrated regions, VLM judges demonstrate exceptional robustness, successfully decoding the image even under fragmented explanations by relying on broad environmental context.

The development process was conducted in Python, utilizing advanced machine learning libraries and modern tools for hardware acceleration and memory management. The CROSS-XAI platform constitutes a first-rate engineering and research infrastructure, providing a practical tool for the development of Safe AI, the deep decoding of machine learning mechanisms, and building solid trust between humans and machines.

תודות

ברצוננו להודות למנחה שלנו, פרופ׳ אבי רוזנפלד, על הדרכתו המסורה והליווי המקצועי לאורך כל שלבי הפרויקט. כמו כן, אנו מבקשים להביע את הערכתנו למרכז האקדמי לב ולסגל המכון על התמיכה האקדמית, וכן על ההשקעה והסיוע ברכישת הציוד ומשאבי המחשוב הנדרשים, אשר אפשרו את קיום המחקר והשלמתו על הצד הטוב ביותר.

תוכן עניינים

תקציר 1
Abstract 2
תודות 4
רשימת איורים 8
רשימת טבלאות 8
רשימת קיצורים 8
1. מבוא 9
1.1 רקע כללי -Explainable AI 9
1.2 הגדרת הבעיה – הערכה מעגלית והטיות במדדי הסברים 10
1.3 הפתרון המוצע והתרומה המחקרית – מסגרת CROSS-XAI 11
1.4 מבנה הדוח 12
2. סקירת ספרות ורקע תיאורטי 13
2.1 התפתחות תחום הראייה הממוחשבת 13
2.2 בינה מלאכותית מוסברת ומפות חום 13
2.3 אתגרים בהערכת נאמנות של הסברים 14
2.4 מודלים שפתיים-חזותיים (VLMs) והתפתחותם 14
..... 15
2.5 ראייה ממוחשבת וסגמנטציה: זיהוי אובייקטים בולטים 15
2.6 מערכי נתונים במחקר 16
2.7 עבודות קודמות ומיצוב המחקר הנוכחי 17
3. הגישות הנבדקות: שיטות Attribution ושיטות סמנטיות 18
3.1 שיטות תלויות-מודל (Gradient & Perturbation-Based) 18
3.1.1 Saliency maps 18
3.1.2 Input x Gradient 18
3.1.3 Guided Backpropagation 19
3.1.4 Integrated Gradients (IG) 19
3.1.5 GradientSHAP 19
3.1.6 Occlusion Sensitivity 19
3.1.7 Grad-CAM 19

3.1.8 Guided Grad-CAM 20
3.1.9 XRAI 20
3.1.10 Random Baseline 20
3.2 שיטות עצמאיות מודל 22
3.2.1 DINOv2 – **ארכיטקטורה ופיתוח השיטות** 22
3.2.2 U^2Net – **זיהוי אובייקטים בולטים** 23
3.2.3 Fusion **- שילוב השיטות** 23
3.3 Wrappers – עיבוד נוסף של מפות החום 25
3.3.1 Continuous Smoothing 25
3.3.2 U^2Net Underlay 25
3.4 אסטרטגיות מילוי (Fill Strategies) 25
4. מתודולוגיית ההערכה: מסגרת CROSS-XAI ושיפוט VLM 26
4.1 עקרון הפעולה: שבירת ההערכה המעגלית באמצעות שופט עצמאי 26
4.2 ארכיטקטורת – CROSS-XAI פייפליין ארבע הפאזות 26
4.2.1 **שלב 1– יצירת הסברים ומיון פיקסלים** 26
4.2.2 **שלב 2 – הסתרה הדרגתית**(Progressive Occlusion) 27
4.2.3 **שלב 3 – שיפוט והסקה**(Judging) 27
4.2.4 **שלב 4 – צבירה, ניתוח והפקת מדדים** 27
4.3 אבולוציית השופטים 28
4.4 מודל LLaMA-3.2-Vision כשופט מתקדם במערכת 28
4.4.1 **מצבי השיפוט**(Evaluation Modes) 29
4.4.2 המרת פלט טקסטואלי למדד כמותי(Parsing) 29
4.4.3 אתגרים הנדסיים – ניהול KV-Cache ומגבלות אצווה 29
5. סביבת פיתוח ומימוש הנדסי 30
5.1 מתודולוגיית הפיתוח וכלים טכנולוגיים 30
5.2 תכנון רכיבי המערכת והקוד 31
5.3 ניהול משאבי עיבוד (GPU) וביצועים 31
5.4 הבטחת שחזור תוצאות 32
רשימת מקורות 33

רשימת איורים

איור 1- מפות חום שונות המיוצרות על ידי שיטות XAI.....10
איור 2- תרשים המתאר שימוש בVLM.....15
איור 3- Superficial-Intermediate.....16
איור 4 - Parabasal 16
איור 5- Football helmet.....16
איור 6- Boston bull 16

רשימת טבלאות

No table of figures entries found.

רשימת קיצורים

1. מבוא

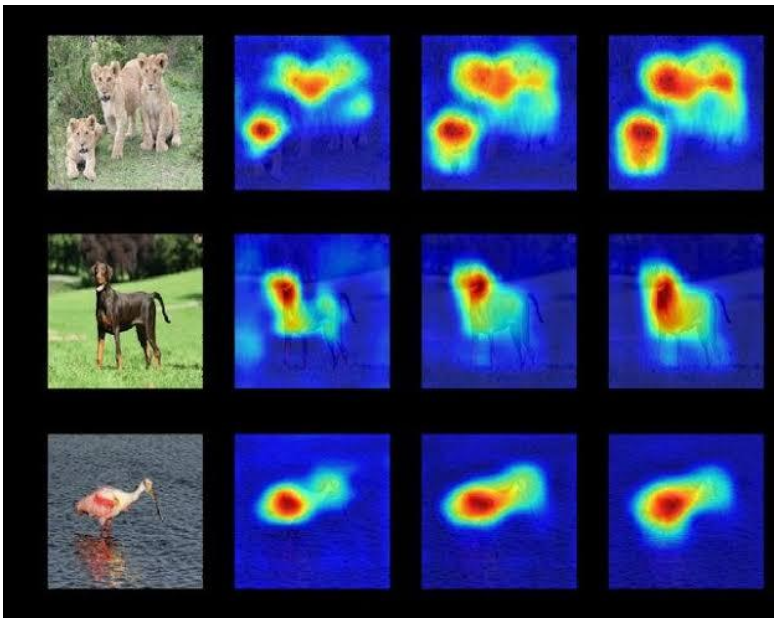
1.1 רקע כללי - Explainable AI

בעשור האחרון, אנו עדים למהפכה טכנולוגית חסרת תקדים בתחום הבינה המלאכותית (AI), ובפרט בלמידה עמוקה (Deep Learning). מודלים עמוקים הפכו לכלי עבודה מרכזי ביישומים קריטיים הנוגעים ישירות לחיי אדם, כגון מערכות עזר לאבחון רפואי, נהיגה אוטונומית, מערכות בקרת טיסה ואף מערכות תומכות החלטה משפטיות. במקרים רבים, ביצועיהם של מודלים אלו עולים על יכולות הסיווג והזיהוי של בני אדם מומחים.

עם זאת, לצד הביצועים יוצאי הדופן, רשתות עצביות עמוקות (Neural Networks) סובלות מחיסרון מובנה מהותי: הן פועלות כ"קופסה שחורה" (Black-Box). מודל למידה עמוקה טיפוסי מורכב ממיליוני או מיליארדי פרמטרים התלויים זה בזה באופן לא ליניארי מורכב. תכונה זו הופכת את תהליך ההסקה הפנימי של המודל לבלתי שקוף לחלוטין. כאשר מערכת אבחון רפואית מסווגת צילום רנטגן כמכיל גידול סרטני, הרופא – שהוא הנושא באחריות המקצועית והמוסרית – אינו מסוגל לדעת מתוך המערכת אילו אזורים בתמונה, או אילו מאפיינים פיזיולוגיים, הובילו להחלטה זו.

חוסר השקיפות יוצר חסם משמעותי בפני אימוץ מערכות בינה מלאכותית בתחומים שבהם נדרשות אמינות, אחריותיות (Accountability) ובטיחות ללא פשרות. עבור גופי רגולציה וחברות טכנולוגיה כאחד, ההסבריות הפכה בשנים האחרונות מנוחות בלבד לדרישת סף מחייבת. דוגמה בולטת לכך היא רגולציית ה-AI Act של האיחוד האירופי, המציבה דרישות שקיפות והסבריות מחמירות עבור מערכות המסווגות כ"מערכות בסיכון גבוה" (High-Risk Systems).

כמענה לאתגרים אלו, הגיח וצמח תחום מחקר מרכזי בשם **explainable AI - XAI**. מטרתו של התחום היא "לפתוח את הקופסה השחורה", או לפחות להאיר את אופן פעולתה, ולספק הסברים קריאים וברורים לאדם עבור החלטות המודל. בעולם הראייה הממוחשבת, ההסבר הנפוץ ביותר לפעולתם של מודלי סיווג הוא **מפת חום** או **מפת בולטות**. מפות אלו מחשבות ומציגות באופן ויזואלי ציון חשיבות עבור כל פיקסל בתמונת הקלט, המייצג את מידת תרומתו והשפעתו של אותו פיקסל על החלטת הסיווג הסופית.



איור 1- מפות חום שונות המיוצרות על ידי שיטות XAI.

כך, באופן תיאורטי, חוקר או משתמש קצה יכולים להבין מהם המאפיינים הוויזואליים ש"משכו את תשומת ליבו" של המודל, ולזהות האם המודל מתבסס על מאפיינים נכונים (למשל, מראה הגידול) או על הטיות (Biases) של מערך הנתונים (למשל, סימן מים המופיע תמיד בתמונות חיוביות). אך למרות ששיטות ה-XAI מספקות מענה לבעיית האטימות, הן מייצרות מיד אתגר משמעותי ומורכב בפני עצמו, אשר עומד במרכזו של מחקר זה: כיצד ניתן לדעת באופן אובייקטיבי האם ההסבר המסופק על ידי שיטת ה-XAI אכן מייצג נאמנה את תהליך קבלת ההחלטות האמיתי של המודל?

1.2 הגדרת הבעיה – הערכה מעגלית והטיות במדדי הסברים

בעוד ששיטות Attribution שונות מייצרות מפות חום ויזואליות המספקות תחושה של הבנה, הספרות המדעית מתמודדת עם קושי מהותי בהערכת אמינותן של מפות אלו. תכונה מרכזית וקריטית של כל הסבר נקראת **נאמנות** - כלומר, המידה שבה ההסבר משקף באופן מדויק את הסיבות האמיתיות שהובילו לתחזית המודל. למרבה הצער, לא קיים "Ground Truth" אוניברסלי למה צריכה להיות מפת החום הנכונה עבור מודל מסוים.

כדי להעריך נאמנות באופן כמותי, רוב המחקרים הקיימים מסתמכים על מתודולוגיות של הוספה או הסתרה של פיקסלים. הרעיון הבסיסי נראה הגיוני: מוחקים מהתמונה את הפיקסלים שדורגו כ"הכי חשובים" על ידי מפת החום. אם הפיקסלים הללו אכן היו חיוניים לסיווג, ציון הביטחון של המודל במחלקה החזויה אמור לצנוח באופן דרמטי. אם הוא לא ירד משמעותית, הרי שההסבר אינו נאמן, שכן המודל עדיין מצליח לסווג בהצלחה גם ללא המידע שלכאורה היה "הכי חשוב" לו.

עם זאת, גישה זו סובלת מכשל מתודולוגי מהותי המכונה **בעיית ההערכה המעגלית**. בגישות המסורתיות, המודל שמעריך את איכות ה-heatmap על ידי התמודדות עם התמונות המוסתרות, הוא בדיוק אותו המודל שעל בסיסו חושבה מפת החום מלכתחילה. מצב זה יוצר ניגוד עניינים מבני והטיית אישור: כאשר פוגעים בפיקסלים קריטיים לפי הגרדיאנט של מודל ספציפי, ברור שהרשת תשנה את הפלט שלה, שכן מפות החום חושבו מראש בדיוק כדי למקסם את השינוי הזה מקומית. לכן, ירידה בביצועי המודל במבחני הסתרה לא בהכרח מעידה על כך שמפת החום הצביעה על האובייקט הסמנטי האמיתי, אלא רק על כך שהיא הצליחה למצוא 'נקודות תורפה' ורגישויות בארכיטקטורה המתמטית של המודל הספציפי.

בנוסף, ייתכן מצב שבו המודל למד להתבסס על מתאמים אקראיים או רעשי רקע, ומפת החום מצביעה בדיוק עליהם. במצב כזה, מבחן ההסתרה המסורתי יעניק ציון "מצוין" לשיטת ה-XAI (משום שהמודל יאבד את יכולת הסיווג), על אף שההסבר רחוק מלהיות הגיוני תפיסתית לאדם, והמודל עצמו פגום. הבעיה מחמירה מכיוון שמפתחים משתמשים במדדי הנאמנות הללו כדי לשפר ולהשוות שיטות XAI, מה שעלול להוביל לOver-fitting של שיטות ההסבר למודל הבודק את עצמו.

כדי לנתק את התלות המעגלית הזו ולקבל אומדן אובייקטיבי לנאמנות ואמינות של מפות חום, נדרש גישור באמצעות גורם שלישי - שופט עצמאי ובלתי תלוי, שאינו שותף לתהליך יצירת ההסבר, ויכול לאמת האם המידע שנשמר או הוסתר על ידי ההסבר אכן מכיל משמעות סמנטית אמיתית.

1.3 הפתרון המוצע והתרומה המחקרית – מסגרת CROSS-XAI

בפרויקט זה אנו מציגים את **CROSS-XAI**: מסגרת מחקרית והנדסית מתקדמת להערכה חוצת-מודלים של שיטות הסבריות בבינה מלאכותית. המסגרת מציעה פתרון מערכתי לבעיית ההערכה המעגלית על ידי הוצאת תהליך השיפוט מידיו של המודל היוצר, והעברתו לידיהם של מודלים שופטים עצמאיים לחלוטין.

עיקרון הפעולה המרכזי ב-CROSS-XAI מבוסס על תהליך של **הסתרה פרוגרסיבית**:

1. בשלב הראשון, מודל המטרה והשיטה הנבדקת מייצרים מפת חום המדרגת את חשיבות הפיקסלים בתמונה.
2. בשלב השני, המערכת מסתירה באופן מדורג את הפיקסלים החשובים ביותר, רמה אחר רמה (מ-0% ועד 100% הסתרה).
3. השלב השלישי והחדשני: התמונות המוסתרות אינן מוזנות חזרה למודל המקורי, אלא מועברות לשורת מודלים חיצוניים בעלי ארכיטקטורות שונות, אשר נדרשים לסווג את התמונה המוסתרת.

הלוגיקה פשוטה אך רבת עוצמה: אם הסתרנו את הפיקסלים שלפי ההסבר הם ה"מרכזיים והחשובים ביותר", ומודל שופט עצמאי לחלוטין מאבד כתוצאה מכך את יכולתו לזהות את

האובייקט בתמונה – הרי שיש בידינו אימות אובייקטיבי ובלתי-תלוי לכך שמפת החום אכן הצביעה על האזורים בעלי המשמעות הסמנטית הגבוהה ביותר בתמונה, ולא על רעש מתמטי.

במסגרת המחקר בדקנו שורה רחבה של שיטות Attribution מסורתיות מבוססות גרדיאנטים. לצד שיטות אלו, הוספנו תרומה מחקרית ייחודית: שילבנו **שיטות סגמנטציה וגילוי אובייקטים בולטים** והשתמשנו בהן כמעין שיטות XAI לכל דבר ועניין. שיטות אלו לא הוכשרו למשימת סיווג ואינן תלויות במודל המקור, אלא נועדו לאתר את האובייקט הסמנטי הבולט באופן טהור. הצבתן כנקודת ייחוס חלופית אפשרה לנו לאשש בצורה אמפירית האם מפות החום הקלאסיות אכן מזהות את האובייקט הבולט, או נופלות ל"הזיות רקע".

כדי להבטיח הערכה עמוקה ומגוונת, מסגרת ה-CROSS-XAI מעמתת את ההסברים מול שלוש פרדיגמות שונות לחלוטין של שופטים: מודלים קונבולוציוניים מקומיים (CNN), מודלים מבוססי תשומת-לב גלובלית (Vision Transformers), והחידוש המרכזי – **מודלים שפתיים-חזותיים רחבים (VLMs)**, דוגמת LLaMA-3.2-Vision. בחירה זו מייצגת את חזית הטכנולוגיה ומאפשרת לשפוט את ההסבר דרך פריזמה של תפיסה מרחבית, הקשרית ושפתית כאחד.

1.4 מבנה הדוח

דוח זה בנוי באופן הדרגתי על מנת לספק את הרקע התיאורטי והטכני המלא להבנת הפרויקט, תוצאותיו ומשמעותן:

- **פרק 2** סוקר את הספרות המקצועית והרקע התיאורטי בתחומי ה-XAI, הסגמנטציה ומודלי ה-VLM, תוך מיצוב מחקר זה ביחס לעבודות קודמות.
- **פרק 3** מציג את הגישות הנבדקות, ומשווה בין שיטות Attribution קלאסיות לבין הגישה החדשה המבוססת על סגמנטציה סמנטית. הפרק מפרט גם על עטיפות (Wrappers) ואסטרטגיות פוסט-פרוססינג שיושמו.
- **פרק 4** מתאר את מתודולוגיית ההערכה, את עקרון הפעולה של מסגרת ה-CROSS-XAI ואת אבולוציית השופטים עד לשימוש ב-VLM.
- **פרק 5** מוקדש לסביבת הפיתוח והמימוש ההנדסי, לרבות ארכיטקטורת התוכנה והתמודדות עם אתגרי עיבוד (GPU).
- **פרק 6** מציג באופן גרפי ומספרי את תוצאות הניסויים על פני פרדיגמות השופטים השונות.
- **פרק 7** כולל דיון וניתוח מנגנוני מעמיק השואף להסביר מדוע שיטות מסוימות (כגון הסגמנטציה ו-Grad-CAM) הציגו ביצועים עודפים, וכיצד הגיב שופט ה-VLM לרעשים.
- **פרק 8** מסכם את תרומת העבודה, מציג את המגבלות ההנדסיות והמחקריות שניצבו בדרכנו, ומציע כיווני המשך עתידיים.

2. סקירת ספרות ורקע תיאורטי

פרק זה סוקר את התפתחות הראייה הממוחשבת ותחום הבינה המלאכותית המוסברת (XAI), ובוחן לעומק את האתגרים הקיימים כיום בהערכת נאמנותם של הסברים. בנוסף, מוצגים בו תחומי ידע משיקים כגון זיהוי אובייקטים בולטים ומודלים שפתיים-חזותיים, המהווים את הנדבך הטכנולוגי והמחקרי שעליו נשענת ארכיטקטורת ההערכה שפותחה בפרויקט זה. מטרת הפרק היא להציג את המגבלות של השיטות הקיימות, ולהסביר מדוע נדרש הפתרון שאנו מציעים במסגרת CROSS-XAI.

2.1 התפתחות תחום הראייה הממוחשבת

בעשור האחרון חלה התקדמות דרמטית ביכולות הראייה הממוחשבת, אשר הונעה בעיקרה על ידי פיתוחן של רשתות קונבולוציה (CNN). כדי להבין כיצד רשתות אלו "רואות", ניתן לדמיין אותן כמי שמסתכל על תמונה דרך זכוכית מגדלת קטנה: הרשת סורקת את התמונה אזור אחר אזור, ומחפשת תבניות מקומיות כמו קווים, זוויות ומרקמים, בעזרת מסננים מתמטיים. ככל שמעמיקים בשכבות הרשת, היא מחברת את הקווים לצורות מורכבות יותר, עד לזיהוי האובייקט השלם[1]

עם זאת, בשנים האחרונות מתרחש מעבר הדרגתי לפרדיגמה המבוססת על מנגנוני תשומת-לב, ובראשם ארכיטקטורת Vision Transformer (ViT). בניגוד לרשתות הקונבולוציה המוגבלות לראייה מקומית, מודל ViT "גוזר" את התמונה לרשת של ריבועים קטנים (אסימונים), בדומה לחלקי פאזל. לאחר מכן, המודל בוחן את כל חלקי הפאזל במקביל ולומד את הקשרים ביניהם, גם אם הם רחוקים זה מזה בתמונה. יכולת זו מאפשרת למודל ללכוד הקשרים גלובליים ולהגיע לביצועים חסרי תקדים **[2]**.

2.2 בינה מלאכותית מוסברת ומפות חום

עם העלייה במורכבות המודלים, נוצר נתק בין יכולת הזיהוי המצוינת של הרשת לבין היכולת שלנו להבין *למה* היא החליטה כפי שהחליטה – תופעת "הקופסה השחורה" **[3]**. כדי לפתור זאת, התפתח תחום ה-XAI. בעולם הראייה הממוחשבת, ההסבר הנפוץ ביותר הוא **מפת חום**.

באופן אינטואיטיבי, אם רשת זיהתה תמונה כ"כלב", מפת חום אידיאלית תצבע באדום עז את הפנים והגוף של הכלב, ותשאיר את צבעי הרקע (כמו דשא או שמיים) בכחול כהה. כך יכול המשתמש לוודא שהמודל באמת זיהה את החיה, ולא למד בטעות שדשא ירוק אומר "כלב".

שיטות יצירת מפות החום נחלקות לשתי משפחות עיקריות:

1. **שיטות מבוססות גרדיאנט:** מנתחות את המשוואות המתמטיות (הנגזרות) של הרשת כדי להבין אילו פיקסלים השפיעו הכי הרבה על התוצאה הסופית

2. **שיטות מבוססות הפרעה:** מסתירות חלקים שונים בתמונה ובודקות כיצד ההסתרה משפיעה על ביטחון המודל.

2.3 אתגרים בהערכת נאמנות של הסברים

הפקת מפת חום היא רק חצי מהעבודה. האתגר הגדול ביותר כיום הוא למדוד באופן כמותי האם ההסבר אמיתי ואמין – תכונה הנקראת נאמנות (Fidelity) **[4]**.

השיטה המקובלת למדוד זאת היא באמצעות "מבחן הסתרה" (Occlusion/Deletion). הרעיון הלוגי פשוט: נניח שמפת החום טוענת שהזנב של הכלב הוא האזור החשוב ביותר. אם נצייר ריבוע שחור שיסתיר את הזנב, ונבקש מהמודל לזהות שוב את התמונה, רמת הביטחון שלו אמורה לצנוח משמעותית. אם הוא עדיין מזהה את הכלב באותו ביטחון גבוה – סימן שמפת החום שיקרה לנו.

מבחינה פורמלית, אם Sc(x) הוא ציון הביטחון של המודל למחלקה c עבור תמונת הקלט המקורית x, ו-xmask היא התמונה לאחר הסתרת הפיקסלים ה"חשובים", מדד הנאמנות המקובל (Drop) מחושב כך:

$$Drop = Sc(x)max(0, Sc(x) - Sc(xmask)) \times 100$$

עם זאת, גישה זו סובלת משני כשלים מחקריים אנושים שאותם אנו באים לפתור בעבודה זו:

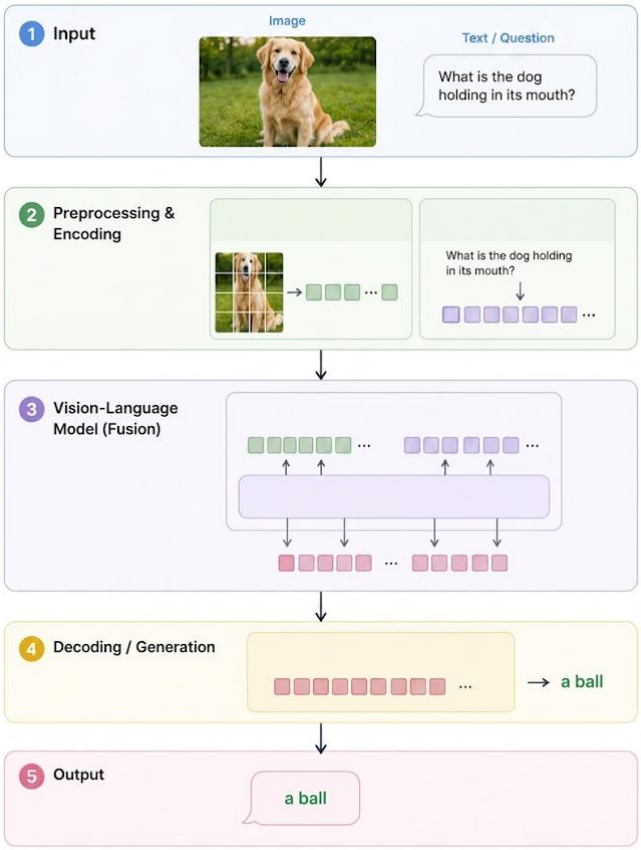
1. **בעיית ההתפלגות החריגה (OOD):** צביעת פיקסלים בשחור מייצרת תמונה מלאכותית שהמודל מעולם לא ראה באימון. המודל עלול להיכשל בזיהוי פשוט בגלל שהתמונה נראית לו "מוזרה", ולא בגלל שהסתרנו מידע קריטי **[5]**.

2. **בעיית ההערכה המעגלית:** זהו הכשל החמור מכל. המודל שנבחן במבחן ההסתרה (זה שמחשב את הציון הסופי) הוא *אותו מודל* שיצר את מפת החום מלכתחילה. מצב זה משול לתלמיד שכותב לעצמו את המבחן ובודק אותו. מפת החום עלולה למצוא "נקודות תורפה" מתמטיות ספציפיות של המודל שיפילו את רמת הביטחון שלו, גם אם מבחינה חזותית האזורים הללו חסרי משמעות לחלוטין לבן אנוש **[6]**.

2.4 מודלים שפתיים-חזותיים (VLMs) והתפתחותם

ההתפתחות המשמעותית ביותר לאחרונה היא צמיחתם של מודלים שפתיים-חזותיים (VLM) **[7]**. מודלים אלו משלבים "עיניים" (מקודד חזותי) יחד עם "מוח שפתי" (מודל שפה גדול).

כדי להבין את כוחם, מספיק להתבונן באופן השימוש בהם: משתמש יכול להזין למודל תמונה ולשאול אותו בטקסט חופשי: "האם אתה עדיין מצליח לזהות את האובייקט המרכזי בתמונה למרות שהסתרתי חלק ממנו?". המודל יענה בשפה טבעית ויסביר את החלטתו. בניגוד למודלי סיווג רגילים שרק פולטים הסתברות יבשה, ה-VLM מסוגל להבין הקשרים (Zero-shot) ללא אימון מוקדם על משימה ספציפית. תכונה זו הופכת אותם למועמדים אידיאליים לשמש כ"שופטים חיצוניים" ובלתי-תלויים לאימות מפות החום שלנו.



איור 2- תרשים המתאר שימוש בVLM

2.5 ראייה ממוחשבת וסגמנטציה: זיהוי אובייקטים בולטים

תחום זיהוי האובייקטים הבולטים מתמקד בפיתוח אלגוריתמים המדמים את מערכת הראייה האנושית. כשאדם מביט בתמונה, המוח שלו מתמקד באופן טבעי באובייקט המרכזי (למשל מכונית בכביש) ומתעלם מרעשי הרקע. מודלים מודרניים כגון U^2Net מסוגלים לבצע פעולה זו בדיוק רב ולייצר "מסכה" (סגמנטציה) החותכת את האובייקט המרכזי במדויק, מבלי לדעת מראש באיזה אובייקט מדובר **[8][9]**

במסגרת עבודה זו, אנו משתמשים במודלים אלו כבסיס-ייחוס (Baseline). השערת המחקר שלנו היא שאם מפת חום קלאסית באמת עובדת טוב, היא אמורה לחפוף כמעט לחלוטין למסכת הסגמנטציה הטבעית של האובייקט הבולט.

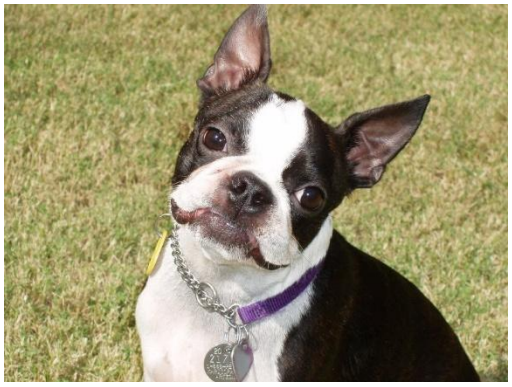
2.6 מערכי נתונים במחקר

בעוד ששיטות XAI רבות מפותחות ונבדקות על סביבות כלליות, יעילותן ומהימנותן הקריטית נבחנות הלכה למעשה בסביבות קצה מורכבות כגון הדמיה רפואית. בעבודה זו, הערכנו את המסגרת שלנו על פני שני תחומי תוכן נפרדים :

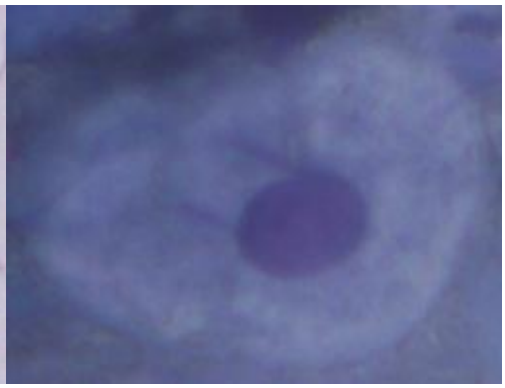
1. **ImageNet ILSVRC:** מערך הנתונים הסטנדרטי והמוכר ביותר בתחום הראייה הממוחשבת. הוא כולל תמונות של סצנות טבעיות ויום-יומיות, לרוב בעלות אובייקט מרכזי ברור, ומשמש כדאטה העיקרי שבו מנוסות שיטות ההסבר הקלאסיות.**[10]**

2. **SIPaKMeD:** מערך נתונים רפואי ייעודי לסיווג וזיהוי תאים במשטחי צוואר הרחם (Pap smear). הדמאות רפואיות מסוג זה מתאפיינות ברעש חזותי גבוה, צפיפות תאים וגבולות סמנטיים מטושטשים בין המחלקות (תאים תקינים לעומת טרום-סרטניים). הצורך בהסבריות אמינה בסביבה זו הוא בגדר קריטי. בחינת שיטות ההסבר על נתונים אלו מהווה מבחן הכללה (Generalization Test) משמעותי הן עבור השיטות עצמן והן עבור יכולת השיפוט של ה-VLM s תחת דומיין שאינו טבעי עבורם (OOD domain).**[11]**

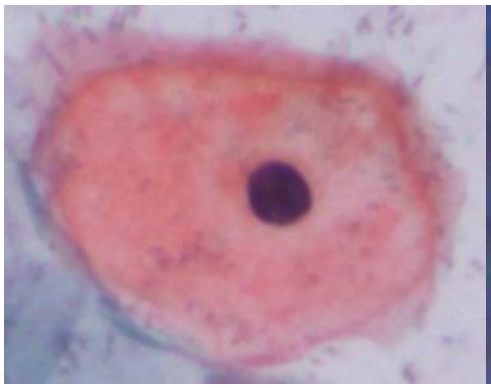
איור 5- Football helmet



איור 6- Boston bull



איור 4 - Parabasal



איור 3- Superficial-Intermediate

2.7 עבודות קודמות ומיצוב המחקר הנוכחי

מחקרים קודמים שניסו לפתור את בעיות ההערכה הציעו למלא את האזורים המוסתרים בטשטוש עדין או ברעש אקראי כדי למנוע את בעיית ה-OOD **[12]**. אולם, עבודות אלו עדיין השתמשו באותו המודל היוצר כדי לבחון את עצמו, ולכן לא פתרו את ההערכה המעגלית. מחקרים אחרים ביקשו מבני אדם לסמן ידנית את האזורים החשובים בתמונה ולהשוות אותם למפת החום, פתרון מדויק אך איטי ויקר להחריד שאינו מאפשר בדיקה אוטומטית של מיליוני תמונות **[13].**

הפתרון המוצע במסגרת CROSS-XAI משלב את הטוב משני העולמות: אנו שומרים על היכולת לבדוק אלפי תמונות באופן אוטומטי ומהיר, אך פותרים את ההערכה המעגלית מהשורש. אנו עושים זאת על ידי שימוש במודלי VLM כשופטים חיצוניים שאינם מכירים את הרשת המקורית, ובמקביל משתמשים באלגוריתמי סגמנטציה אוטומטיים כתחליף לתיוג אנושי יקר. גישה משולבת זו מאפשרת לנו להעריך את מפות החום בצורה אובייקטיבית וסמנטית חסרת תקדים.

3. הגישות הנבדקות: שיטות Attribution ושיטות סמנטיות

פרק זה מפרט את הבסיס התיאורטי והמתמטי של השיטות שהופעלו במסגרת מערכת CROSS-XAI לייצור מפות חום (Heatmaps). השיטות מחולקות לשתי קטגוריות מרכזיות: שיטות תלויות-מודל (Model-Dependent), השואבות מידע או מודדות רגישות ישירות מתוך "הקופסה השחורה" הנבדקת, ושיטות עצמאיות-מודל (Model-Independent), המבוססות על מודלי סגמנטציה ופועלות באופן חיצוני ובלתי תלוי בארכיטקטורת הסיווג.

3.1 שיטות תלויות-מודל (Gradient & Perturbation-Based)

שיטות אלו מתבססות על המודל הנבדק עצמו. רובן מחלצות מידע מתוך משקלי הרשת והאקטיבציות הפנימיות (מבוססות-גרדיאנט), ואחרות מודדות את תגובת המודל לשינויים בקלט (מבוססות-הפרעה).

3.1.1 Saliency maps

זוהי אחת השיטות הפשוטות והוותיקות ביותר להסבר רשתות עצביות. השיטה מחשבת את הגרדיאנט של ציון המחלקה שנחזתה ביחס לכל פיקסל בתמונה. ערך גרדיאנט גבוה מצביע על כך ששינוי קטן באותו פיקסל ישפיע משמעותית על החלטת המודל, ולכן הוא נחשב חשוב. עבור תמונה x ומחלקה c, מפת החום $M_c(x)$ מוגדרת כערך המוחלט של הגרדיאנט:

$$M_c(x) = \left| \frac{\partial S_c(x)}{\partial x} \right|$$

הגרדיאנט מייצג עד כמה שינוי מינורי בפיקסל ישפיע על פלט המחלקה. שיטה זו לרוב מניבה מפות "רועשות" מכיוון שאינה מתחשבת בהקשר הלוקאלי הרחב.[16]

3.1.2 Input x Gradient

שיטה זו מרחיבה את Saliency Map באמצעות הכפלת הגרדיאנט בערך המקורי של כל פיקסל בתמונה. כך מתקבלת מפה שמדגישה פיקסלים שגם השפעתם גבוהה וגם ערכם המקורי משמעותי, ובדרך כלל מפחיתה רעש ביחס לגרדיאנט הבסיסי. [17]

$$M_c(x) = x \odot \frac{\partial S_c(x)}{\partial x}$$

3.1.3 Guided Backpropagation

בשיטה זו מתבצע שינוי קטן בתהליך החישוב לאחור (Backpropagation). כאשר הגרדיאנט שלילי, הוא מאופס במעבר דרך שכבות ReLU. כתוצאה מכך נשמרים רק אותות התורמים באופן חיובי לחיזוי, ומתקבלת מפת חום חדה וברורה יותר. [18]

3.1.4 Integrated Gradients (IG)

רשתות עמוקות נוטות לסבול מ"רוויית גרדיאנטים" (Gradient Saturation) – מצב שבו הפונקציה משתטחת והגרדיאנט מתאפס, למרות שהמאפיין קריטי להחלטה. IG פותרת זאת על ידי ביצוע איטגרל של הגרדיאנטים לאורך מסלול ליניארי מתמונת ייחוס ריקה ((Baseline x' אל התמונה המקורית x: [19]

$$IG_i(x) = (x_i - x'_i) \times \int_0^1 \frac{\partial S_c(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

3.1.5 GradientSHAP

GradientSHAP מרחיבה את Integrated Gradients באמצעות הוספת רעש אקראי לתמונה וחישוב ממוצע של מספר רב של אינטגרלי גרדיאנטים. התוצאה היא מפות חום חלקות ויציבות יותר, שפחות רגישות לשינויים קטנים בקלט. [20]

3.1.6 Occlusion Sensitivity

זוהי שיטה המבוססת על הפרעה לתמונה. חלון קטן מכסה בכל פעם אזור אחר בתמונה, ובכל צעד נמדדת הירידה בביטחון המודל. אם הסתרת אזור מסוים גורמת לירידה משמעותית בביצועי המודל, האזור נחשב חשוב לחיזוי. [21]

$$M_c(x)_{i,j} = S_c(x) - S_c(x_{masked(i,j)})$$

3.1.7 Grad-CAM

שיטה המתמקדת אך ורק בשכבת הקונבולוציה האחרונה של הרשת, המאופיינת בעושר סמנטי גבוה על אף רזולוציה מרחבית גסה. תחילה מחושב משקל ערוצי α_k^c לכל מפת מאפיינים A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial S_c}{\partial A_{i,j}^k}$$

לאחר מכן מתבצעת סכימה ממושקלת המועברת דרך פונקציית ReLU, כדי לשמור רק על האזורים שתורמים פוזיטיבית לסיווג: [22]

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right)$$

Guided Grad-CAM 3.1.8

מיזוג בין הרזולוציה המרחבית של Grad-CAM, לבין הרזולוציה העדינה ברמת הפיקסל של Guided Backpropagation. השילוב מתבצע על ידי הכפלה נקודתית (Element-wise) של מפת ה--Grad CAM (לאחר Upsampling) במפת ה-Guided Backprop: [22]

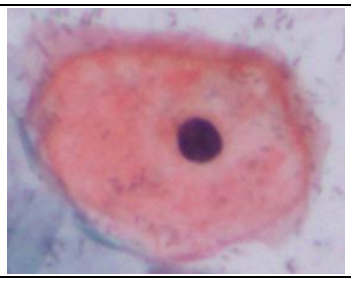
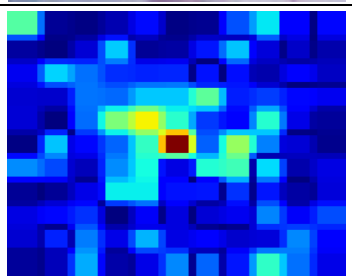
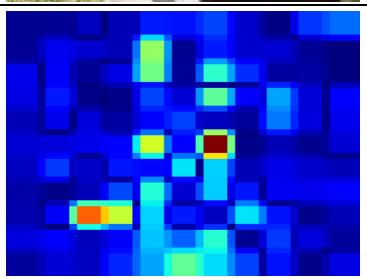
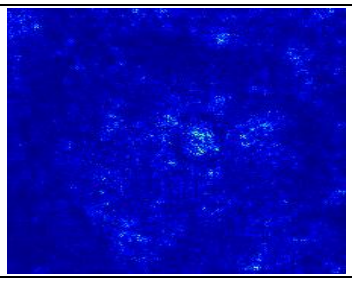
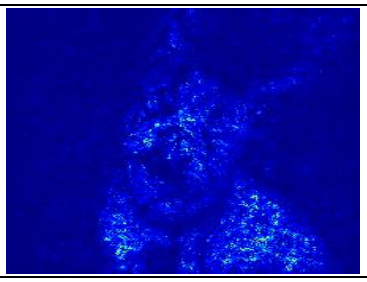
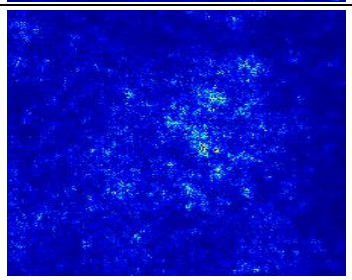
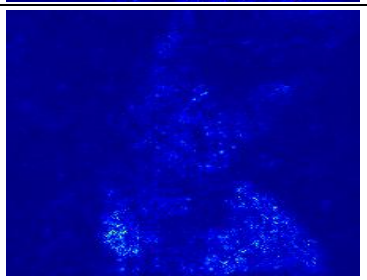
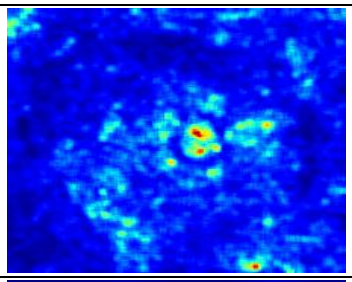
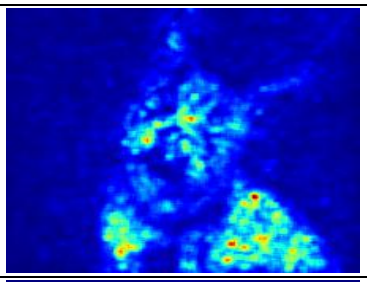
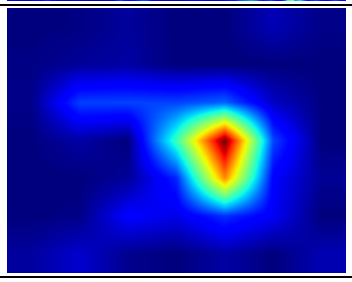
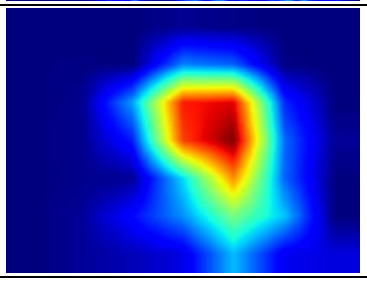
$$L^c_{Guided-Grad-CAM} = L^c_{Grad-CAM} \uparrow \odot R_{Guided-Backprop}$$

XRAI 3.1.9

XRAI מחלקת את התמונה לאזורים בעלי דמיון חזותי (Superpixels). לאחר חישוב מפת Integrated Gradients, האלגוריתם מאחד אזורים סמוכים ובעלי חשיבות דומה, וכך מתקבלת מפת חום המדגישה אובייקטים שלמים במקום פיקסלים בודדים. [23]

Random Baseline 3.1.10

שיטה המשמשת אך ורק כנקודת ייחוס (Sanity check) להערכת המערכת. מפת החום מיוצרת על ידי דגימה מקרית לחלוטין מתוך התפלגות אחידה רציפה $U(0,1)$ לכל פיקסל בתמונה. מטרתה לבחון האם שיטות ה-XAI אכן מציגות ביצועים העולים על ניחוש אקראי.

		Input image
		Occlusion
		Integrated Gradients
		Saliency
		XRAI
		Grad cam

3.2 שיטות עצמאיות מודל

התרומה המרכזית של מחקר זה היא פיתוח קבוצת שיטות המבוססות על **DINOv2** ו-**U²Net**. בניגוד לשיטות Attribution, שיטות אלו אינן מנסות להבין כיצד המודל קיבל את החלטתו, אלא מזהות באופן ישיר את האובייקטים המשמעותיים בתמונה. מפות הסגמנטציה המתקבלות משמשות כבסיס אובייקטיבי להשוואה מול שיטות ההסבר הקלאסיות.

3.2.1 DINOv2 – ארכיטקטורה ופיתוח השיטות

א. כיצד DINOv2 מייצג תמונה?

DINOv2 הוא מודל מסוג ViT. במקום לעבד את התמונה פיקסל אחר פיקסל, הוא מחלק אותה לריבועים קטנים (Patches), בגודל של 14×14 פיקסלים. **[9]**

כל Patch עובר דרך הרשת ומיוצג באמצעות וקטור מספרי (Token) המתאר את התכונות החזותיות שלו. בנוסף, המודל מחזיק וקטור מיוחד בשם **CLS Token**, אשר מרכז מידע מכל הטלאים ומייצג את התמונה כולה.**[9]**

כדי להפיק מפת חום, ממירים כל וקטור לערך מספרי יחיד, מסדרים את הערכים לפי מיקומם המקורי בתמונה ומגדילים אותם חזרה לגודל התמונה באמצעות Upsampling. התוצאה היא מפת חום שבה כל אזור מקבל ציון המייצג את חשיבותו.

במחקר נעשה שימוש בגרסת DINOv2 הכוללת Registers, בשל בעיה מוכרת במודלי ViT שנפתרה באמצעות רגיסטרים.**[14]**

ב. מפת הבסיס ותיקון הקוטביות

בלב חלק מהאלגוריתם שפיתחנו נמצא שימוש באלגוריתם הPCA לכיווץ המימדים של כל הווקטורים שהפיק המודל. הרכיב הראשי המתאים ביותר נבחר לפי מידת השונות שהוא מתאר.(או ביצוע שיטות אחרות שבוחרות כמה רכיבים ומשלבות אותן ביחד בצורה מסויימת)

מאחר שכיוון הווקטור ב-PCA הוא שרירותי, פותח מנגנון אוטומטי לתיקון הקוטביות. המנגנון משווה בין כל Patch לבין וקטור ה-CLS באמצעות דמיון קוסינוס. אם מתקבל מתאם שלילי, כיוון הרכיב מתהפך.

$$\mathrm{Corr} = \rho\left(Scores, \cos_sim(Patch_{Tokens}, CLS)\right)$$

כך מתקבלת מפת בסיס עקבית, רציפה ומנורמלת לטווח [0,1], ללא צורך בהפיכתה למסכה בינארית.

ג. המתודות שפותחו על בסיס DINOv2

על בסיס מנגנון זה פותחו שיטות שונות. להלן השיטות העיקריות שנוצרו:

- **DINO_ATTN** – משתמשת במנגנון תשומת הלב (Attention) של המודל כדי למדוד עד כמה כל Patch משפיע על ה-CLS, ובכך למצוא חשיבות של איזור בתמונה.
- **DINO_PC1** – מבצעת PCA ולוקחת את הרכיב עם השונות הכי גבוהה, וממנו יוצרת מפת חום.
- **DINO_PC_EV** – משלבת את שלושת הרכיבים הראשיים הראשונים עם השונות הכי גבוהה, כאשר כל רכיב מקבל משקל לפי אחוז השונות שהוא מסביר.
- **COMBO_ENT** – מקצה באופן אוטומטי משקל לכל אחת מהמפות בהתאם לרמת הביטחון שלה, הנמדדת באמצעות אנטרופיית שאנון. מפות ממוקדות מקבלות משקל גבוה יותר ממפות מפוזרות.

$$H(M) = -\frac{\sum_{i=1}^{N} p_i \log(p_i)}{\log(N)}$$

- **COMBO_ENT_SMOOTH** – מרחיבה את COMBO_ENT באמצעות החלקה גאוסית, סינון נוסף ומיזוג סופי השומר על גבולות האובייקטים.

3.2.2 U²Net – זיהוי אובייקטים בולטים

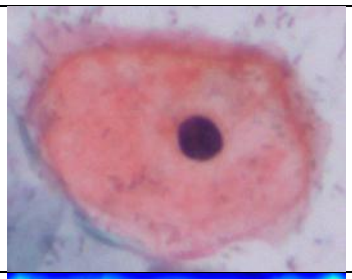
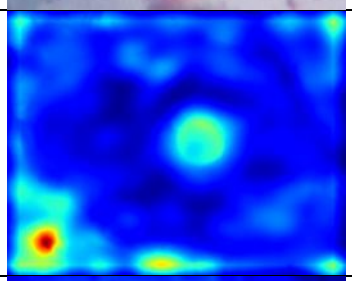
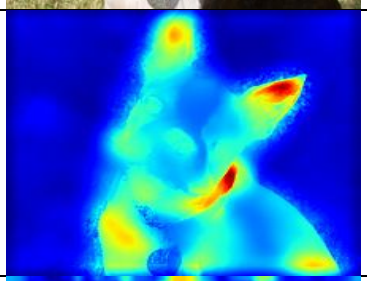
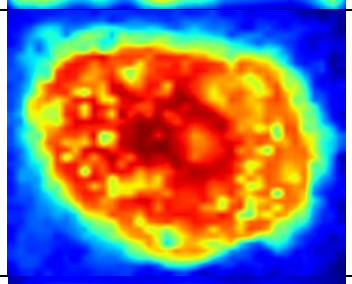
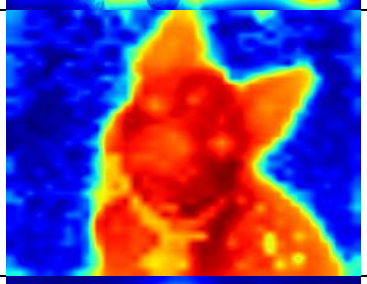
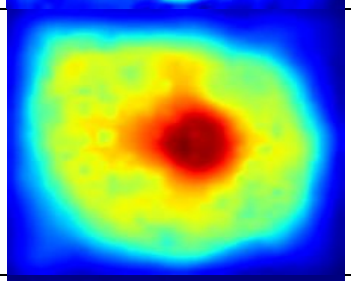
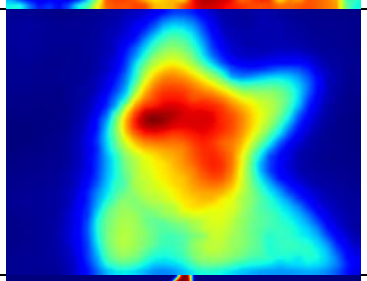
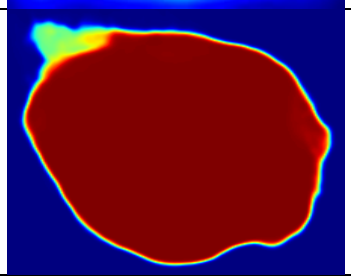
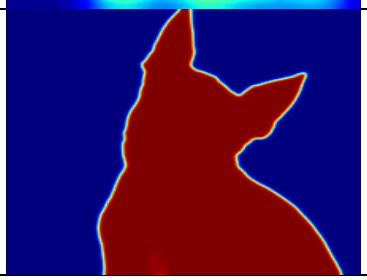
בניגוד למודל DINOv2, המחלק את התמונה לטלאים, U²Net פועל ישירות ברמת הפיקסל.

הרשת בנויה בארכיטקטורת Encoder–Decoder. בשלב הראשון היא מקטינה בהדרגה את התמונה כדי להבין את המבנה הכללי שלה, ובשלב השני מחזירה אותה לגודלה המקורי תוך שמירה על גבולות מדויקים.

פלט המודל הוא מפת הסתברויות בטווח [0,1], שבה כל פיקסל מקבל ציון המבטא את ההסתברות שהוא שייך לאובייקט המרכזי בתמונה. כתוצאה מכך מתקבלות מסכות חדות במיוחד.**[8]**

3.2.3 Fusion - שילוב השיטות

כדי למקסם את ההבנה הסמנטית והיציבה של DINOv2 יחד עם חדות הקצוות של U²Net ברמת הפיקסל, בדקנו אסטרטגיות משולבות. שתי המפות עוברות נרמול ומשולבות יחד פיקסלית בצורות שונות

		Input image
		DINO_ATTN
		DINO_PC1
		COMBO_ENT_SMOOTH
		U^2Net
		DINOv2 + U^2Net

3.3 Wrappers – עיבוד נוסף של מפות החום

בנוסף לשיטות עצמן, פותחו רכיבי עיבוד נוספים שמטרתם לשפר או לנתח מפות חום קיימות.

3.3.1 Continuous Smoothing

מפת החום עוברת פילטר גאוסי המטשטש רעשים ומאחד אזורים סמוכים, כך שהתוצאה רציפה ונקייה יותר.

3.3.2 U^2Net Underlay

בשיטה זו מכפילים מפת חום קיימת (למשל Grad-CAM) במסכת הסגמנטציה של U^2Net.

המשמעות היא שכל אזור שנמצא מחוץ לאובייקט המרכזי מאופס, גם אם שיטת ההסבר סימנה אותו כחשוב. כך ניתן לבדוק האם הסרת אזורי רקע משפרת את איכות ההסבר.

3.4 אסטרטגיות מילוי (Fill Strategies)

במבחני ההסתרה (Progressive Occlusion) יש למלא את האזורים המוסתרים באופן שלא יכניס מידע מלאכותי שעלול להטעות את המודל.[15]

לשם כך נבחנו שש אסטרטגיות שונות:

- **Black / White** – החלפת האזור בשחור או לבן מלא. שיטה פשוטה אך עלולה ליצור קצוות מלאכותיים.
- **Gray** – החלפה בצבע אפור אחיד.
- **Mean** – החלפה בערך הממוצע של הפיקסלים בתמונה או במאגר הנתונים.
- **Blur** – טשטוש גאוסי, השומר על צבעי הרקע אך מסיר את הפרטים.
- **Random Noise** – החלפת האזור ברעש אקראי.

השוואת אסטרטגיות אלו מאפשרת לבחון עד כמה אופן ההסתרה משפיע על הערכת איכות מפות החום.

4. מתודולוגיית ההערכה: מסגרת CROSS-XAI ושיפוט VLM

פרק זה מתמקד בחידוש המחקרי והמתודולוגי המרכזי של עבודה זו.
בפרק יוסבר עקרון "השיפוט העצמאי" שנועד לפתור את בעיית ההערכה המעגלית, ויתואר הPipeline המורכב מארבע שלבים עוקבים. לאחר מכן, נסקור את השופטים השונים במערכת, החל ממודלי CNN ועד למודלים שפתיים-חזותיים רחבים (VLMs) ונעמיק באתגרים ההנדסיים שכרוכים בהמרת פלט שפתי חופשי למדד כמותי רציף.

4.1 עקרון הפעולה: שבירת ההערכה המעגלית באמצעות שופט עצמאי

הבסיס הרעיוני שעליו מושתתת פלטפורמת CROSS-XAI הוא ההפרדה המוחלטת וההרמטית בין תהליך יצירת ההסבר לבין תהליך הערכתו. בשיטות המסורתיות הקיימות כיום, המודל שיצר את מפת החום הוא גם המודל שמשמש כדי להעריך אותה במבחני הסתרה – כשל מתודולוגי חמור המוכר כ"בעיית ההערכה המעגלית". במסגרת שלנו, אנו עוקרים בעיה זו מהשורש על ידי הכנסת ישות חדשה למערכת שנקרא **השופט העצמאי**.

שופט עצמאי מוגדר כמודל סיווג שאינו "מכיר" את שיטת ההסבר (Attribution) שייצרה את מפת החום, לא היה שותף בשום שלב לאימון או להסקה של מודל המקור, והוא מדרג את התמונות המוסתרות באופן עיוור לחלוטין. הוא פועל כגורם אובייקטיבי שנועד לבחון האם המידע החזותי שנשמר או הוסתר על ידי מפת החום אכן מכיל משמעות סמנטית.

הלוגיקה המנחה פשוטה ורבת עוצמה: אם אנו מסירים מהתמונה את הפיקסלים ה"חשובים ביותר" (כפי שקבעה מפת החום), ושופט עצמאי לחלוטין מאבד עקב כך את יכולתו לזהות את האובייקט – הרי שנוכל להסיק בוודאות אובייקטיבית כי מפת החום אכן הצביעה על אזורים בעלי משמעות סמנטית קריטית, ולכן ההסבר נאמן למציאות.

4.2 ארכיטקטורת – CROSS-XAI פייפליין ארבע הפאזות

מערכת CROSS-XAI תוכננה כצינור נתונים (Pipeline) מודולרי המורכב מארבע שלבים הפועלים בטור. תכנון ארכיטקטוני זה מבטיח גמישות מחקרית, הדירות מוחלטת (Reproducibility) וניהול יעיל של משאבי חומרה, שכן כל שלב יכול לרוץ, להיכשל ולעבור אופטימיזציה בנפרד מבלי להשפיע על שאר השלבים.

4.2.1 שלב 1– יצירת הסברים ומיון פיקסלים

השלב הראשון אחראי על הפקת ההסברים ממודל המקור. עבור כל שילוב אפשרי של תמונת קלט, מודל מסווג, ושיטת XAI(כגון GradCam או DINOv2). המערכת מחשבת מפת חום המדרגת את חשיבותם של הפיקסלים. הפלט של שלב זו עבור כל תמונה מורכב משני תוצרים:

1. **תמונה ויזואלית (PNG):** מיועדת לבקרה אנושית ולהצגה ויזואלית של התוצאות.

2. **מערך נתונים מתמטי (.npy):** קובץ המכיל את אינדקסי הפיקסלים ממוינים בסדר עולה, מהחשוב פחות לפחות יותר. זהו ה"מרשם" שעל פיו תתבצע ההסתרה בהמשך.

מכיוון שחילוץ גרדיאנטים ומפות חום דורש משאבי חישוב וזיכרון נרחבים, שלב זו מורץ פעם אחת בלבד. נתוניה נשמרים במטמון לשימוש חוזר בכל השלבים הבאים.

4.2.2 שלב 2 – הסתרה הדרגתית(Progressive Occlusion)

בשלב השני, המערכת קוראת את מערכי המיון משלב 1 ומתחילה בייצור פיזי של התמונות המוסתרות. תהליך ההסתרה מתבצע ב-21 רמות שונות – החל מ-0% (תמונה מקורית) ועד ל-100% הסתרה, בקפיצות רזולוציה של 5%, כאשר המטרה של 0% ושל 100% היא לוודא שכל השיטות הנפיקו את אותם ביצועים בשלבים הבאים.

עבור כל רמת הסתרה, הפיקסלים שסומנו כמועמדים למחיקה מוחלפים בהתאם לאסטרטגיית המילוי שהוגדרה מראש בניסוי (כפי שפורט בסעיף 3.4 : אפור ניטרלי, ממוצע, טשטוש וכו'). כדי למנוע צוואר בקבוק בפעולות קלט/פלט (I/O) עקב ייצור עשרות וריאציות לכל תמונה, המערכת עושה שימוש בעיבוד אצוות (Batch Processing) דינמי על גבי הGPU.

4.2.3 שלב 3 – שיפוט והסקה(Judging)

זהו שלב הליבה של מסגרת ה-CROSS-XAI שבמסגרתה מופעלים השופטים העצמאיים על גבי התמונות ה"פגומות" שיוצרו בשלב הקודם. המערכת מזרימה את התמונות המוסתרות אל השופט הנבחר וקולטת את רמת הביטחון (Confidence) שלו בזיהוי האובייקט. כדי למנוע אובדן נתונים יקרים במקרה של קריסת מערכת במהלך ריצות ארוכות, התוצאות נכתבות באופן רציף לקובצי CSV באמצעות מנגנון של Checkpoint במידה ונצטרך לעצור באמצע. התוצאות הנשמרות הן :

top1 acc: האם המחלקה הראשונה שהמודל השופט זיהה היא המחלקה שלנו.

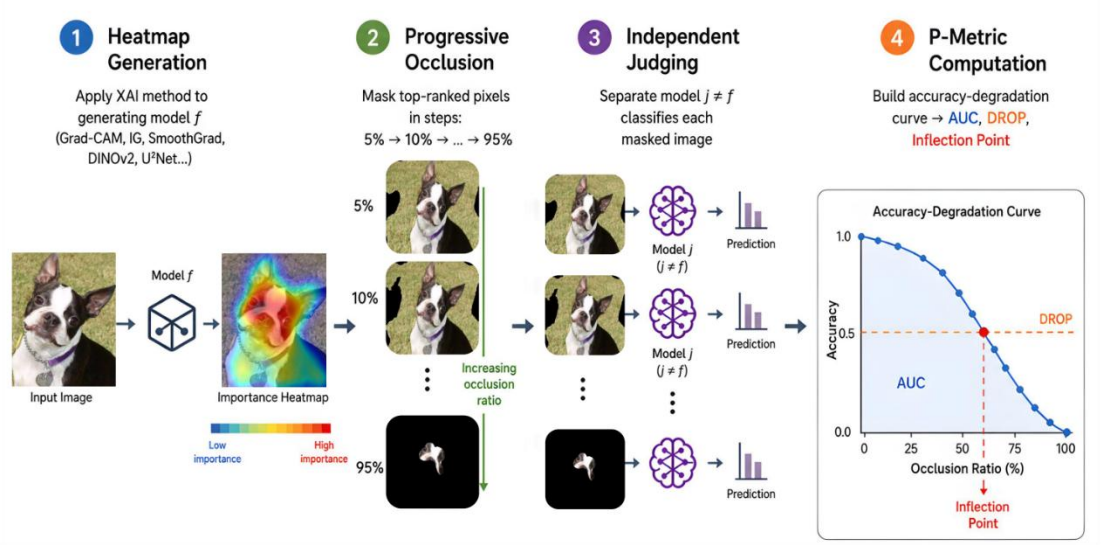
top5 acc: האם אחת המחלקות בחמישייה הראשונה שהמודל השופט זיהה היא המחלקה שלנו.

4.2.4 שלב 4 – צבירה, ניתוח והפקת מדדים

השלב האחרון קורא את כלל תוצאות השיפוט שנאספו, מצליב אותן ומחשב את המדדים הסטטיסטיים הסופיים. המדדים המרכזיים הם :

- **ציון AUC (שטח מתחת לעקומה):** מייצג את עמידות וביצועי המודל על פני כל רמות ההסתרה באופן אינטגרלי. ערך נמוך יותר מעיד על הסתרה יעילה יותר (ולכן על הסבר נאמן יותר).
- **ציון DROP:** מייצג נקודתית את צניחת הביצועים בין מצב הבסיס למצב ההסתרה המקסימלי שנבחן.

פלט שלב זו הוא דוחות נתונים מפורטים וגרפים ויזואליים להשוואה של עקומות דיוק-מול-הסתרה.



4.3 אבולוציית השופטים

בחינת תוצאות המערכת מחייבת הבנה של פריזמת השיפוט. במהלך המחקר שילבנו שלושה דורות של ארכיטקטורות שופטים, המגיבים באופן שונה מהותית למנגנוני ההסתרה:

1. **מודלי CNN (כגון ResNet):** ניחנים בשדה קלט מקומי. הם מחפשים תבניות ורציפות מרחבית, ולכן מסתמכים במידה רבה על טקסטורות וקצוות. פגיעה נקודתית והחדרת רעש מקומי נוטים להפיל את ביטחון המודל במהירות.

2. **מודלי ViT (כגון Swin Transformer):** מייצגים פרדיגמת ביניים. מחד גיסא, הם פולטים הסתברות מספרית ישירה למחלקה סגורה (כגון CNN). מאידך גיסא, העיבוד נשען על חלוקה לטוקנים ומנגנוני קשב (Attention) גלובליים. כתוצאה מכך, הם רגישים יותר לצורניות כללית וליחסי מרחב, ונפגעים פחות מחסימות פיקסליות נקודתיות ויותר מפיזור נרחב של החסימה.

3. **מודלים שפתיים-חזותיים (VLMs):** נמצאים בפסגת יכולת ההסקה. בניגוד לאופטימיזציה הצרה של זיהוי אובייקט במערך נתונים סגור, ה-VLM אומן על מסדי נתונים עצומים המשלבים טקסט ותמונות, ורכש הבנה סמנטית עמוקה של הקשרים. בעוד ששופט קלאסי יוצא מאיזון כשהוא מקבל קלט קטוע, ה-VLM מסוגל להישען על ידע עולם הקיים במודל השפה שלו, ו"להשלים את התמונה" גם משאריות דלות של האובייקט.

4.4 מודל LLaMA-3.2-Vision כשופט מתקדם במערכת

כדי לממש את שופט ה-VLM במערכת, נבחר המודל הפתוח מבית חברת מטא - **LLaMA-3.2-Vision**. בחירה זו נועדה להבטיח שקיפות מחקרית, שחזור מלא של התוצאות, ויכולת הרצה מקומית ומבוקרת על גבי חומרת GPU, מבלי להסתמך על ממשקי API סגורים (כגון אלו של OpenAI).

ארכיטקטורת המודל משלבת מקודד חזותי עמוק המעביר טוקנים חזותיים דרך שכבות השלכה (Projection layers) אל תוך מרחב הייצוג של מודל שפה גדול. תהליך זה מקודד הלכה למעשה את התמונה לרצף של "מילים" וירטואליות, המאפשרות למודל להגיב להנחיה טקסטואלית ולשפוט את התמונה המוסתרת.

4.4.1 מצבי השיפוט(Evaluation Modes)

כדי להפיק את המרב מהמודל מול ההסתרות השונות, הוגדרו במערכת שלושה מצבי תשאול:

1. **Binary Mode:** גישה ישירה שבה המודל נשאל שאלת כן/לא (למשל, "האם אתה רואה אובייקט מסוג כלב בתמונה?"). מתמקדת נטו בשרידות האובייקט.
2. **Class-ID Mode:** המודל מקבל את רשימת כלל המחלקות במערך הנתונים כשהן ממוספרות, ומתבקש להחזיר אך ורק את מספר האינדקס המדויק של המחלקה המוצגת.
3. **Cosine Similarity Mode (Open-ended):** גישה מחקרית פתוחה שבה המודל נשאל "מה אתה רואה בתמונה?". תשובתו הטקסטואלית החופשית מושווית סמנטית למחלקות המקוריות כדי למדוד את מידת החפיפה התפיסתית.

4.4.2 המרת פלט טקסטואלי למדד כמותי(Parsing)

האתגר המרכזי בשימוש במודל שפה כשופט הוא המרת הפלט הבלתי-מובנה לערך מספרי רציף. תהליך זה דרש פיתוח מנגנון עיבוד (Parsing) חסין שגיאות:

- **Structured Outputs באמצעות Pydantic:** עבור מצבי Binary ו Class-ID, המודל אולץ להחזיר מבנה JSON קפדני. בוטל הצורך בביטויים רגולריים (Regex) מועדים לפורענות; סכמת Pydantic וידאה שהמודל מחזיר שדה בוליאני או שלם, וכללה מנגנון התאוששות (Fallback) במקרה של סטייה.
- **חישובי Embedding:** עבור מצב התיאור החופשי, פלט ה-VLM הומר לווקטור באמצעות מודל הקידוד המרחבי nomic-embed-text. חושב הדמיון הקוסינוס (Cosine Similarity) בין התיאור המופק לבין וקטור המחלקה האמיתית. אם הציון עבר סף (Threshold) שהוגדר מראש, הזיהוי הוגדר כמוצלח.

4.4.3 אתגרים הנדסיים – ניהול KV-Cache ומגבלות אצווה

שילוב LLaMA-3.2-Vision דרש התמודדות עם מגבלה ארכיטקטונית כבדה. המערכת נשענה על תשתית שרת מקומית אשר מנהלת את מטריצות המטמון של הטרנספורמר (KV-Cache) בצורה מצטברת כדי להאיץ שיחות.

במהלך הפיתוח התגלה כי עיבוד תמונות שונות ברצף גורם ל"זיהום" של ה- KV-Cache. המודל החל להזות אובייקטים מתמונות קודמות לתוך התמונה הנוכחית. הפתרון ההנדסי ההכרחי היה **אכיפה נוקשה של גודל אצווה יחיד (Batch Size = 1) וניקוי אקטיבי מאולץ של ההקשר בין הרצה להרצה**. למרות המחיר החישובי והארכת זמני ההסקה, צעד זה הבטיח שיפוט סטרילי ונקי לחלוטין, שאפשר לנו להריץ את הניסוי על גבי מאגרי הנתונים במלואם עם תוקף מדעי מובהק.

5. סביבת פיתוח ומימוש הנדסי

פרק זה מפרט את המימוש ההנדסי והארכיטקטורה של פלטפורמת CROSS-XAI. המערכת תוכננה מראשיתה לספק תשתית מחקרית מודולרית, מדרגית (Scalable) ויעילה, המסוגלת להתמודד עם עיבוד נתונים רחב היקף. הפיתוח כלל אינטגרציה מורכבת של הרצת מודלי למידה עמוקה עתירי משאבים לצד שילוב מודלי שפה חזותיים גדולים (VLMs). במסגרת הפיתוח הושם דגש קפדני על אופטימיזציה וניצול מיטבי של משאבי החומרה (GPU), וכן על יישום מנגנונים נוקשים להבטחת שחזור מדויק של תוצאות הניסויים.(Reproducibility)

5.1 מתודולוגיית הפיתוח וכלים טכנולוגיים

סביבת המחקר פותחה בשפת Python, תוך הישענות על חבילות תוכנה וספריות חזיתיות מהתעשייה והאקדמיה בתחומי הבינה המלאכותית והראייה הממוחשבת:

- **PyTorch:** תשתית למידת המכונה המרכזית בפרויקט, המשמשת לאימון, הסקה וניהול מתקדם של טנסורים בסביבת GPU. עיבוד הנתונים וטעינת המודלים נתמכו באמצעות הספריות torchvision ו- torchaudio.
- **Captum:** ספרייה ייעודית מבית פרויקט Pytorch, המהווה את תשתית הליבה למימוש אלגוריתמי ההסבריות. חבילה זו אפשרה יישום יעיל ומתוקנן של שיטות מבוססות גרדיאנטים (כגון GradCam ו- Integrated Gradients).
- **TIMM ו- Transformers:** ספריות אלו שולבו על מנת לספק גישה למודלי ראייה מתקדמים כגון Vision Transformers ו-DINOv2 ולהבטיח תמיכה במגוון רחב של ארכיטקטורות קצה.
- **עיבוד תמונה ונתונים:** הספריות OpenCV ו-Pillow יושמו לטובת מניפולציות ויזואליות מדויקות (כגון חיתוך ויישום מנגנוני Occlusion). בנוסף, שולבו NumPy ו- Pandas לצורך עיבוד מתמטי, חישוב מדדים וניהול טבלאי של נתוני הניסוי.
- **הדמיה:** הספריות Matplotlib ו Seaborn-הופעלו לצורך יצירת ויזואליזציה של מפות חום והפקת גרפים סטטיסטיים להשוואת ביצועים.
- **Ollama:** פלטפורמה להרצה מקומית של מודלי שפה חזותיים (כגון Llama-3.2-Vision) אשר שימשה כבסיס לחלק מה"שופטים העצמאיים". הבחירה בהרצה מקומית נועדה לבטל תלות ב APIs-חיצוניים, להבטיח אבטחת מידע, וקריטי מכך – לשמור על דטרמיניזם מוחלט בתוצאות.

5.2 תכנון רכיבי המערכת והקוד

ארכיטקטורת התוכנה של המערכת תוכננה על בסיס עקרונות של תכנות מודולרי והפרדת רשויות (Separation of Concerns). תכנון זה מאפשר החלפה שקופה של מודלי מטרה, מסדי נתונים ושיטות XAI, מבלי לדרוש התאמות או שכתוב של ליבת הפייפליין (אשר תוארה בפרק 4).

המערכת בנויה כך שכל שלב בתהליך ההערכה מנוהלת על ידי מודול תוכנה נפרד ועצמאי, אשר קורא וכותב נתונים באופן מוסכם ומוגדר מראש. כדי לנהל את הצינור המורכב הזה ללא התנגשויות, מרכז הבקרה של סביבת הריצה מנוהל באמצעות קובץ תצורה מרכזי (config.py). קובץ זה מהווה "Single Source of Truth" עבור כלל משתני הניסוי.

גישה זו מנתקת לחלוטין בין הלוגיקה ההנדסית של הקוד לבין פרמטרי הניסוי המחקרי, ומאפשרת להריץ עשרות ניסויים שונים במקביל פשוט על ידי שינוי מערכים בקובץ הקונפיגורציה, כפי שניתן לראות בדוגמת הקוד הבאה:

קטע קוד 1: דוגמה למבנה תצורת הניסוי מתוך קובץ הconfig.py-

```
#the models we want to work on
GENERATING_MODELS = ["sipakmed_ResNet50.pth", ....]
ATTRIBUTION_METHODS = [#the methods we want to work on
    "grad_cam",
    "u2net",
    "dino_combo_ent",
    ...
]
OCCLUSION_LEVELS = list(range(0, 105, 5))# the occlusion levels
FILL_STRATEGIES = ["gray", "black", "mean"]# the fills
```

5.3 ניהול משאבי עיבוד (GPU) וביצועים

הפעלה רחבת היקף של אלגוריתמי XAI על מאגרי תמונות גדולים מהווה צוואר בקבוק חישובי משמעותי. לאור זאת, הוטמעו במערכת מנגנוני אופטימיזציה מתקדמים לניצול יעיל של זיכרון ה-VRAM:

- **ניהול זיכרון דינמי:** המערכת מנטרת את משאבי הזיכרון הזמינים ומסווגת אותם היררכית. גודל האצווה (Batch Size) מותאם בזמן אמת לקיבולת החומרה, מה שמונע שגיאות חריגת זיכרון (OOM) וממקסם את תפוקת העיבוד.
- **אופטימיזציית הקצאות (CUDA Allocator):** יושמה תצורת סביבה אשר מצמצמת משמעותית את תופעת הפרגמנטציה בזיכרון ה-CUDA ומייעלת את טעינת המשקלים.
- **דיוק מעורב (Mixed Precision) והאצת מטריצות:** לשם האצת שלב ההסקה, הופעל מנגנון USE_FP16_INFERENCE. כמו כן, נוצלה תמיכת כרטיסי המסך המודרניים (בחישובי TensorFloat-32 באמצעות פקודת ההאצה הייעודית ב- Pytorch.

- **עיבוד מקבילי (Concurrency):** פעולות צוואר-בקבוק בתהליכי קלט/פלט (I/O) ובתשאול מודלי ה-LLM מנוהלות אסינכרונית באמצעות ריבוי תהליכונים, כדוגמת MAX_WORKERS=8 לטעינת נתונים, דבר המקטין את זמני ההמתנה של ה-GPU ל-0.

5.4 הבטחת שחזור תוצאות

היכולת לשחזר תוצאות ניסוי באופן דטרמיניסטי היא נדבך קריטי במחקר מדעי מבוסס למידה עמוקה. כדי לעמוד בסטנדרט זה, יושמו במערכת מספר שכבות בקרה:

- **קיבוע אקראיות:** עבור כל ניסוי נקבע גרעין (Seed) סטטי קבוע (כגון 42) המוזן לכלל הספריות(PyTorch, NumPy, Random) . הליך זה קריטי במיוחד באינטגרציה מול מודלי השפה; שימוש בפרמטר OLLAMA_SEED מבטיח כי השופט החזותי יפיק פלט רציף וזהה לחלוטין עבור קלט נתון בכל ריצה.

- **תצורת ניסוי נעולה:** הארכיטקטורה כופה ניהול ניסויים דרך קובץ התצורה המרכזי בלבד, מה שמונע סטנסטציה של הגדרות לוקאליות בלתי-מתועדות ומאפשר מעקב גרסאות מסודר של תנאי הניסוי.

- **שמירת מצב אוטומטית:** לאור העובדה ששלבי ההערכה עשויים להימשך שעות ואף ימים, המערכת מגבה את תוצאותיה באופן תדיר. מנגנון זה מבטיח כי במקרה של כשל מערכתי, הניסוי יחודש מנקודת העצירה המדויקת ללא אובדן נתונים או כפילויות עיבוד.

- **בידוד סביבתי:** יושמו הגדרות כגון KMP_DUPLICATE_LIB_OK="TRUE" למניעת התנגשויות ברמת ספריות ה ++C-המקומפלות מתחת למנוע ה- Python על מנת להבטיח יציבות בסביבות הרצה שונות.

רשימת מקורות

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1-42.
4. Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*.
5. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32.
6. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31.
7. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Touret, S. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.09014*.
8. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., & Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106, 107404.
9. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211-252.
11. Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., & Charchanti, A. (2018). SIPAKMED: A new dataset for feature and image based classification of normal and dysplastic pap smear images. In *2018 IEEE International Conference on Image Processing (ICIP)* (pp. 3144-3148). IEEE.

12. Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3429-3437).

13. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3), 1-45.

14. Darcet, T., Oquab, M., Mairal, J., & Bojanowski, P. (2023). Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

15. Sturmfels, P., Lundberg, S., & Lee, S. I. (2020). Visualizing the impact of feature attribution baselines. Distill, 5(1), e22.

16. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

17. Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

18. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

19. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.

20. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

21. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818-833). Springer, Cham.

22. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

23. Kapishnikov, A., Bolukbasi, T., Viégas, F., & Terry, M. (2019). Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4948-4957).