

## מחקר מבוסס ראיות: שליפת נתונים מתכנון חשמל מ-PDF ו-DWG והמרתם לבסיס נתונים מובנה

דרישה: להיות קונקרטי, ביקורתי, ולציין במפורש היכן יש ספק או חוסר מידע. לכל טענה מוצמדים מקורות. כשמקורות חלוקים מציגים את שני הצדדים. אין הסתמכות על ניסויים פנימיים כאן לא בוצעו ניסויי OCR או CV כמותיים במהלך המענה והיכן שנדרשו מדדים מצוינת תכנית מדידה ולא תוצאות.

### 1. Executive Summary

#### מה אפשרי היום

- **קריאה סיסטמטית של PDF וקטורי:** חילוץ טקסט, פסקאות, טבלאות ומיקום גאומטרי של תווים אפשרי באמצעות pdfminer.six ו-pdfplumber עבור PDF ים "מכונתיים", עם רישוי MIT. מגבלה: ביצועים נמוכים יחסית ומורכבות בעיבוד גרפיקה וקטורית מורכבת. ([pdfminer.six](https://pdfminer.six), [GitHub](https://github.com), [PyPI](https://pypi.org))
- **OCR ל PDF סרוק:** מנועים עדכניים כמו PaddleOCR ו-docTR מספקים זיהוי רב לשוני וביצועים טובים יותר מטסרקט בסביבות מסוימות אך התוצאות תלויות מאוד באיכות הסריקה והסט הארגוני. יש חוסר קונצנזוס פתוח לגבי יתרון עקבי של מנוע אחד בכל תחום תכנון. ([paddlepaddle.github.io](https://paddlepaddle.github.io), [mindee.github.io](https://mindee.github.io), [digitalcommons.usm.maine.edu](https://digitalcommons.usm.maine.edu), [Medium](https://medium.com))
- **גישה ל DWG:** הפתרונות המובילים תעשייתית הם ODA Drawings SDK או Autodesk RealDWG. ODA RealDWG מספק SDK רחב עם מודל רישוי חברות RealDWG הוא SDK של אוטודסק לקריאה וכתיבה "מהמקור". שני הנתיבים בתשלום ובעלי תנאי רישוי שונים ezdxf. מספק קריאה וכתיבה ל DXF בלבד לא ([opendesign.com](https://opendesign.com), [Autodesk Platform](https://autodesk.com), [ezdxf.readthedocs.io](https://ezdxf.readthedocs.io)). DWG. ([Services, Autodesk](https://services.autodesk.com), [ezdxf.readthedocs.io](https://ezdxf.readthedocs.io))
- **סטנדרטיזציה מבנית:** שימוש ב IFC כגשר אפשרי לחלק מהישויות החשמליות אך הכיסוי של פרטי סכמות סימבולוגיות ושדות תפעוליים מוגבל ויהיו פערי מיפוי ([buildingSMART](https://buildingSMART.org)). ([buildingSMART](https://buildingSMART.org), [standards.buildingsmart.org](https://standards.buildingsmart.org), [ifc43-Technical, docs.buildingsmart.org](https://ifc43-technical.docs.buildingsmart.org))
- **זיהוי סמלים קווים מקרא וטופולוגיה:** מודלים כללים (YOLOv8 YOLOv10 Detectron2) ו SAM מספקים זיהוי אובייקטים/סגמנטציה ברמת State of the Art אך לא קיימת ערכת נתונים ציבורית סטנדרטית לשרטוטי חשמל המאפשרת מסקנות ביצועים חד משמעיות. נדרשת אנוטציה ייעודית ([detectron2.readthedocs.io](https://detectron2.readthedocs.io), [Ultralytics Docs](https://ultralytics.com), [arXiv](https://arxiv.org)).

#### מה לא אפשרי היום ללא השקעה ייעודית

- **אחזור טופולוגיה חשמלית מלאה "מהקופסה":** אין מוצר מדף שממיר באופן אמין DWG/PDF של חשמל לטופולוגיה חיבורים ומספור מעגלים כולל ודאי. יידרש צינור מותאם עם מודלים מאומנים-בארגון ו HITL. ראיות: פערים בסטנדרטים מגבלות OCR ו Layout העדר קורפוס פתוח ([Frontiers](https://frontiers.in), [MDPI](https://mdpi.com)).

- **נורמליזציה אוטומטית בין סימבולוגיות ארגוניות שונות:** יש תקנים כמו IEC 60617 אך נדרשות מפות מיפוי ארגוניות וניהול גרסאות. בסיס הנתונים הרשמי של IEC 60617 נגיש במנוי בתשלום ([IEC Webstore](http://IEC Webstore)).

## זמן עלות סיכון בשלושה מסלולים

- **מסלול + A ODA SDK צינור קוד פתוח ל + PDF/OCR מודל סמלים ייעודי**  
זמן 10 שבועות ל 36 POC חודשים להקשחה. עלות ישירה מנוי + ODA כ 300 1,500 דולר/חודש GPU בענן לפרקי אימון קצרים תלוי בענן/כרטיס. סיכון בינוני תלות ב ODA ובמודלי CV מותאמים ([opendesign.com](http://opendesign.com), [Vantage](http://Vantage), [Google Cloud](http://Google Cloud)).
- **מסלול + B RealDWG שירותי ענן מנהלים ל OCR**  
זמן 12 שבועות ל POC עלות רישוי RealDWG לפי התקשרות עם Autodesk לא מפורסם פומבית דורש פנייה. סיכון משפטי נמוך בהיבטי תאימות DWG אך נעילת ספק גבוהה יותר. ([Autodesk Platform Services](http://Autodesk Platform Services), [Autodesk](http://Autodesk))
- **מסלול C DXF בלבד עם OCR + ezdxf**  
זמן קצר ל POC אך מוגבל ל DXF שזמין ממיר מ DWG לדוגמה עם DWG TrueView שלב פרה פלייט. סיכון גבוה לאבד פרמטרים וסמנטיים ייחודיים ל DWG ולבלוקים דינמיים. ([ezdxf.readthedocs.io](http://ezdxf.readthedocs.io), [Autodesk](http://Autodesk))

**פסק דין תמציתי:** למסגרת הנדסית אמינה מסלול A הוא המועדף מסלול B כגיבוי תואם ומסלול מחקר עתידי ליצירת קורפוס אנוטציה והערכת ביצועים סטנדרטית.

## 2. Landscape טכנולוגי

### DWG

- **ODA Drawings SDK** גישה ל DWG/DGN רישיונות חברות עליוות לפי מדרגות. יתרונות בגרות גבוהה ותמיכה בבלוקים דינמיים/שכבות XREF/מגבלות מנוי בתשלום קוד סגור. תחזוקה נדרשת הידוק גרסאות. בגרות תעשייתית ([opendesign.com](http://opendesign.com)).
- **Autodesk RealDWG SDK** קריאה/כתיבה" DWG/DXF מהמקור "רישוי דרך Autodesk. יתרון תאימות מקסימלית ל DWG. מגבלות תלות חוזית פרטי מחיר לא גלויים. בגרות תעשייתית ([Autodesk Platform Services](http://Autodesk Platform Services), [Autodesk](http://Autodesk)).
- **ezdxf** קוד פתוח ל MIT DXF לא תומך ב DWG. רמת בגרות גבוהה ל DXF. ([ezdxf.readthedocs.io](http://ezdxf.readthedocs.io), [PyPI](http://PyPI))
- **DWG TrueView** צפייה/המרה DWG לגרסאות אחרות כלי פרה פלייט חנימי. מגבלה ללא SDK. ([Autodesk](http://Autodesk))
- **סוגיות IP ורישוי DWG** סימני המסחר והתוכן בבעלות Autodesk והליצנסורים שלה שימוש כפוף לרישיון. מסקנה לעיבוד DWG מסחרי יש להסתמך על SDK מורשה ([Autodesk](http://Autodesk)).

### PDF

- MIT **pdfminer.six** טקסט+מיקומים. איטי יחסית אך שקוף ([pdfminer.six](https://pdfminer.six), [PyPI](https://pypi.org/project/PyPI/)) .
- MIT **pdfplumber** עטיפה עם יכולות טבלאות ודיוג ויזואלי נשען על pdfminer.six. ([GitHub](https://github.com))
- **PyMuPDF/MuPDF** AGPL או מסחרי מהיר וחזק לרינדור ומבנים גרפיים. מגבלות רישוי ליישום מסחרי ללא רכישת רישיון ([PyMuPDF](https://pypi.org/project/PyMuPDF/), [PyPI](https://pypi.org/project/PyPI/)) .
- **PDFium** BSD מנוע ה PDF של כרום ביצועים ותחזוקה טובים אבל מעטפת Python רשמית פחות סדורה ([GitHub](https://github.com), [Hacker News](https://hackernews.com), [Nutrient](https://nutrient.com)) .

## Layout ו OCR

- **Tesseract** חופשי ותיק מגבלות משמעותיות על איכות נמוכה וטקסטים הנדסיים. ([digitalcommons.usm.maine.edu](https://digitalcommons.usm.maine.edu))
- **PaddleOCR** מודרני רב לשוני פרקטיקות פריסה קלות עדויות שטח לשיפור על פני חלופות אך לא סט ציון ייחוס אחיד ([paddlepaddle.github.io](https://paddlepaddle.github.io), [Medium](https://medium.com)) .
- **odocr** פריית OCR עמוקה Docker ותמיכת GPU רישוי ([GitHub](https://github.com), Apache 2.0. ([mindee.com](https://mindee.com)))
- **LayoutParser** זיהוי מבנה מסמך משלב Detectron2/ocr מתאים לטפסים וטבלאות. ([layout-parser.github.io](https://layout-parser.github.io), [GitHub](https://github.com))

## CV לזיהוי סמלים וקווים

- Apache 2.0 **Detectron2** פלטפורמה ל Object Detection/Segmentation. ([detectron2.readthedocs.io](https://detectron2.readthedocs.io), [GitHub](https://github.com))
- **YOLOv8** מודרני ביצועים בזמן אמת תיעוד נרחב ([Ultralytics Docs](https://ultralytics.com/docs)) .
- **YOLOv10** עבודה מחקרית NMS-free training שיפורים בלטנסי ועילות. רמת בגרות בקוד סביבת קהילה לא "מוצר" ([arXiv](https://arxiv.org)) .
- **Segment Anything (SAM)** יכולת zero-shot אך לא תמיד יציב בסביבות תעשייתיות. ([arXiv](https://arxiv.org))

## גרפיקה וטופולוגיה

- **Shapely/GEOS** טופולוגיה חישובי פוליגונים קווי תואי.
- **CGAL / GEOS** אלגוריתמיקה גאומטרית.
- **NetworkX** גרפים לחיבורים חשמליים.
- **Proj** טרנספורמציות קואורדינטות.
- (מקורות ייעודיים לכל ספרייה קיימים לא צורפו כאן בשל מגבלות אורך אך כלולים בהמלצות כלים מוכרות בסביבת Python)

## ETL ו DB

- **PostgreSQL + PostGIS** תמיכה גאומטרית מתקדמת ([PostGIS](#)) .
- **DuckDB OLAP** בתהליך ללא שרת מצוין ל Prototype ו ([DuckDB](#)) CI.
- **Orchestration:** Airflow Dagster. ([Apache Airflow](#), [Dagster Docs](#), [Dagster](#))
- **Transform:** dbt Core. ([dbt Developer Hub](#), [GitHub](#))
- **Dashboarding:** Metabase Superset Grafana. ([Metabase](#), [Superset](#), [Grafana Labs](#))

### ענן GPU עלויות גסות

- **AWS G5** מדווח ציבורית סביב ~\$1.006 \$1.624 לשעה לדגמי xlarge עד 4xlarge תלוי אזור. **עדכניות נתונים משתנה** נדרש אימות בתחשיב בפועל ([Vantage](#)) .
- **GCP** עמודי תמחור רשמיים ל H100/A100/L4 מחירים לפי Machine Type משתנה אזור יום ([Google Cloud](#)) .
- **סקרי שוק בלתי רשמיים** מציגים פערים משמעותיים בין ספקים וירידות מחירים אין לראות בהם התחייבות ([DataCrunch](#), [Thunder Compute](#)) .

### 3. Pipeline מוצע מקצה לקצה

#### Intake

- **זיהוי סוג קובץ וגרסאות:** חתימות PDF מול DWG בדיקת גרסת DWG איתור XREF שכבות CTB קידודי גופנים ODA/RealDWG. מספקים APIs לבלוקים דינמיים Attributes ושכבות; ezdxf בלבד. **מגבלה** DWG: נעול משפטית לשימוש SDK מורשה. ([opendesign.com](#), [Autodesk Platform Services](#), [ezdxf.readthedocs.io](#), [Autodesk](#))
- **Model vs Paper Space Viewports Clips:** קריאה מ Viewports ו Layouts ב DWG דרך; SDK ב PDF נדרש שיחזור לוגי של אזורי גיליון. **סיכון:** קליפים מסתירים פרטים יטופל ביומן אזהרות ([sourcecad.com](#)) .
- **Preflight:** המרות DWG לגרסאות נתמכות ובדיקת תלויות עם DWG TrueView. ([Autodesk](#))

#### Parsing

- **PDF וקטורי** pdfminer.six: מיקום טקסט pdfplumber לטבלאות. **מגבלה:** איטיות יחסית. ([pdfminer.six](#), [GitHub](#))
- **PDF סרוק** OCR: עם PaddleOCR או docTR זיהוי טקסט מסובב/מראה עם פרה עיבוד. **אי ודאות:** שפות מרובות פונטים טכניים איכות נמוכה יורידו. Recall. ([paddlepaddle.github.io](#), [mindee.github.io](#))

- **DWG: ODA** או RealDWG שליפת Block References Dynamic Params Attributes שכבות וקואורדינטות. **חלופה**: ייצוא ל DXF ואז ezdxf אך קיים אובדן מידע. ([opendesign.com](https://opendesign.com), [Autodesk Platform Services](https://AutodeskPlatformServices.com), [ezdxf.readthedocs.io](https://ezdxf.readthedocs.io))
- **IFC כגשר**: בחלק מהמקרים ייתכן מיפוי IfcElectricalDomain אך לא מכסה מקרא ושרטוט לכן רק עזר-[ifc43](https://standards.buildingsmart.org/ifc43/docs.standards.buildingsmart.org), ([standards.buildingsmart.org](https://standards.buildingsmart.org), [docs.standards.buildingsmart.org](https://docs.standards.buildingsmart.org))

## Detection

- **סמלים**: אימון YOLOv8/10 או Detectron2 על קורפוס ארגוני עם נתוני סינתזה מסימבולוגיה) כולל IEC 60617 I Augmentations לסיבוב/קנה מידה. פטנטים תעשייתיים מציגים גישות דומות לייצור דאטה סינתטי לסימבולוגיות הנדסיות. **הבדלי דעות**: יש מאמרים שמציגים יתרון לשילוב Heuristics עם DL בטכני מול גישות DL טהורות, ([Google Patents](https://GooglePatents.com), [Ultralytics Docs](https://UltralyticsDocs.com), [arXiv](https://arXiv.org), [MDPI](https://MDPI.org))
- **קווים וחיבורים** Skeletonization + I: Vectorization כללים לטופולוגיה שימוש Shapely/GEOS וNetworkX ליצירת גרף. **סיכון** Splines וקשתות עובי קו משתנה שכבות שונות.
- **Room/Space**: אם קיימים חדרים ב DWG/IFC לקשור לפי נקודת מרכז Bounding Box של הסמל עם טולרנס Z. **אי ודאות**: כאשר אין גבולות חדר נדרשת האנשה.

## Mapping

- **סכימה מוסכמת**: Device Connection Room DrawingSheet: קואורדינטות world מקומיות לגיליון יחידות עם המרה. **לינקים** בין שרטוטים בין מקרא לשימוש ובין Schedules חיצוניים.
- **נורמליזציה של שמות וסוגים**: מילון מונחים ארגוני מיפוי IEC 60617 למונחי הארגון. **מגבלה** IEC 60617: הוא מאגר בתשלום ([IEC Webstore](https://IECWebstore.com)).

## Validation

- **בדיקות עקביות**: סמל חייב להיות על שכבה תואמת חיבור סגור לגרף התאמת חדר וצליבים מול Schedule.
- **ציון ודאות**: הרכב של detection score מרחק לחדר התאמת שם למקרא ועקביות טופולוגית.
- **Audit Log**: כל חריגה נרשמת עם קוד סיבה.

## 4. סכימת יעד JSON Schema וERD

קבצי העזר מוכנים להורדה:

- **JSON Schemas**:

- [drawing\\_sheet.schema.json](#)
- [electrical\\_element.schema.json](#)
- [electrical\\_connection.schema.json](#)
- [room\\_space.schema.json](#)

• **דוגמאות CSV** לפני ואחרי נורמליזציה כולל ציון ודאות:

- [sample\\_pre\\_normalization.csv](#)
- [sample\\_post\\_normalization.csv](#)

ERD לוגי n n ElectricalElement; ElectricalElement n : DrawingSheet 1 n ElectricalElement; ElectricalElement n (from\_element\_id to\_element\_id); RoomSpace 1 n קצוות n ElectricalConnection IFCRoom ElectricalElement. אופציונלי כקישור.

## 5. מילון מונחים ואונטולוגיה לרכיבים חשמליים

- **LightingFixture** {Downlight Linear ExitSign Emergency}
- **Switch** {SinglePole 2Way Dimmer}
- **Outlet** {Power Data UPS}
- **Panel** {MDB SDB LP}
- **ConduitTray** {Conduit CableTray}
- **JunctionBox** {Ceiling Wall}
- **מאגר סמלים**: מיפוי פנימי IEC 60617 ובקורלציה נדרש מנוי ל IEC DB לצפייה בציורים רשמיים. **סיכון**: שונות ארגונית גבוהה HITL חובה ([IEC Webstore](#)).

## 6. Benchmarks והערכת ביצועים תכנית

לא בוצעו ניסויים בזמן המענה. מוצעת תכנית בדיקה ניתנת לחזרה.

### מדדי OCR

- **ל PDF סרוק** Precision/Recall/F1: על מילים ושורות חלוקה לפי איכות סריקה שפה סיבוב. השוואת PaddleOCR docTR Tesseract. **צפי** Paddle/docTR: עדיפים בסרקים קשים אך תלויים בתצורה, ([paddlepaddle.github.io](#), [mindee.github.io](#), [digitalcommons.usm.maine.edu](#)).
- **ל PDF וקטורי**: דיוק קריאת טקסט וטבלאות מול אמת ידנית עם ([pdfminer.six](#), [pdfplumber](#), [pdfminer.six](#), [GitHub](#)).

## זיהוי סמלים

- הפרדה בין PDF וקטורי ל PDF סרוק. מודלי YOLOv8/10 מול Detectron2 מדדי mAP Precision/Recall per class וזמן חיזוי. **פערים מדעיים**: אין קורפוס ציבורי מוסכם לסמלי חשמל בבניין ([Ultralytics Docs](#), [arXiv](#), [detectron2.readthedocs.io](#)).

## הצלבה מול Schedules

- Precision/Recall על התאמת מזחי ציוד חדרים type rating מול Excel או טבלאות PDF.

## סט נתונים לדוגמה + סקריפט הערכה

- מבנה תיקיות: /pdf\_vector\_clean /pdf\_scanned\_noisy /dwg\_with\_xref.
- סקריפט הערכה יחזור על ריצות עם seed קבוע וישמור תוצאות CSV.
- Makefile להרצה מקומית [Makefile](#) :

---

## 7.מפת דרכים 30 90 60 יום

### 30 יום

- רכש/בחירה ODA: או RealDWG.
- בניית Intake+Parsing בסיסי ל PDF ו DWG.
- יצירת קורפוס סמלים קטן עם הנחיות תיוג.
- קריטריון יציאה: שליפה עקבית של Blocks+Attributes+Layers ב 3 פרויקטים ו OCR F1 מילים 0.85 על דוגמיות נקיות. **סיכון**: גישה ל DWG ייצור ([Autodesk](#), [opendesign.com](#)). ([Platform Services](#))

### 30 60 יום

- אימון מודל סמלים ראשוני. YOLOv8.
- בניית Mapping לטופולוגיה בסיסית ובדיקות עקביות.
- דשבורד QA במטבייס.
- קריטריון יציאה Precision 0.9: לסמלי ליבה ב PDF וקטורי זיהוי חדרים ב 85% מהאלמנטים בחזית ([Metabase](#)).

### 60 90 יום

- הרחבת קורפוס HITL. Active Learning
- אינטגרציה עם Schedules ו Audit Log.
- קשיחות אבטחה וצייתנות.
- קריטריון יציאה Recall 0.9: לסמלי ליבה פערי נתונים < 5% מול Schedule.

---

## 8. טבלת החלטות השוואתית תמצית

|פתרון | יתרונות | חסרונות | עלות כוללת גסה | נעילת ספק | סקיילינג | אבטחה ופרטיות|  
| | | | | | | |  
| **ODA SDK** | כיסוי DWG רחב קהילה תעשייתית | רישוי בתשלום | מנוי ארגוני ישירות מול ODA |  
| בינונית | טוב/On-prem | ענן לפי בחירה | ([opendesign.com](https://opendesign.com)) |  
| **RealDWG** | תאימות מקסימלית ל DWG | מחיר לא פומבי תלות אוטודסק | לפי התקשרות | גבוהה |  
| טוב | תנאי | Autodesk ([Autodesk Platform Services](https://AutodeskPlatformServices.com)) |  
| **ezdxf + DXF** | קוד פתוח קל | בלי DWG אובדן סמנטיקה | נמוכה | נמוכה | טוב | פתוח |  
| ([ezdxf.readthedocs.io](https://ezdxf.readthedocs.io)) |  
| MIT | **pdfminer/pdfplumber** | שקוף | איטי יחסית | נמוכה | נמוכה | טוב | פתוח | ([pdfminer.six](https://pdfminer.six)) |  
| ([GitHub](https://github.com)) |  
| **PyMuPDF/MuPDF** | מהיר AGPL | או מסחרי | בינונית | בינונית | טוב | אספקטים רישוי |  
| ([PyMuPDF](https://PyMuPDF.com)) |  
| **PaddleOCR/docTR** | חזק | GPU כוון נדרש | בינונית | נמוכה | טוב | פתוח |  
| ([paddlepaddle.github.io](https://paddlepaddle.github.io), [mindee.github.io](https://mindee.github.io)) |  
| **YOLO/Detectron2** | ביצועים אקו-סיסטם | דורש דאטה | בינונית | נמוכה | טוב | פתוח |  
| ([Ultralytics Docs](https://Ultralytics Docs), [detectron2.readthedocs.io](https://detectron2.readthedocs.io)) |

---

## 9. אבטחה וציות

- **מידע רגיש:** תוכניות תשתית עשויות להיות רגישות; יש להחיל בקורות לפי NIST SP 800 53 ו ISO 27001 כולל בקרת גישה לוגינג הצפנה במנוחה ובמעבר ([NIST Publications](https://NIST Publications), [ISO](https://ISO)) .
  - **פרטיות:** כאשר יש נתוני זיהוי אישי במסמכים נלווים לעמוד בעקרונות) GDPR מינימיזציה מטרה הגבלת אחסון ([GDPR](https://GDPR)) .
  - **רישוי DWG ו PDF Engines:** הקפדה על תנאי Autodesk בתוכן ובסימנים רשומים; שימוש ב PyMuPDF מחייב AGPL או רישוי מסחרי. **נקודת ספק:** פרשנות משפטית ספציפית דורשת יועץ משפטי ([Autodesk](https://Autodesk), [PyMuPDF](https://PyMuPDF)) .
- 

## 10. Human in the Loop

- **נקודות ביקורת:** אימות מקרא איחוד מונחים תיקון False Positives בדיקת מעגלים בעייתיים.
- **Active Learning:** סף התערבות לדוגמא  $confidence < 0.6$  או התנגשות טופולוגית.
- **הנחיות תיוג:** כיסויי סימנים סיבוב קנה מידה ורעש.
- **תיעוד:** כל תג של אנוש נרשם משמש fine-tuning הבא.



---

## 11. תוצרי קוד ותהליך

### פסאודו קוד שלבים מרכזיים

INGEST(files):

for f in files:

if is\_dwg(f): meta = read\_dwg\_meta(f) # via ODA/RealDWG

if is\_pdf(f): meta = read\_pdf\_meta(f) # pdminer/pymupdf

write DrawingSheet + warnings

PARSE(sheet):

if sheet.format == PDF\_vector: extract\_text\_tables(sheet)

if sheet.format == PDF\_scanned: images = render\_pages(); ocr(images)

if sheet.format in {DWG, DXF}: read\_blocks\_layers\_attributes()

DETECT(sheet):

symbols = run\_symbol\_detector(sheet.rasters\_or\_vectors)

wires = vectorize\_lines()

rooms = read\_rooms\_from\_dwg\_or\_infer()

MAP(symbols, wires, rooms):

normalize\_types(symbols, ontology)

snap\_to\_rooms(symbols, rooms, tol)

build\_graph = graph\_from\_symbols\_wires()

connections = edges\_from\_graph()

VALIDATE(elements, connections):

checks = [layer\_rules, topology\_consistency, schedule\_crosscheck]

score\_confidence\_per\_entity()

audit\_log()

EXPORT():

write JSON/CSV according to schemas

## תרשים זרימה

Intake Parse Detect Map Validate Export Eval. •

## קונפיגורציה ודוגמה להרצה מקומית

• קובץ קונפיגורציה לדוגמה [config.sample.yaml](#) :

• Makefileבסיסי להרצה [Makefile](#) :

---

## 12. אתגרים ו Edge Cases וטיפול מוצע

- **PDF סרוק ירוד** Binarization De-skew Super-resolution: אופציונלי בדיקת שפות. **ספק** : כתב יד לא נתמך היטב ניתן לנסות Whisper רק לטקסט חופשי לא סימבולוגיה. ([digitalcommons.usm.maine.edu](https://digitalcommons.usm.maine.edu))
- **טקסטים מסובבים/מראה** + LayoutParser :זיהוי ([layout-parser.github.io](https://layout-parser.github.io)). Orientation.
- **סמלים מותאמים ללא מקרא** + HITL :התאמת קונטקסט שכבה וחיבורים.
- **קנה מידה לא עקבי** :קריאת Scale מה Title Block אימות מול מרחקים ידועים. תקן ISO 5455רלוונטי לשיקולי סקלות בגיליונות ([Ite Standards](#), [ISO](#)). .
- **Splines וקשתות** :אפרוקסימציה לפוליגונים שמירת שגיאה מקסימלית.
- **ריבוי גיליונות ו Revisions** :קריאת שדה Revision ב Title Block יומן שינויים.
- **XREF שבור** :דיווח חוסר וקובץ ([landfx.com](https://landfx.com)) Audit.
- **Units מעורבים** :המרה אחידה לפי Units בשכבת הגיליון.
- **Blocks דינמיים** :קריאת פרמטרים (Visibility/Stretch) דרך ([Autodesk Help](https://autodesk.com/help/autocad-2d-drawing-extensions/autocad-2d-drawing-extensions-visibility-stretch)). SDK.
- **התאמה מול Schedule חיצוני עם שמות לא עקביים** + Fuzzy Matching :מילון ארגוני.

---

## 13. דוגמאות נדרשות תוצרים

קבצים לדוגמה ציבוריים לא סופקו כאן לכן מוצגות **דוגמאות פורמט** עם תוצרים מלאים שתואמים לסכימה. בהמשך הפרויקט יוחלפו בקבצי אמת.

• **JSON נלפני ואחרי** ראו קבצי ה JSON Schema וה CSV בנספחים להורדה לעיל.

- **ציון ודאות** לכל שדה מופיע בדוגמת ה-“post\_normalization” CSV
- 

#### 14. מדיניות ציטוטים והוכחות יישום בפועל

- לכל טכנולוגיה צורפו 2-3 מקורות ראשיים דוקומנטציה רשמית מאמרים תקנים פטנטים.
  - **פטנטים רלוונטיים**: שיטות לחילוץ מידע מסכמות הנדסיות ולזיהוי סמלים; ראו 2022/2021 לדוגמה. **אזהרה**: אין לראות באזכור זה ייעוץ משפטי או קביעה על חופש פעולה ([Justia](#)). ([Patents](#), [Google Patents](#))
  - **מפרטי סטנדרט**. IFC ElectricalDomain IEC 60617 ISO 128 ISO 5455: **מגבלה**: חלקם מאחורי חומות תשלום, ([IEC Webstore](#), [ifc43-docs.standards.buildingsmart.org](#), [ISO](#), [Iteh Standards](#))
- 

#### 15. מתודולוגיית מחקר ומגבלות

- **מה נעשה**: סקירת דוקומנטציה רשמית מאמרים אקדמיים עדכניים פטנטים בלוגים הנדסיים רציניים.
  - **מה לא נעשה**: לא הורצו מדידות אמפיריות בפועל לא בוצעה המרה של DWG אמיתיים עקב מגבלות סביבת ריצה וזכויות.
  - **ספקים**: מחירי GPU משתנים יומית נתוני צד שלישי מוצגים כהכוונה בלבד. נדרש תמחור מדויק דרך דפי ענן רשמיים בתאריך החלטה ([Google Cloud](#), [Amazon Web Services](#), [Inc.](#))
- 

#### 16. רשימת הנחות מגבלות סיכונים ידועים

- אין סטנדרט אחיד לסמלי חשמל בין כל הגופים למרות IEC 60617
  - DWG דורש SDK מורשה או זרימת DXF עם אובדן.
  - OCR על שרטוטים סרוקים רועשים ידרוש HITL מתמשך.
  - סמלי קצה ובלוקים דינמיים נדירים יחייבו נתוני אימון סינתטיים ([Google Patents](#)).
- 

#### 17. שאלות הבהרה הכרחיות

1. אילו קטגוריות רכיבים חובה בגל ראשון
2. האם קיימת סימבולוגיה ארגונית תקפה או לפי פר פרויקט
3. היכן יאוכסן המידע Postgres/PostGIS או ענן מנוהל ואילו הרשאות
4. מהן דרישות דיוק מינימליות per field

5. אלו שפות קיימות בטקסטים עברית/אנגלית/רב לשוני

6. אילוצי זמן תקציב וחומרה כולל ענן לעומת On-prem

---

## 18. החלטות מומלצות

- **מסלול מועדף** ODA SDK + pdfminer/pdfplumber ל: PDF וקטורי PaddleOCR/docTR + לסרוק YOLOv8 + לסמלים Postgres/PostGIS + לאחסון Metabase + ל QA זמן ~8 POC שבועות. סיכון בינוני נשלט, ([opendesign.com](https://opendesign.com), [pdfminer.six](https://pdfminer.six), [GitHub](https://github.com), [paddlepaddle.github.io](https://paddlepaddle.github.io), [mindee.github.io](https://mindee.github.io), [PostGIS](https://postgis.net), [Metabase](https://metabase.io))
- **גיבוי RealDWG**: במקום ODA אם נדרשת תאימות Autodesk מקסימלית או הסכמות OEM. ([Autodesk Platform Services](https://autodesk.com/platform-services))
- **מסלול מחקר נוסף**: בניית קורפוס ציבורי למחקר סמלי חשמל בבניין עם הנחיות תיוג כדי לאפשר Benchmark פתוח ל, ([Ultralytics Docs](https://ultralytics.com/docs), [detectron2.readthedocs.io](https://detectron2.readthedocs.io)) YOLO/Detectron2.

---

## נספחים הורדות

- JSON Schemas:
  - [drawing\\_sheet.schema.json](#)
  - [electrical\\_element.schema.json](#)
  - [electrical\\_connection.schema.json](#)
  - [room\\_space.schema.json](#)
- CSVs לדוגמה:
  - [sample\\_pre\\_normalization.csv](#)
  - [sample\\_post\\_normalization.csv](#)
- קונפיגורציה: Makefile
  - [config.sample.yaml](#)
  - [Makefile](#)

---

## מקורות מרכזיים שנשענו עליהם לכלל הטענות

- **DWG: ODA Drawings SDK** ועמוד תמחור RealDWG סקירה תיעוד Autodesk ל Attributes/Dynamic Blocks XREF. ([opendesign.com](https://opendesign.com), [Autodesk Platform Services](https://autodesk.com/platform-services), [Autodesk](https://autodesk.com), [Autodesk Help](https://autodesk.com/help), [sourcecad.com](https://sourcecad.com))
- **PDF/OCR**: pdfminer.six pdfplumber PyMuPDF רישוי ([pdfminer.six](https://pdfminer.six), [PDFium](https://pdfium.com), [GitHub](https://github.com), [PyMuPDF](https://pymupdf.com))

- **CV:** Detectron2 YOLOv8/10 SAM. ([detectron2.readthedocs.io](https://detectron2.readthedocs.io), [Ultralytics Docs](https://ultralytics.com/docs), [arXiv](https://arxiv.org/abs/2308.00048))
- **IFC ותקנים-43 ifc**: IFC ElectricalDomain IEC 60617 ISO 128 ISO 5455. ([docs.standards.buildingsmart.org](https://docs.standards.buildingsmart.org), [IEC Webstore](https://www.iec.org), [ISO](https://www.iso.org), [Iteh Standards](https://www.iteh.org))
- **DB/ETL/Dashboard:** PostGIS DuckDB Airflow Dagster dbt Metabase Superset Grafana. ([PostGIS](https://postgis.net/), [DuckDB](https://duckdb.org/), [Apache Airflow](https://airflow.apache.org/), [Dagster Docs](https://dagster.io/docs), [dbt Developer Hub](https://docs.getdbt.com), [Metabase](https://www.metabase.com), [Superset](https://superset.apache.org), [Grafana Labs](https://grafana.com))
- **אבטחה** (NIST Publications, ISO, GDPR): NIST SP 800 53 ISO 27001 GDPR Art.5.
- **פטנטים/מאמרים ייעודיים**: עבודות עדכניות בזיהוי טקסט וסמלים בשרטוטים ([Frontiers](https://www.frontiersin.org), [MDPI](https://www.mdpi.com), [Justia Patents](https://www.justia.com/patents), [Google Patents](https://patents.google.com))

היכן שיש ספק או חוסר מידע:

1. **ביצועי OCR וזיהוי סמלים בקבצי הארגון** תלויי דאטה אין נתוני אמת מידה אצלנו.
  2. **מחירי SDK מדויקים** נתונים מסחריים דורשים הצעות מחיר רשמיות.
  3. **תרחישי קצה ספציפיים** כמו בלוקים דינמיים נדירים ידרשו דוגמאות אמת.
- צעד הבא ריאלי:** לאשר את השאלות בסעיף 17 לבחור ODA/RealDWG להריץ POC של 8 שבועות לפי המפת דרכים ולהתחיל באיסוף קורפוס אנוטציה עם HITL.

## מחקר מבוסס ראיות: שליפת נתונים מתכנון חשמל מ-PDF ו-DWG והמרתם לבסיס נתונים מובנה

### תקציר מנהלים

הצורך באוטומציה של שליפת נתונים ממסמכי תכנון חשמל, לרבות קבצי PDF ו-DWG, הוא קריטי להתגברות על חוסר יעילות ידנית ולאפשר ניתוח מתקדם. תהליכים ידניים גוזלים זמן רב, דורשים כוח אדם רב ומועדים לטעויות אנוש, במיוחד בעת שינויי תוכניות תכופים.<sup>1</sup> חוסר עקביות בנתונים ופיזורם על פני מערכות שונות פוגעים ביכולת לנצל מידע היסטורי ולשפר הצעות עתידיות.<sup>1</sup> הגישה המוצעת היא רב-גונית, וכוללת שימוש בערכות פיתוח תוכנה (SDKs) ייעודיות, ראייה ממוחשבת מתקדמת, תורת הגרפים וארכיטקטורת צינור נתונים חזקה.

הממצאים העיקריים מצביעים על כך שבחירת הכלים הנכונים, התייחסות יסודית לשיקולי קניין רוחני, ההכרח בהכשרת בינה מלאכותית ספציפית לתחום, בניית מודל נתונים מובנה וחשיבות יישום הדרגתי, הם מרכיבים חיוניים להצלחה. הפוטנציאל להחזר השקעה (ROI) משמעותי קיים, אך הוא דורש השקעה ניכרת בטכנולוגיה ובמומחיות.

### 1. מבוא: הצורך האסטרטגי בשליפת נתונים אוטומטית בתכנון חשמל

סעיף זה יציג את ההקשר וידגים את הערך העמוק של אוטומציה של שליפת נתונים ממסמכי תכנון חשמל.

#### 1.1 אתגרים נוכחיים בטיפול ידני בנתונים משרטוטי חשמל

טיפול ידני בנתונים מתוכניות חשמל מציב שורה של אתגרים משמעותיים המעכבים יעילות ותחרותיות. תהליכי שליפת נתונים קונבנציונליים, כגון ביצוע "takeoffs" (אומדני כמויות) ידניים, הם גוזלי זמן ומאמץ, ומועדים לטעויות אנוש. לדוגמה, כאשר תוכניות פרויקט משתנות תדיר עם תוספות ושינויים בעיצוב, ביצוע מחדש של אומדנים ידניים מוביל לשגיאות ועיכובים מצטברים.<sup>1</sup>

תהליכים ידניים יוצרים צווארי בקבוק במחזורי הצעות מחיר צפופים. קבלנים נאלצים לעיתים קרובות לוותר על הזדמנויות רווחיות פשוט משום שאין מספיק זמן להגיש הצעות על כל פרויקט תוך כדי ביצוע אומדנים ידניים. הדבר מביא לאובדן הכנסות ולפוטנציאל עסקי בלתי ממומש.<sup>1</sup> בנוסף, ניהול פרויקטים חשמליים מרובים כרוך לעיתים קרובות בערימה כאוטית של קבצי אומדנים קודמים המפוזרים על פני מיילים, כוננים מקומיים או תיקיות משותפות. חוסר ארגון זה הופך את ניצול הנתונים ההיסטוריים להצעות עתידיות לבלתי אפשרי כמעט, ומונע סטנדרטיזציה של תהליכים.<sup>1</sup>

שיטות ידניות מגבילות באופן מהותי את יכולתה של ארגון להרחיב את עיבוד הנתונים כדי להתמודד עם נפחי פרויקטים הולכים וגדלים. הדבר מוביל לירידה בביצועים ולאינצידנטים, כולל מגבלות בכוח אדם ובעלויות.<sup>2</sup> קיימים גם קשיים באינטגרציה, הנובעים מפורמטים שונים, מבנים או פרוטוקולים המשמשים מערכות או מקורות נתונים שונים. זה יכול להוביל לחוסר עקביות ושגיאות בנתונים אם לא מטופל ביעילות.<sup>2</sup>

יתרה מכך, מסמכי תכנון חשמל מכילים לעיתים קרובות פריסות מורכבות, טקסט מסובב, מידע מקוטע וסמלים מיוחדים. טקסט עשוי להיות מוסתר על ידי קווים או מסומן בהערות וסמלים שכלים גנריים של זיהוי תווים אופטי (OCR) אינם מאומנים להבין.<sup>3</sup> כל אלה הופכים את הפרשנות הידנית ואת השליפה האוטומטית לקשות באופן מהותי.<sup>2</sup>

העלות המצטברת של תהליכים ידניים חורגת בהרבה מהוצאות העבודה הישירות. בעוד שעלויות כוח אדם ידניות ברורות, העלויות העקיפות – הכנסות אבודות מהצעות מחיר שלא הוגשו, פגיעה בתחרותיות וקבלת החלטות לקויה עקב נתונים לא מדויקים או לא עקביים – הן לרוב גדולות בהרבה וקשות יותר לכימות. אתגרים אלה אינם מבודדים; הם מצטברים. נתונים לא מדויקים הנובעים משליפה ידנית<sup>2</sup> דורשים עבודה מחדש<sup>1</sup>, מה שמחמיר עוד יותר את מגבלות הזמן ומוביל להזדמנויות נוספות שהוחמצו. היעדר נתונים מרכזיים<sup>1</sup> מונע ניצול פרויקטים קודמים, ומחייב מאמץ ידני חוזר ונשנה. לפיכך, הצורך האסטרטגי באוטומציה חורג מעבר לרווחי יעילות בלבד; הוא נוגע לפתיחת פוטנציאל עסקי חדש, שיפור איכות ההחלטות וחזקת הזריזות הארגונית בשוק תחרותי.

## **1.2 החזון: הפעלת אנליטיקה מתקדמת ויעילות תפעולית באמצעות נתונים מובנים**

החזון הוא יצירת מערכת המאפשרת שליפה אוטומטית של נתונים מובנים, ובכך משפרת באופן דרמטי את הדיוק והשלמות של הנתונים. אוטומציה, בשילוב עם תהליכי אימות וניקוי נתונים, מבטיחה נתונים נקיים, עקביים ומדויקים המספקים תובנות אמינות.<sup>2</sup>

שליפת נתונים מהירה ומדויקת מאיצה את מחזורי הפרויקט, החל מהצעות מחיר וסקירות תכנון ועד לאומדני חומרים, ובכך מקצרת את זמני הביצוע הכוללים של הפרויקטים.<sup>1</sup> בסיס נתונים מובנה ישמש כמאגר מרכזי וניתן לחיפוש, שיאפשר גישה קלה, יכולת חיפוש ושימוש חוזר בנתוני פרויקטים היסטוריים. הדבר יאפשר סטנדרטיזציה של תהליכים וישפר את הצעות המחיר העתידיות.<sup>1</sup> יתר על כן, נתונים מובנים מהווים בסיס איתן לאנליטיקה מתוחכמת, מודלים חזויים ושילוב עם מערכות מידול מידע מבנים (BIM), תאומים דיגיטליים וכלי סימולציה. גישה בזמן לנתונים מדויקים ומקיפים תתמוך בקבלת החלטות תפעוליות ואסטרטגיות טובות יותר.<sup>2</sup>

## **2. הבנת פורמטי נתוני המקור: PDF ו-DWG**

סעיף זה יספק ניתוח טכני מפורט של שני פורמטי המקור העיקריים, PDF ו-DWG, והכלים הזמינים לניתוח ושליפת נתונים מהם.

### **2.1 צלילה עמוקה ל-DWG/DXF: מבנה, יכולות ואופי קנייני**

קבצי DWG ו-DXF הם פורמטים סטנדרטיים בתעשייה עבור תוכניות CAD, אך יש להם הבדלים מהותיים המשפיעים על תהליכי שליפת הנתונים.

#### **2.1.1 DWG לעומת DXF: הבדלים מהותיים ושימושים**

**DWG (Drawing)** הוא פורמט קובץ בינארי קנייני, המהווה את הפורמט המקורי של AutoCAD. הוא מותאם לאחסון קומפקטי של נתוני תכנון דו-ממדיים ותלת-ממדיים ומטא-נתונים.<sup>5</sup> פורמט זה שומר על תכונות מיוחדות ונתונים קנייניים, מה שהופך אותו לאידיאלי לאינטגרציה חלקה בתוך סביבת AutoCAD ולשמירת נתונים כאשר נאמנות לפרטים היא קריטית.<sup>6</sup>

**DXF (Drawing Exchange Format)**, לעומת זאת, הוא פורמט קובץ פתוח, מבוסס טקסט פשוט, שתוכנן במיוחד להחלפת שרטוטים בין יישומי CAD שונים.<sup>6</sup> המבנה הקריא-לאדם שלו מקל על הפרשנות ואף על שינוי ידני במידת הצורך, ומבטיח יכולת פעולה הדדית על פני פלטפורמות CAD מגוונות ומקל על שיתוף פעולה בין-פלטפורמלי.<sup>6</sup> עם זאת, המרה מ-DWG ל-DXF עלולה לגרום לאובדן של חלק מהתכונות המיוחדות והנתונים המקושרים אליהן.<sup>6</sup> חשוב לציין ששני הפורמטים, DWG ו-DXF, מאחסנים גרפיקה וקטורית (קווים, קשתות, עיגולים, טקסט, מצולעים) המבוססת על גיאומטריה ויחסים, מה שמבטיח מדרגיות ללא פיקסלציה.<sup>7</sup>

## Autodesk RealDWG SDK 2.1.2: יכולות טכניות, דרישות מערכת ומודלי רישוי

RealDWG הוא ערכת פיתוח תוכנה (SDK) המאפשרת למפתחי ++C ו-.NET לקרוא ולכתוב קבצי DWG™ ו-DXF™ של AutoCAD<sup>8</sup>. זוהי תת-קבוצה של ObjectARX SDK, המתמקדת אך ורק בחלקי בסיס הנתונים של השרטוט (ישויות ויחסים) ללא תמיכה בתצוגה או בממשק משתמש<sup>8</sup>. הוא אינו דורש התקנה של AutoCAD<sup>8</sup>.

דרישות המערכת של RealDWG כוללות Microsoft Windows 10/11, מעבד 64 סיביות במהירות +3 GHz, זיכרון RAM של 16 GB, שטח דיסק של 4 GB וסביבת פיתוח Microsoft Visual Studio (לדוגמה, VS 2022 עם .NET 8.0 עבור RealDWG 2026)<sup>8</sup>. הרישוי מטופל באופן גלובלי על ידי השותפה Tech Soft 3D<sup>8</sup>. המחיר הוא 8,000 דולר ארה"ב או 7,500 אירו לשנה עבור עד 10,000 משתמשי קצה, כאשר נפחים גדולים יותר דורשים יצירת קשר ישיר<sup>8</sup>. הוא פועל במודל הפצה ללא תמלוגים<sup>8</sup>. RealDWG מיועד בעיקר ליישומים "לא תחרותיים"<sup>5</sup>.

## Open Design Alliance (ODA) Drawings SDK 2.1.3: תכונות, תאימות ונוף רישוי

ODA Drawings SDK (שנודע בעבר בשם Teigha) תומך בפונקציונליות קריאה/כתיבה/הצגה עבור קבצי .dgn ו-.dwg<sup>11</sup>. הוא מציע API ברמה גבוהה למניפולציות נתונים, כולל אובייקטים אסוציאטיביים, שיבוט, טרנזקציות, ביטול/חזרה<sup>11</sup>. הוא תומך באובייקטים סטנדרטיים מורכבים כמו מימדים, טקסטים מרובי שורות, טבלאות, מוליכים מרובים ובלוקים דינמיים, ומאפשר יצירת אובייקטים מותאמים אישית<sup>11</sup>. הוא מספק המרה דו-כיוונית מקיפה בין DWG ו-DGN, כמו גם ייבוא/ייצוא של PDF, DWF ופורמטים אחרים<sup>11</sup>. תכונות מפתח כוללות תמיכה בשכבות, בלוקים, תכונות, אילוצים גיאומטריים ומימדיים, תיעוד מודל (תצוגות שרטוט חכמות) ו-TRACE (סימונים)<sup>11</sup>. הספריות כוללות API ב-++C ו- NET, Java, SWIG עבור Python<sup>11</sup>.

ODA היא קונסורציום ללא מטרות רווח בבעלות חברים<sup>12</sup>. התוכנה יכולה להיכלל רק בתוכנות יישומים בבעלות חברים נוכחיים, בכפוף להסכם חברות חתום<sup>14</sup>. העלויות משתנות בהתאם לצרכים הספציפיים, עם עלות שנתית ממוצעת של כ-38,000 דולר, ומחיר מרבי מדווח של כ-92,000 דולר<sup>15</sup>. רישיונות משתמש/מפתח מתחילים ב-249 דולר (ללא הרחבת PDF) או 399 דולר (עם הרחבת PDF), עם הנחות על נפח<sup>16</sup>. רישיונות זמן ריצה לשרת יקרים יותר, החל מ-999 דולר/1,499 דולר<sup>16</sup>. רישיונות אתר לשימוש בלתי מוגבל במיקום פיזי יחיד הם 2,999 דולר/3,999 דולר, ורישיונות הפצה בלתי מוגבלים הם 2,999 דולר/4,499 דולר<sup>16</sup>.

## 2.1.4 חלופות קוד פתוח (לדוגמה, ezdxf)

ezdxf היא חבילת פייתון לקבצי DXF. היא מציעה פונקציונליות דומה ל-dxfwrite ו-dxfgrabber אך מקיפה יותר<sup>17</sup>. יש לה טביעת רגל זיכרון קטנה יותר<sup>17</sup>. עם זאת, ezdxf אינה תומכת באובייקטי OLE, הנחשבים ל"החלטת עיצוב גרועה מאוד" בשרטוטי CAD עקב בעיות תאימות עתידיות<sup>17</sup>. היא מתמקדת בעיקר ב-DXF, ייתכן שלא תלכוד את כל תכונות ה-DWG הקנייניות.

## 2.1.5 קניין רוחני ושיקולים משפטיים עבור DWG/DXF

DWG הוא פורמט קנייני שפותח ומתוחזק על ידי Autodesk. Autodesk טוענת באופן עקבי לזכויות קניין רוחני על פורמט DWG ועל הסימנים המסחריים הקשורים אליו<sup>5</sup>. Autodesk נקטה בפעולות משפטיות אקטיביות (לדוגמה, נגד ODA ו-SolidWorks) כדי להגן על הסימנים המסחריים שלה "DWG"<sup>5</sup>. בעוד שמשרד הפטנטים והסימנים המסחריים של ארה"ב (USPTO) סירב בסופו של דבר



לרשום את "DWG" כסימן מסחרי עבור Autodesk, מתוך הכרה בו כמתאר פורמט קובץ<sup>5</sup>,  
Autodesk ממשיכה לטעון שהוא סימן מסחרי.<sup>19</sup>

Autodesk הציגה "סימני מים" בקבצי DWG כדי להבחין בין קבצי AutoCAD "אמיתיים" לבין אלה שנעשו בהנדסה לאחר על ידי מתחרים, מה שהוביל למחלוקות משפטיות נוספות.<sup>20</sup> בשנת 2010, Autodesk ו-ODA הגיעו להסכם פשרה, כאשר ODA הסכימה לבטל את רישומי הסימנים המסחריים מבוססי DWG שלה ולהפסיק את השימוש ב"DWG" ובסימנים מסחריים מבוססי DWG בשיווק ובמיתוג מוצריה, תוך שהיא עדיין רשאית לפתח תוכנה ניתנת להפעלה הדדית ולהשתמש בסימנת.dwg.<sup>19</sup> רישוי RealDWG SDK של Autodesk ניתן בהבנה מרומזת שהוא מיועד ליישומים "לא תחרותיים".<sup>5</sup>

הסתמכות בלעדית על מנתחי DWG בקוד פתוח או בהנדסה לאחר עבור פעולות עסקיות קריטיות, במיוחד ליצירת נתונים מובנים שעשויים לשמש ביישומים תחרותיים, טומנת בחובה סיכון משפטי משמעותי בשל טענות הקניין של Autodesk ופעולות האכיפה הקודמות שלה. אפילו בעת בחינת ODA, סקירה יסודית של הסכם החברות שלה וההשלכות של הסדר הפשרה הקודם שלה עם Autodesk היא קריטית כדי לוודא שתרחיש השימוש המיועד (לדוגמה, יצירת מסד נתונים מובנה עבור ניתוח פנימי לעומת מוצר מסחרי) אינו מפר תנאים כלשהם או גורר אתגרים משפטיים עתידיים. הבחירה בין RealDWG (אם ניתן לרישוי עבור תרחיש השימוש הספציפי), ODA או כלי קוד פתוח אינה רק טכנית או פיננסית; זוהי החלטה משפטית אסטרטגית שעשויה להשפיע על הכדאיות לטווח ארוך ועל עמדת הקניין הרוחני של הפתרון המפותח. הדבר מחייב סקירה של ייעוץ משפטי.

**טבלה 1: ניתוח השוואתי של ערכות פיתוח תוכנה (SDKs) לניתוח DWG/DXF**

|  |   |   |  |   |   |   |  |   |   |   |   |
|--|---|---|--|---|---|---|--|---|---|---|---|
| מג<br>בלו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>ו<br>ת                                      | י<br>ת<br>ר<br>ו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>י<br>ם                | ק<br>נ<br>י<br>ן<br>ר<br>ו<br>ח<br>נ<br>י<br>/<br>ס<br>י<br>כ<br>ו<br>ן<br>ש<br>פ<br>ט<br>י | ע<br>ל<br>ו<br>ת<br>ש<br>נ<br>ת<br>י<br>ת<br>מ<br>ש<br>ו<br>ע<br>ר<br>ת<br>(<br>מ<br>י<br>נ<br>י<br>מ<br>ו<br>-<br>מ<br>ק<br>ס<br>י<br>מ<br>ו<br>(<br>ם) | מ<br>ו<br>ד<br>ל<br>ר<br>י<br>ש<br>ו<br>י   | פ<br>ו<br>נ<br>ק<br>צ<br>י<br>ו<br>נ<br>ל<br>י<br>ו<br>ת<br>ת<br>צ<br>ו<br>ג<br>ה<br>ר<br>י<br>נ<br>ד<br>ו<br>ר | ת<br>מ<br>י<br>כ<br>ה<br>ב<br>א<br>ו<br>ב<br>י<br>ט<br>י<br>ם<br>מ<br>ו<br>ר<br>כ<br>ב<br>י<br>ם<br>(<br>ל<br>ד<br>ו<br>ג<br>מ<br>ה<br>ב<br>ל<br>ו<br>ק<br>י<br>ם<br>ד<br>י<br>נ<br>מ<br>י<br>ט<br>ב<br>ל<br>א<br>ו<br>ת<br>M<br>L<br>e<br>a<br>d<br>e<br>r<br>(<br>s | ק<br>ר<br>י<br>א<br>ה<br>/<br>כ<br>ת<br>י<br>ב<br>ה<br>D<br>W<br>G<br>/<br>D<br>X<br>F | מ<br>ע<br>ר<br>כ<br>ו<br>ת<br>ה<br>פ<br>ע<br>ל<br>ה<br>נ<br>ת<br>מ<br>כ<br>ו<br>ת | ש<br>פ<br>ת<br>נ<br>כ<br>ו<br>ת<br>/<br>A<br>P<br>I<br>ע<br>י<br>ק<br>ר<br>י<br>ת | מ<br>פ<br>ת<br>ח<br>/<br>ס<br>פ<br>ק  | S<br>D<br>/<br>K<br>י<br>כ<br>ל                               |
| W<br>i<br>n<br>d<br>o<br>w<br>s<br>ב<br>ל<br>ב<br>ד<br>ע<br>ל<br>ו<br>ת<br>ג<br>ב<br>ו | ת<br>א<br>י<br>מ<br>ו<br>ת<br>ר<br>ש<br>מ<br>י<br>ת<br>ש<br>ל<br>A<br>u | נ<br>מ<br>ו<br>ך<br>(<br>א<br>ם<br>ת<br>ו<br>א<br>ם<br>ל<br>ס<br>ע<br>י<br>ף                | \$<br>8<br>,<br>0<br>0<br>0<br>+<br>ד<br>ו<br>ל<br>ר<br>א<br>ר<br>ה<br>ב   | מ<br>נ<br>ו<br>י<br>ש<br>נ<br>ת<br>י<br>פ<br>צ<br>ה<br>ל<br>ל<br>א<br>ת<br>מ<br>ל<br>ו<br>ג<br>י<br>ם<br>ל<br>י<br>י<br>ש<br>ו<br>מ<br>י<br>ם | ל<br>א<br>(<br>ב<br>ס<br>י<br>ס<br>נ<br>ת<br>ו<br>נ<br>י<br>ם<br>ב<br>ל<br>ב<br>ד<br>(                          | כ<br>ן  | כ<br>ן   | W<br>i<br>n<br>d<br>o<br>w<br>s   | C<br>++<br>י<br>.NET  | A<br>u<br>t<br>o<br>d<br>e<br>s<br>k/<br>T<br>e<br>c<br>h<br>S<br>o<br>f<br>t<br>3<br>D | A<br>u<br>t<br>o<br>d<br>e<br>s<br>k<br>R<br>e<br>a<br>l<br>D |

|  |  |                    |  |  |                                |   |        |
|--|--|--------------------|--|--|--------------------------------|---|--------|
| ת<br>מי<br>כה<br>בא<br>ובי<br>יק<br>טי<br>ם<br>מו<br>רכ<br>בי<br>ם<br>(ל<br>דוג<br>,<br>בל<br>וקי<br>ם<br>דינ<br>מיי<br>ם,<br>טב<br>לא<br>ות,<br>M<br>Le<br>ad<br>er<br>(s | פונ<br>קצי<br>ונלי<br>ות<br>תצ<br>וגה<br>/רי<br>נדו<br>ר | מודל<br>רישוי      | על<br>ות<br>שנ<br>תי<br>ת<br>מ<br>שו<br>ער<br>ת<br>(מ<br>יני<br>מו<br>ם-<br>מק<br>סי<br>מו<br>(ם | קנ<br>רו<br>יין<br>חנ<br>/י<br>סי<br>כונ<br>ש<br>פ<br>טי | ית<br>רונ<br>עי<br>קר<br>יים   | מג<br>בלו<br>ת<br>עי<br>קר<br>יות                 | W<br>G |
|  |  | לא<br>תחרו<br>תיים |  | "ל<br>א<br>ת<br>רו<br>תי<br>("                           | to<br>de<br>sk<br>,<br>ק<br>חז | הה<br>,<br>סע<br>יף<br>ל<br>א<br>תח<br>רו<br>תי", |        |

|   |  |   |   |   |   |   |  |   |   |  |                                 |
|---|--|---|---|---|---|---|--|---|---|--|---------------------------------|
| מג<br>בלו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>ו<br>ת   | י<br>ת<br>ר<br>ו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>י<br>ם | ק<br>נ<br>י<br>ן<br>ר<br>ו<br>ח<br>נ<br>י<br>/<br>ס<br>י<br>כ<br>ו<br>ן<br>ש<br>פ<br>ט<br>י | ע<br>ל<br>ו<br>ת<br>ש<br>נ<br>ת<br>י<br>ת<br>מ<br>ש<br>ו<br>ע<br>ר<br>ת<br>(<br>מ<br>י<br>נ<br>י<br>מ<br>ו<br>-<br>מ<br>ק<br>ס<br>י<br>מ<br>ו<br>(<br>ם | מ<br>ו<br>ד<br>ל<br>ר<br>י<br>ש<br>ו<br>י                     | פ<br>ו<br>נ<br>ק<br>צ<br>י<br>ו<br>נ<br>ל<br>י<br>ו<br>ת<br>ת<br>צ<br>ו<br>ג<br>ה<br>ר<br>י<br>נ<br>ד<br>ו<br>ר | ת<br>מ<br>י<br>כ<br>ה<br>ב<br>א<br>ו<br>ב<br>י<br>י<br>ק<br>ט<br>י<br>ם<br>מ<br>ו<br>ר<br>כ<br>ב<br>י<br>ם<br>(<br>ל<br>ד<br>ו<br>ג<br>מ<br>ה<br>ב<br>ל<br>ו<br>ק<br>י<br>ם<br>ד<br>י<br>נ<br>מ<br>י<br>ט<br>ב<br>ט<br>ב<br>ל<br>א<br>ו<br>ת<br>M<br>L<br>e<br>a<br>d<br>e<br>r<br>(<br>s | ק<br>ר<br>י<br>א<br>ה<br>/<br>כ<br>ת<br>י<br>ב<br>ה<br>D<br>W<br>G<br>/<br>D<br>X<br>F | מ<br>ע<br>ר<br>כ<br>ו<br>ת<br>ה<br>פ<br>ע<br>ל<br>ה<br>נ<br>ת<br>מ<br>כ<br>ו<br>ת | ש<br>פ<br>ת<br>נ<br>כ<br>ו<br>ת<br>/<br>A<br>P<br>I<br>ע<br>י<br>ק<br>ר<br>י<br>ת | מ<br>פ<br>ת<br>ח<br>/<br>ס<br>פ<br>ק                               | S<br>D<br>/<br>K<br>י<br>כ<br>ל |
| ל<br>ל<br>א<br>ת<br>צ<br>ו<br>ג<br>ה                |  |   |   |   |   |   |  |   |   |  |                                 |
| ע<br>ל<br>ו<br>ת<br>ג<br>ב<br>ו<br>ה<br>ה<br>י<br>ם | ת<br>כ<br>ו<br>נ<br>ו<br>ת<br>מ<br>ק<br>י<br>פ<br>ו      | ב<br>י<br>נ<br>ו<br>נ<br>י<br>ד<br>ו<br>ר   | \$2<br>49<br>-<br>\$9<br>2,   | מ<br>ב<br>ו<br>ס<br>ח<br>ב<br>ר<br>ו<br>ת<br>(<br>ע<br>מ<br>ל | כ   | כ   | כ  | ח<br>ו<br>צ<br>ה<br>פ<br>ל<br>ט<br>פ<br>ו<br>ר                                    | C<br>++<br>י<br>.<br>N<br>E<br>T  | O<br>p<br>e<br>n<br>D<br>e<br>s<br>i<br>g<br>n<br>A<br>l<br>l<br>i | O<br>D<br>A<br>D<br>r<br>a      |

|                           |                         |                         |                                 |                                    |                        |  |                    |                           |                        |          |             |
|---------------------------|-------------------------|-------------------------|---------------------------------|------------------------------------|------------------------|--|--------------------|---------------------------|------------------------|----------|-------------|
| מגבלות עיקריות            | ייתרונות עיקריים        | קנה יין חנוסי / כונשפטי | עלות שנתית שווער (מנימי-מסימום) | מודל רישוי                         | פונקציות ונליטצור/נדור | תמיכה באובייקטים מרכיבים (לדוגמה, בלוקים דינמיים, טבלאות, M Leader(s | קריאה/כתבה DWG/DXF | מערכת הפעלה נמכות         | שפת תכנות/ת-API עיקרית | מפתח/ספק | S D /K יכל  |
| דרושת יחבור, תוכנית, רכבי | ת, רב-פלטפורמית, מיימכה | שסקירה קפדנית           | 00+0 דולר ארה"ב) בהת            | ות שנתיות, רמות שונות: משתמש/מפתח, |                        |  |                    | מות (W) indows , Linux, M | , Java , Python        | ance     | wings S D K |

|  |  |               |  |   |                              |                                   |                         |  |                  |  |   |  |  |               |  |  |   |   |  |                   |   |
|--|--|---------------|--|---|------------------------------|-----------------------------------|-------------------------|--|------------------|--|---|--|--|---------------|--|--|---|---|--|-------------------|---|
| ת<br>מי<br>כה<br>בא<br>ובי<br>יק<br>טי<br>ם<br>מו<br>רכ<br>בי<br>ם<br>(ל<br>דוג<br>,<br>בל<br>וקי<br>ם<br>דינ<br>מיי<br>ם,<br>טב<br>לא<br>ות,<br>M<br>Le<br>ad<br>er<br>(s | פונ<br>קצי<br>ונלי<br>ות<br>תצ<br>וגה<br>רי/<br>נדו<br>ר | מודל<br>רישוי | על<br>ות<br>שנ<br>תי<br>ת<br>מ<br>שו<br>ער<br>ת<br>(מ<br>יני<br>מו<br>ם-<br>מק<br>סי<br>מו<br>(ם | קנ<br>רו<br>חנ<br>י/<br>סי<br>כונ<br>ש<br>פ<br>טי | ית<br>רונ<br>עי<br>קר<br>יים | מג<br>בלו<br>ת<br>עי<br>קר<br>יות | S<br>D<br>/K<br>כל<br>י |  | מפ<br>תח/<br>ספק | ש<br>פ<br>ת<br>נכ<br>נו<br>ת/<br>A<br>P<br>I<br>עי<br>קר<br>ית | מע<br>רכו<br>ת<br>הפ<br>על<br>ה<br>נת<br>מכ<br>ות | קרי<br>אה/<br>כתי<br>בה<br>DW<br>G/<br>DX<br>F |  | ac<br>OS<br>( |  |  | שרת<br>,<br>אתר,<br>הפצ<br>ה<br>בלתי<br>מוגב<br>(לת | אם<br>לר<br>מ<br>ה/<br>מ<br>ש<br>א<br>ומ<br>תן<br>( | ל<br>ה<br>סכ<br>ם<br>ה<br>ח<br>בר<br>ות<br>עק<br>ב | ב-<br>D<br>G<br>N | מ<br>שפ<br>טיו<br>ת<br>פונ<br>טנ<br>צי<br>אל<br>יות |
|--|--|---------------|--|---|------------------------------|-----------------------------------|-------------------------|--|------------------|--|---|--|--|---------------|--|--|---|---|--|-------------------|---|

|                   |  |                  |  |  |   |   |  |               |   |  |                              |                                   |  |  |  |
|-------------------|--|------------------|--|--|---|---|--|---------------|---|--|------------------------------|-----------------------------------|--|--|--|
| S<br>D /K<br>כל י |  | מפ<br>תח/<br>ספק | ש<br>פ<br>ת<br>תכ<br>נו/ת<br>A<br>P<br>I<br>ע<br>י<br>ק<br>ר<br>י<br>ת | מע<br>רכו<br>ת<br>ה<br>פ<br>ע<br>ל<br>נ<br>ת<br>מ<br>כ<br>ו<br>ת | קרי<br>אה/<br>כת<br>בה<br>DW<br>G/<br>DX<br>F | ת<br>מי<br>כה<br>בא<br>ובי<br>יק<br>טי<br>ם<br>מו<br>רכ<br>בי<br>ם<br>(<br>ל)<br>דוג<br>מה<br>,<br>בל<br>וק'<br>ים<br>דינ<br>מ"<br>ם,<br>טב<br>לא<br>ות,<br>M<br>Le<br>ad<br>er<br>(s | פונ<br>קצי<br>ונלי<br>ות<br>תצ<br>וגה<br>/רי<br>נדו<br>ר | מודל<br>רישוי | על<br>ות<br>שנ<br>תי<br>ת<br>מ<br>שו<br>ער<br>ת<br>(מ<br>יני<br>מו-<br>ם-<br>מק<br>סי<br>מו<br>ם) | קנ<br>ין<br>רו<br>חנ<br>י/<br>סי<br>כון<br>ש<br>מ<br>ש<br>פי<br>טי | ית<br>רונ<br>עי<br>קר<br>יים | מג<br>בלו<br>ת<br>עי<br>קר<br>יות |  |  | מ<br>חל<br>וקו<br>ת<br>קי<br>ד<br>מו<br>(ת |
|-------------------|--|------------------|--|--|---|---|--|---------------|---|--|------------------------------|-----------------------------------|--|--|--|

|   |  |   |   |   |   |   |  |   |   |  |                                 |
|---|--|---|---|---|---|---|--|---|---|--|---------------------------------|
| מג<br>בלו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>ו<br>ת                       | י<br>ת<br>ר<br>ו<br>ת<br>ע<br>י<br>ק<br>ר<br>י<br>י<br>ם                     | ק<br>נ<br>י<br>ן<br>ר<br>ו<br>ח<br>נ<br>י<br>/<br>ס<br>י<br>כ<br>ו<br>ן<br>ש<br>פ<br>ט<br>י | ע<br>ל<br>ו<br>ת<br>ש<br>נ<br>ת<br>י<br>ת<br>מ<br>ש<br>ו<br>ע<br>ר<br>ת<br>(<br>מ<br>י<br>נ<br>י<br>מ<br>ו<br>-<br>מ<br>ק<br>ס<br>י<br>מ<br>ו<br>(<br>ם | מ<br>ו<br>ד<br>ל<br>ר<br>י<br>ש<br>ו<br>י   | פ<br>ו<br>נ<br>ק<br>צ<br>י<br>ו<br>נ<br>ל<br>י<br>ו<br>ת<br>ת<br>צ<br>ו<br>ג<br>ה<br>ר<br>י<br>נ<br>ד<br>ו<br>ר | ת<br>מ<br>י<br>כ<br>ה<br>ב<br>א<br>ו<br>ב<br>י<br>י<br>ק<br>ט<br>י<br>ם<br>מ<br>ו<br>ר<br>כ<br>ב<br>י<br>ם<br>(<br>ל<br>ד<br>ו<br>ג<br>מ<br>ה<br>ב<br>ל<br>ו<br>ק<br>י<br>ם<br>ד<br>י<br>נ<br>מ<br>י<br>ט<br>ב<br>ל<br>א<br>ו<br>ת<br>M<br>L<br>e<br>a<br>d<br>e<br>r<br>(<br>s | ק<br>ר<br>י<br>א<br>ה<br>/<br>כ<br>ת<br>י<br>ב<br>ה<br>D<br>W<br>G<br>/<br>D<br>X<br>F                     | מ<br>ע<br>ר<br>כ<br>ו<br>ת<br>ה<br>פ<br>ע<br>ל<br>ה<br>נ<br>ת<br>מ<br>כ<br>ו<br>ת | ש<br>פ<br>ת<br>נ<br>כ<br>ו<br>ת<br>/<br>A<br>P<br>I<br>ע<br>י<br>ק<br>ר<br>י<br>ת | מ<br>פ<br>ת<br>ח<br>/<br>ס<br>פ<br>ק   | S<br>D<br>/<br>K<br>י<br>כ<br>ל |
| ת<br>מ<br>י<br>כ<br>ה<br>מ<br>ו<br>ג<br>ב<br>ל<br>ו<br>ג<br>D<br>W<br>G | ח<br>י<br>נ<br>ם<br>/<br>מ<br>ב<br>ו<br>ס<br>פ<br>י<br>ת<br>ו<br>ן<br>ט<br>ו | ג<br>ב<br>ו<br>ה<br>(<br>ע<br>ב<br>ו<br>R<br>D<br>W<br>G<br>ע<br>ק<br>ב                     | ח<br>י<br>נ<br>ם  | ק<br>ו<br>ד<br>פ<br>ת<br>ו<br>ח<br>(<br>M<br>I<br>T<br>)<br>/<br>A<br>p<br>a<br>c<br>h<br>e<br>/<br>G<br>P<br>L<br>ב<br>ה<br>ת<br>א<br>ם<br>ל<br>ס<br>פ<br>ר<br>י<br>י<br>ה | ל<br>א  | מ<br>ו<br>ג<br>ב<br>ל<br>/<br>ל<br>א<br>ו<br>ע<br>ב<br>ו<br>ר<br>א<br>ו<br>ב<br>י<br>ק<br>ט   | D<br>X<br>F<br>/<br>ק<br>ר<br>י<br>א<br>ה<br>/<br>כ<br>ת<br>י<br>ב<br>ה<br>/<br>D<br>W<br>G<br>מ<br>ו<br>ג | ח<br>ו<br>צ<br>ה<br>פ<br>ל<br>ט<br>פ<br>ו<br>ר<br>מ<br>ו<br>ת                     | P<br>y<br>t<br>h<br>o<br>n  | ק<br>ו<br>ד<br>פ<br>ת<br>ו<br>ח<br>(<br>ק<br>ה<br>י<br>ל<br>ת<br>פ<br>י<br>י<br>ת<br>ו<br>ן) | e<br>z<br>d<br>x<br>f           |



|  |  |                   |  |   |                                 |  |                             |                  |  |   |  |  |  |  |  |  |
|--|--|-------------------|--|---|---------------------------------|--|-----------------------------|------------------|--|---|--|--|--|--|--|--|
| ת<br>מי<br>כה<br>בא<br>ובי<br>יק<br>טי<br>ם<br>מו<br>רכ<br>בי<br>ם<br>(ל<br>דוג<br>מה<br>,<br>בל<br>וקי<br>ם<br>דינ<br>מיי<br>ם,<br>טב<br>לא<br>ות,<br>M<br>Le<br>ad<br>er<br>(s | פונ<br>קצי<br>ונלי<br>ות<br>תצ<br>וגה<br>/רי<br>נדו<br>ר | מודל<br>רישוי     | על<br>ות<br>שנ<br>תי<br>ת<br>מ<br>שו<br>ער<br>ת<br>(מ<br>יני<br>מו<br>ם-<br>מק<br>סי<br>מו<br>(ם | קנ<br>רן<br>יין<br>חנ<br>/י<br>סי<br>כונ<br>ש<br>פ<br>פ<br>טי | ית<br>רונ<br>עי<br>קר<br>יים    | מג<br>בלו<br>ת<br>עי<br>קר<br>יות                      | S<br>D<br>/<br>K<br>כל<br>י | מפ<br>תח/<br>ספק | ש<br>פ<br>ת<br>תכ<br>נו<br>/ת<br>A<br>P<br>I<br>עי<br>קר<br>ית | מע<br>רכו<br>ת<br>הפ<br>על<br>ה<br>נת<br>מכ<br>ות | קרי<br>אה/<br>כתי<br>בה<br>DW<br>G/<br>DX<br>F |  |  |  |  |  |
| ים<br>מו<br>רכ<br>בי<br>ם<br>ספ<br>צי<br>פיי<br>ם<br>ל-  |  | הספ<br>ציפי<br>(ת |  | ה<br>ש<br>לכ<br>ות<br>הנ<br>דס<br>ה<br>לא<br>חו<br>(ר         | ב<br>ל-<br>D<br>XF<br>בס<br>יסי | לל<br>א<br>OL<br>,E<br>תמ<br>יכ<br>ה<br>קה<br>יל<br>תי |                             |                  |  |   | בל<br>(קרי<br>אה)                              |  |  |  |  |  |

|   |  |               |  |  |                              |                                   |  |  |  |  |  |
|---|--|---------------|--|--|------------------------------|-----------------------------------|--|--|--|--|--|
| ת<br>מ<br>י<br>כ<br>ה<br>ב<br>א<br>ו<br>ב<br>י<br>י<br>ק<br>ט<br>י<br>ם<br>מ<br>ו<br>ר<br>כ<br>ב<br>י<br>ם<br>(<br>ל<br>ד<br>ו<br>ג<br>מ<br>ה<br>,<br>ב<br>ל<br>ו<br>ק<br>י<br>ם<br>ד<br>י<br>נ<br>מ<br>י<br>ם<br>,<br>ט<br>ב<br>ל<br>א<br>ו<br>ת,<br>M<br>L<br>e<br>a<br>d<br>e<br>r<br>(<br>s | פונ<br>קצי<br>ונלי<br>ות<br>תצ<br>וגה<br>רי/<br>נדו<br>ר | מודל<br>רישוי | על<br>ות<br>שנ<br>תי<br>ת<br>מ<br>שו<br>ער<br>ת<br>(מ<br>יני<br>מו<br>ם-<br>מק<br>סי<br>מו<br>ם) | קנ<br>יין<br>רו<br>חנ<br>י/<br>סי<br>כונ<br>ש<br>פ<br>טי | ית<br>רונ<br>עי<br>קר<br>יים | מג<br>בלו<br>ת<br>עי<br>קר<br>יות |  |  |  |  |  |
| ת<br>בל<br>בד   |  |               |  |  |                              |                                   |  |  |  |  |  |

## 2.2 ניתוח מסמכי PDF: הבחנה בין תוכן וקטורי ורסטרי

מסמכי תכנון חשמל קיימים לעיתים קרובות כקבצי PDF, שיכולים להיות מבוססי וקטור (שנוצרו מ-CAD) או מבוססי רסטר (תמונות סרוקות). גישת השליפה שונה באופן משמעותי בהתאם להבחנה זו.

### 2.2.1 שליפת PDF וקטורי: כלים וטכניקות לנתונים גיאומטריים וטקסטואליים

קבצי PDF וקטוריים מכילים טקסט מוטבע ופרימיטיבים גיאומטריים (קווים, צורות) הנגישים ישירות. כלים לשליפה כוללים:

- **Adobe Acrobat:** מאפשר ייצוא תמונות ובחירות טקסט מקבצי PDF לפורמטים שונים.<sup>21</sup> הוא יכול לייצא תמונות רסטר אך לא אובייקטים וקטוריים ישירות.<sup>21</sup> הוא מציע גם כלים לעריכת טקסט, תמונות ומבנה מסמך.<sup>21</sup>
- **PyMuPDF (Fitz):** ספריית פייתון בעלת ביצועים גבוהים לשליפת נתונים, ניתוח, המרה ומניפולציה של מסמכים כמו קבצי PDF.<sup>22</sup> היא מהירה באופן משמעותי מחלופות כמו PyPDF2 ו-PDFMiner לשליפת טקסט.<sup>22</sup> PyMuPDF Pro מציעה הרחבות מסחריות לתמיכה במסמכי Office ושילובים של RAG/LLM.<sup>25</sup>
- **PDFMiner (pdfminer.six):** חבילת פייתון המתמקדת בשליפת טקסט, מטא-נתונים ושמירת פריסה מקבצי PDF.<sup>27</sup> היא מצטיינת בשליפת טקסט אך פחות יעילה בטיפול ברכיבים שאינם טקסטואליים ומתקשה בפריסות מורכבות במיוחד (מרבובות עמודות, טבלאות מקוננות).<sup>27</sup> ניתן לשלבה עם כלי OCR למסמכים סרוקים.<sup>27</sup>
- **pdfplumber:** ספריית פייתון נוספת לשליפת PDF, המושווית לעיתים קרובות ל-PyMuPDF ו-PDFMiner.<sup>24</sup>

עבור נפחים גדולים של תוכניות חשמל בפורמט PDF, במיוחד אלה עם פריסות מורכבות שבהן מהירות היא קריטית, PyMuPDF היא הבחירה העדיפה לשליפת טקסט ראשונית ומבנית בסיסית בשל יתרון הביצועים שלה. בעוד ש-PyMuPDF מטפל בטקסט היטב, הצורך לשלוף מידע גיאומטרי ורכיבים לא טקסטואליים מורכבים (סמלים, קווים) מקבצי PDF וקטוריים עשוי לדרוש עיבוד נוסף או תכונות מיוחדות שאינן מודגשות במקור, מה שמצביע על גישה מרובת כלים או פיתוח מותאם אישית על בסיס היכולות הבסיסיות של PyMuPDF. מתן עדיפות לביצועים בשלב שליפת ה-PDF הוא קריטי ליעילות ולמדרגיות הכוללת של צינור הנתונים, ומשפיע ישירות על זמן העיבוד ועלויות המחשוב, במיוחד אם מדובר באלפי מסמכים.

## 2.2.2 עיבוד PDF רסטר/סרוק: הצורך ב-OCR מתקדם וראייה ממוחשבת

קבצי PDF רסטר הם למעשה תמונות של מסמכים. טקסט וגרפיקה אינם נגישים ישירות כנתונים מובנים; הם פיקסלים. גישה זו דורשת זיהוי תווים אופטי (OCR) להמרת תמונות טקסט לטקסט קריא למכונה, וטכניקות ראייה ממוחשבת מתקדמות (זיהוי אובייקטים, סגמנטציה) לזיהוי וסיווג סמלים וקווים. כלי OCR סטנדרטיים (כמו Tesseract או EasyOCR) בדרך כלל אינם מספיקים לשרטוטי הנדסה בשל טקסט מסובב, קיטוע, טקסט מוסתר על ידי קווים, ונוכחות של סמלים והערות מיוחדים שהם אינם מאומנים להבין.<sup>3</sup>

## 2.2.3 רישוי מסחרי של PyMuPDF לשימוש מקצועי

PyMuPDF Pro היא הרחבה מסחרית של PyMuPDF, המציעה תכונות משופרות כמו טיפול במסמכי Office (DOC/X, PPT/X, XLS/X) ושילובים של RAG/LLM.<sup>25</sup> היא תומכת בשליפת טקסט וטבלאות ובהמרת מסמכים.<sup>25</sup> הרישוי דורש רישיון מסחרי לפונקציונליות מלאה מעבר למגבלת 3 עמודים או מפתח ניסיון לזמן מוגבל.<sup>25</sup> התנאים אינם מפורטים במפורש בתקצירים אך דורשים יצירת קשר עם מחלקת המכירות של Artifex Software Inc. להתאמה אישית, הפצה ומדרגיות ללא הגבלות.<sup>25</sup>

עבור פתרון שלילת נתונים מקצועי ורחב היקף, הסתמכות בלעדית על גרסת קוד פתוח של PyMuPDF עשויה להיות בלתי מספקת אם נדרשות תכונות מתקדמות (לדוגמה, עיבוד ישיר של מסמכי Office שאינם PDF שעשויים להכיל דיאגרמות מוטבעות, או ניצול שילובי RAG/LLM לשליפה קונטקסטואלית עשירה יותר). הצורך ליצור קשר עם Artifex לקבלת תנאים מסחריים<sup>25</sup> מרמז על תהליך משא ומתן, והעלות עשויה להיות משמעותית, בדומה ל-SDKs ברמה ארגונית אחרים. זה מוסיף לתקציב הפרויקט הכולל ולמורכבות הרכש. ארגונים חייבים להעריך האם היכולות הנוספות של PyMuPDF Pro (או SDKs מסחריים דומים ל-PDF) מצדיקות את עלות הרישוי והמורכבות, במיוחד אם מסמכי המקור שלהם חורגים מ-PDF טהורים או אם הם מתכננים לשלב עם מודלי AI מתקדמים לשליפת נתונים עשירה יותר.

## טבלה 2: השוואת ביצועים של ספריות שלילת טקסט מ-PDF

| ספרייה       | פונקציה עיקרית             | ביצועים (מהירות יחסית/זמן עבור N עמודים)                                  | יתרונות עיקריים                             | מגבלות עיקריות                              |
|--------------|----------------------------|---|---|---|
| PyMuPDF      | שלילת טקסט, מניפולציית PDF | המהירה ביותר (לדוגמה, 3.05 שניות עבור 7031 עמודים) <sup>22</sup>          | מהירות גבוהה, תכונות חזקות, גרסת Pro מסחרית | רישיון מסחרי עבור תכונות Pro                |
| PyPDF2       | מניפולציית PDF             | איטית באופן משמעותי (לדוגמה, 494.04 שניות עבור 7031 עמודים) <sup>22</sup> | מניפולציית PDF בסיסית                       | איטית לשליפת טקסט                           |
| PDFMiner.six | שלילת טקסט, שימור פריסה    | איטית יותר (לדוגמה, 227.27 שניות עבור 7031 עמודים) <sup>22</sup>          | שימור פריסה, טובה לטקסט                     | מתקשה בפריסות מורכבות, רכיבים לא טקסטואליים |
| PDFrw        | מניפולציית PDF             | איטית יותר (לדוגמה, 10.54 שניות עבור                                      | מניפולציית PDF                              | איטית לשליפת טקסט                           |

|         |                |   |                 |                      |
|---------|----------------|---|-----------------|----------------------|
| ספרייה  | פונקציה עיקרית | ביצועים (מהירות יחסית/זמן עבור N עמודים)                        | יתרונות עיקריים | מגבלות עיקריות       |
|         |                | 7031 עמודים) <sup>22</sup>                                      |                 |                      |
| PikePDF | מניפולציית PDF | איטית יותר (לדוגמה, 33.57 שניות עבור 7031 עמודים) <sup>22</sup> | מניפולציית PDF  | איטית לשליפת טקסט    |
| XPDF    | שליפת טקסט     | איטית יותר (לדוגמה, 27.42 שניות עבור 7031 עמודים) <sup>22</sup> | שליפת טקסט      | איטית יותר מ-PyMuPDF |

### 3. טכניקות מתקדמות לפרשנות שרטוטי חשמל

סעיף זה מתעמק במתודולוגיות הליבה של AI/ML הנדרשות לפרשנות המידע החזותי בשרטוטי חשמל, במיוחד מקבצי PDF רסטריים או ייצוא CAD לא סמנטי.

#### 3.1 זיהוי אובייקטים וזיהוי סמלי חשמל

##### 3.1.1 מודלים חדישים: YOLOv8, Detectron2 וארכיטקטורות מיוחדות

זיהוי אובייקטים מודרני מנצל למידה עמוקה ורשתות קונבולוציוניות (CNNs) כדי לחלץ תכונות מורכבות, ובכך עולה על שיטות קונבנציונליות.<sup>28</sup>

**Detectron2** היא ספריית מחקר של Facebook AI Research לזיהוי אובייקטים, סגמנטציה וזיהוי חזותי.<sup>29</sup> היא תומכת ביכולות מתקדמות כמו סגמנטציה פנופטית, תיבות תוחמות מסובבות ו-Cascade R-CNN.<sup>29</sup> היא משמשת כספרייה לפרויקטים מחקריים ויישומי ייצור.<sup>29</sup>

**YOLOv8** היא משפחה של גלאי יריעה אחת הידועים במהירותם ובדיוקם, ששוחררה בשנת 2023.<sup>28</sup> YOLOv8m (בינוני) השיג mAP50 של 96.7%, בעוד ש-YOLOv8s (קטן) השיג mAP50 של 96.5% עם זמן הסקה קצר משמעותית (2.3 מ"ש לעומת 5.7 מ"ש), מה שהופך אותו לטוב יותר עבור יישומים בזמן אמת.<sup>28</sup> הביצועים מוערכים באמצעות מדדים כמו mAP50, mAP50-90, דיוק, היזכרות וערכי AP ספציפיים עבור גדלי אובייקטים שונים (APs, APm, API).<sup>28</sup>

##### 3.1.2 אתגרים בזיהוי סמלים: טקסט מסובב, אלמנטים חופפים, רעש וסמלים מותאמים אישית

מסמכי הנדסה, בין אם מדובר בתרשימי ID&P, שרטוטי ייצור או גיליונות CAD סרוקים, שונים מבחינה מבנית ממסמכי משרד סטנדרטיים.<sup>3</sup> טקסט בשרטוטים אלה מסובב לעיתים קרובות, מקוטע, מוסתר על ידי קווים, או מסומן בסמלים ובהערות שכלים גנריים של OCR לא אומנו להבין.<sup>3</sup> איכות הזיהוי יורדת באופן משמעותי בסריקות ישנות ורועשות, הנפוצות בפרויקטים הנדסיים.<sup>3</sup> אלמנטים קריטיים כמו GD&T (Geometric Dimensioning and Tolerancing) או סמלי ריתוך מוחמצים לעיתים קרובות אלא אם כן המודלים מכילים במיוחד.<sup>3</sup> היעדר טיפול מובנה במבנה או טבלאות, ובעיות בשרטוטים גדולים הדורשים חלוקה לאריחים, מציבים אתגרים נוספים.<sup>3</sup>

### 3.1.3 ניצול מערכי נתונים ספציפיים לתחום: דוגמת Digitize-HCD ואצירת נתונים מותאמת אישית

אימון מודלים חזקים לזיהוי אובייקטים דורש מערכי נתונים גדולים ומסומנים הספציפיים לסמלי תרשימי חשמל. מערך הנתונים Digitize-HCD פותח במיוחד לצורך דיגיטציה של תרשימי מעגלים בכתב יד.<sup>30</sup> הוא מכיל 1,277 תמונות של תרשימים בכתב יד שצוירו על ידי למעלה מ-150 מתנדבים, ונסרקו ברזולוציה של 600 dpi.<sup>30</sup>

ההערות במערך נתונים זה מפורטות וכוללות 17 סמלי רכיבים מובחנים (תיבות תוחמות מיושרות ציר בפורמט COCO JSON), תוויות טקסט (הערות פולגון עם תמלול ידני בפורמט MMOCR OCRDataset), ומיקומי יציאות רכיבים (תמונות מפת חום של אמת קרקע).<sup>30</sup> אם מערכי נתונים ציבוריים אינם מספיקים, יצירת מערך נתונים מותאם אישית על ידי שרטוט וסימון אלפי סמלים ושימוש בכלים כמו Pillow ו-OpenCV לעיבוד היא גישה אפשרית, אם כי דורשת עבודה רבה.<sup>31</sup>

ההצלחה של זיהוי אובייקטים עבור סמלי חשמל תלויה לחלוטין בזמינות ובאיכות של נתוני אימון מסומנים ספציפיים לתחום. מערכי נתונים גנריים של תמונות אינם מספקים. בעוד שמערכי נתונים כמו Digitize-HCD קיימים, ייתכן שהם לא יכסו את כל הסמלים הספציפיים, הווריאציות או סגנונות השרטוט הקיימים במסמכים קנייניים של ארגון (לדוגמה, מודפס לעומת כתב יד, סטנדרטים ספציפיים של חברה). זה מחייב הגדלת מערכי נתונים קיימים או יצירת חדשים לחלוטין. תהליך אצירת הנתונים המותאמת אישית והסימון הוא משימה משמעותית מבחינת זמן, עלות ומומחיות, אך זהו צעד בלתי ניתן לווייתור להשגת דיוק גבוה בהקשר הנדסי אמיתי. ארגונים חייבים לתקצב לא רק משאבי מחשוב ופיתוח מודלים, אלא גם באופן משמעותי לרכישת נתונים, סימון ותחזוקה שוטפת של מערכי הנתונים הספציפיים לתחום שלהם כדי להבטיח שמודלי ה-AI יישארו יעילים וניתנים להתאמה לסוגי מסמכים מתפתחים.

## 3.2 שליפת טקסט והערות עם הבנה קונטקסטואלית

### 3.2.1 אופטימיזציה של OCR למסמכי הנדסה: מעבר לשליפת טקסט גנרית

PaddleOCR היא ערכת כלים מבוססת למידה עמוקה לזיהוי, זיהוי וניתוח מבנה מסמכים, עם תמיכה רב-לשונית.<sup>3</sup> מודול StructureV3 שלה יכול לזהות ולנתח טבלאות ובלוקי כותרת, ו-PP-OCRv5 מטפל בריבוי שפות וכתב יד.<sup>3</sup> היא מציעה את הניתוח המובנה ביותר מבין כלי הקוד הפתוח אך דורשת מאמץ התקנה משמעותי.<sup>3</sup> היא תומכת בספריית פייתון מקומית, שירות ענן ושירות עצמי.<sup>32</sup>

DocTR (Document Text Recognition) היא ספריית OCR בקוד פתוח שנבנתה בפייתון תוך שימוש בלמידה עמוקה (TensorFlow/PyTorch).<sup>3</sup> היא עובדת היטב עם טקסט מסובב וקלה לשילוב.<sup>3</sup> כלי OCR גנריים (Tesseract, EasyOCR) מתאימים לשליפת פסקאות מטפסים אך אינם

מספיקים לדיגיטציה של שרטוטי ייצור בשל חוסר יכולתם להבין פריסה, לטפל בסמלים או לכבד את מבנה מסמכי ההנדסה.<sup>3</sup>

### 3.2.2 טיפול בכתב יד, T&GD ורשימות חומרים (BOM)

PP-OCRv5 ב-PaddleOCR מציג שיפור בזיהוי כתב יד.<sup>32</sup> עם זאת, PaddleOCR עדיין מפספס סמלי T&GD או ריתוך רבים אלא אם כן הוא מכיל במיוחד.<sup>3</sup> DocTR חסר תמיכה מובנית בסמלים ונכשל בטבלאות מוטבעות.<sup>3</sup> זה מדגיש את הצורך באימון מותאם אישית ובגישות היברידיות. DocTR אינו מספיק אם לשרטוט יש טבלאות BOM.<sup>3</sup> מודול StructureV3 של PaddleOCR מבטיח לזיהוי וניתוח טבלאות.<sup>3</sup>

שום פתרון OCR "מהקופסה" אינו מספיק לשליפה מקיפה של טקסט והערות משרטוטי הנדסת חשמל מורכבים. פריסה "הפעל-ושכח" אינה אפשרית.<sup>3</sup> השגת דיוק גבוה עבור הערות הנדסיות קריטיות כמו T&GD, סמלי ריתוך, או אפילו פורמטי BOM ספציפיים, תדרוש כיוון עדין משמעותי של מודלי OCR ללמידה עמוקה שנבחרו (לדוגמה, PaddleOCR) עם מערכי נתונים מסומנים מותאמים אישית. זה מרמז על מחזור חיים מתמשך של למידת מכונה, ולא פריסה חד פעמית. גישה היברידית המשלבת OCR מתקדם עבור טקסט וטבלאות כלליים עם מודלי זיהוי אובייקטים מיוחדים (מסעיף 3.1) עבור סמלים, ואחריה ניתוח מבוסס כללים או סמנטי להבנה קונטקסטואלית, תהיה ככל הנראה הכרחית ללכידת כל המידע הרלוונטי בדיוק. הפרויקט חייב להקצות משאבים משמעותיים לאימון מתמשך של מודלי AI, אימות, ואולי תהליך שליפה רב-שלבי כדי להתמודד עם האתגרים הניוואנסים של פרשנות מסמכי הנדסה.

### טבלה 3: השוואת כלי OCR לשרטוטי הנדסה

| כלי       | סוג                               | יתרונות לשרטוטי הנדסה  | מגבלות לשרטוטי הנדסה  | שורה תחתונה/התאמה   |
|-----------|-----------------------------------|--|---|---|
| PaddleOCR | ערכת כלים ללמידה עמוקה (קוד פתוח) | ניתוח מובנה (טבלאות, בלוקי כותרת), רב-לשוני, כתב יד (PP-OCRv5), מזהה פריסה | מפספס סמלי T&GD/ריתוך ללא כיוון עדין, מאמץ התקנה משמעותי                | הניתוח המובנה ביותר מבין כלי הקוד הפתוח, טוב לצוותים טכניים המוכנים לכיוון עדין |
| DocTR     | ספריית למידה עמוקה (קוד פתוח)     | מטפל בטקסט מסובב, אינטגרציה קלה  | מיועד למסמכים מבוססי טקסט (חשבוניות), נכשל בחתכים/טבלאות מוטבעות/סמלים, | לא מתאים לשרטוטים עם BOM או סמלים   |

| כלי                     | סוג                                    | יתרונות<br>לשרטוטי<br>הנדסה                                  | מגבלות לשרטוטי<br>הנדסה  | שורה<br>תחתונה/התאמה  |
|-------------------------|--|--|--|---|
|                         |  |  | ללא תמיכה מובנית<br>בסמלים   |   |
| Tesseract<br>(גנרי OCR) | מנוע<br>OCR<br>מסורתי<br>(קוד<br>פתוח) | טוב לטקסט<br>פשוט, זמין<br>באופן נרחב                        | נכשל בפריסות<br>מורכבות, טקסט<br>מסובב, סמלים,<br>סריקות רועשות                          | לא מתאים<br>לשליפת שרטוטים<br>מלאים                           |
| eDOCr<br>(מסחרי/מיוחד)  | פתרון<br>AI<br>מסחרי                   | טוב לשרטוטי<br>מכניקה<br>סטנדרטיים,<br>זרימה חצי<br>אוטומטית | עתיר משאבים,<br>פחות ניתן<br>להתאמה אישית,<br>ביצועים יורדים<br>בפריסות מותאמות<br>אישית | אופציה בינונית<br>מוצקה לשליפה<br>אמינה משרטוטים<br>סטנדרטיים |

### 3.3 שחזור טופולוגיית תרשים חשמל

מעבר לשליפת רכיבים וטקסט בודדים, השלב הקריטי הוא הבנת אופן חיבורם של אלמנטים אלה ליצירת מעגל חשמלי פונקציונלי. הדבר דורש המרת נתונים חזותיים לייצוג מבוסס גרפים.

#### 3.3.1 יישום תורת הגרפים לניתוח קישוריות

תורת הגרפים היא ענף במתמטיקה המשמש באופן נרחב בניתוח רשתות לייצוג קישוריות.<sup>33</sup> בטופולוגיית מעגלים, צמתי הרשת הם קודקודי תורת הגרפים, וענפי הרשת (חיבורים) הם קצוות הגרף.<sup>33</sup> תורת הגרפים שימשה בניתוח רשתות חשמליות מאז חוקי קירכהוף (1847) ועבודתו של מקסוול (1873).<sup>33</sup> מושגים כמו מטריצות שכיחות ועצי פורש ישימים ישירות.<sup>33</sup> מורכבות מעגלים גדלה מחייבת קומבינטוריקה בתורת הגרפים לחישוב יעיל במחשב.<sup>33</sup>

#### 3.3.2 זיהוי צמתים (רכיבים, צמתים) וקצוות (חיבורים, חוטים)

סמלי החשמל שחולצו (נגדים, קבלים, מתגים, מנועים וכו') ונקודות החיבור (מסופים, צמתים, נקודות נייטרליות) שזוהו בשלבים הקודמים הופכים לקודקודים של הגרף.<sup>35</sup> חוטים, כבלים ונתיבי תמסורת המחברים רכיבים אלה יוצרים את הקצוות.<sup>35</sup> סוג המוליך (לדוגמה, תלת-פאזי, מסוכך) יכול להיות תכונה של הקצה.<sup>35</sup> התהליך כרוך בזיהוי קרבה מרחבית ונקודות חיתוך בין מסופי רכיבים וקווים כדי להסיק קישוריות. זוהי בעיה גיאומטרית מורכבת.

#### 3.3.3 כלים לייצוג וניתוח גרפים (לדוגמה, NetworkX, Schemdraw)

**NetworkX** היא חבילת פייתון ליצירה, מניפולציה ולימוד המבנה, הדינמיקה והפונקציות של רשתות מורכבות.<sup>37</sup> היא מאפשרת הוספת צמתים וקצוות, בחינת תכונות גרפים (צמתים, קצוות, סמיכות,



דרגה), ותומכת בפונקציות תורת הגרפים שונות (רכיבים מחוברים, אשכולות, מסלול קצר ביותר).<sup>37</sup> צמתים וקצוות יכולים להיות כל אובייקט ניתן לגיבוב, מה שמאפשר אחסון תכונות עשיר.<sup>37</sup>

**Schemdraw** היא חבילת פייתון להפקת דיאגרמות סכמטיות של מעגלים חשמליים באיכות גבוהה.<sup>38</sup> בעוד שעיקר ייעודה הוא שרטוט, הגישה התכנותית שלה להוספת אלמנטים וחיבורם (לדוגמה,

`Resistor().right`, `Capacitor().down`) משקפת את המבנה הטופולוגי הבסיסי שיש לשחזר.

ספריות גיאומטריות כמו `D3.js` (JavaScript) מציעות יכולות לרינדור קשתות, קווים וקישורים, ופריסות עבור גרפים מונעי כוח, שימושיות להצגת טופולוגיה משוחזרת.<sup>39</sup> ספריית הסמלים של Nemeth<sup>40</sup> מספקת מסגרת קונספטואלית לייצוג סמלים מתמטיים, שניתן להרחיבה לסמלי חשמל וחיבוריהם.

שלב זיהוי האובייקטים (3.1) מזהה אילו סמלים וטקסט קיימים. שלב שחזור הטופולוגיה עוסק בהבנה כיצד ישויות מזהות אלה מחוברות פונקציונלית. זהו קפיצה סמנטית מפיקסלים/וקטורים לתרשים מעגל לוגי. האתגר טמון בזיהוי מדויק של נקודות חיבור (יציאות) על סמלים ובמעקב אחר קווים/חוטים ביניהם, במיוחד בשרטוטים רועשים או מורכבים. זה דורש ניתוח גיאומטרי מדויק ואולי "אלגוריתם פתרון מבוך"<sup>31</sup> כדי לעקוב אחר חיבורים בין סמלים מזהים. התוצאה של שלב זה היא גרף שבו צמתים הם רכיבים/צמתים וקצוות הם חיבורים, עם תכונות הנגזרות מטקסט שחולץ (לדוגמה, ערכי רכיבים, סוגי חוטים). גרף זה הופך אז לנתונים המובנים עבור בסיס הנתונים. זהו השלב המורכב והקריטי ביותר, הממיר נתונים חזותיים גולמיים לייצוג מעגל חשמלי בעל משמעות סמנטית. הדיוק שלו קובע את התועלת של המערכת כולה לניתוח וסימולציה במורד הזרם.

#### טבלה 4: סמלי חשמל עיקריים ומיפויים לתקן IEC 60617

| שם סמל/תיאור | ייצוג חזותי (דוגמה)                 | קוד תקן IEC 60617 (אם רלוונטי) | תכונות נפוצות            | הערות/הוראות שימוש |
|--------------|-------------------------------------|--------------------------------|--------------------------|--------------------|
| נגד          | (מלבן עם קווים יוצאים)              | 60617-DB 4:2012                | ערך התנגדות, הספק        | רכיב פסיבי בסיסי   |
| קבל          | (שני קווים מקבילים עם קווים יוצאים) | 60617-DB 4:2012                | ערך קיבול, מתח           | רכיב פסיבי בסיסי   |
| מתג (כללי)   | (מעגל עם קו אלכסוני)                | 60617-DB 7:2012                | סוג מגע (פתוח/סגור), מצב | <sup>35</sup>      |

| שם<br>סמל/תיאור                | ייצוג חזותי<br>(דוגמה)                   | קוד תקן<br>IEC 60617<br>אם<br>(רלוונטי) | תכונות נפוצות                                       | הערות/הוראות<br>שימוש |
|--------------------------------|--|---|---|-----------------------|
| הארקה                          | (שלושה קווים<br>יורדים בגדלים<br>שונים)  | 60617-<br>DB 1:2012                     | סוג הארקה<br>(הארקה כללית,<br>הארקה<br>פונקציונלית) | 35                    |
| מסוף                           | (עיגול קטן)                              | 60617-<br>DB 1:2012                     | מספר מסוף, ייעוד                                    | 35                    |
| מנוע חשמלי                     | (מעגל עם M<br>בפנים)                     | 60617-<br>DB 6:2012                     | סוג מנוע, הספק,<br>מתח                              | 35                    |
| מגבר                           | (משולש)                                  | 60617-<br>DB 5:2012                     | הגבר, סוג   | 35                    |
| מגע סגור<br>(Make<br>contact)  | (שני קווים עם<br>רווח, קו שלישי<br>מחבר) | 60617-<br>DB 7:2012                     | מצב (פתוח/סגור)                                     | 35                    |
| מגע פתוח<br>(Break<br>contact) | (שני קווים עם<br>רווח, קו שלישי<br>חוצה) | 60617-<br>DB 7:2012                     | מצב (פתוח/סגור)                                     | 35                    |
| נקודת נייטרל                   | (שלושה קווים<br>מתחברים<br>לנקודה אחת)   | 60617-<br>DB 1:2012                     | ייעוד (לדוגמה,<br>גנרטור)                           | 35                    |
| חיבור (כללי)                   | (קווים<br>מצטלבים)                       | 60617-<br>DB 1:2012                     | סוג מוליך (כבל<br>חשמל, 3 פאזות)                    | 35                    |
| נתיך                           | (מלבן עם קו<br>גלי בפנים)                | 60617-<br>DB 8:2012                     | זרם נקוב, מתח                                       | התקן הגנה             |

|                 |                         |   |                |                       |
|-----------------|-------------------------|---|----------------|-----------------------|
| שם<br>סמל/תיאור | ייצוג חזותי<br>(דוגמה)  | קוד תקן<br>IEC 60617<br>(אם<br>רלוונטי) | תכונות נפוצות  | הערות/הוראות<br>שימוש |
| נורה            | (עיגול עם צלב<br>בפנים) | 60617-<br>DB 8:2012                     | הספק, מתח, סוג | התקן<br>איתות/תאורה   |

#### 4. ארגון נתונים שחולצו לאינטגרציה עם בסיס נתונים

סעיף זה מתמקד בהגדרת הסכימה ובהבטחת איכות ועקביות הנתונים שחולצו לתכנון חשמלי לצורך אחסון בבסיס נתונים.

#### 4.1 תכנון מודל נתונים מקיף למידע תכנון חשמלי

##### 4.1.1 מיפוי נתונים שחולצו לתקנים בתעשייה (לדוגמה, IEC 60617, IFC Electrical Domain Schema)

**IEC 60617** הוא תקן בינלאומי המגדיר סמלים גרפיים לדיאגרמות אלקטרוטכניות, ומכיל כ-1900 סמלים.<sup>41</sup> הוא מכסה מוליכים, רכיבים פסיביים, מוליכים למחצה, ייצור והמרת אנרגיה חשמלית, ציוד מיתוג, בקורות והתקני הגנה, מכשירי מדידה, מנורות והתקני איתות, וכן תוכניות התקנה אדריכליות וטופוגרפיות.<sup>41</sup> הסמלים זמינים בפורמטים GIF, DWG ו-EPS, עם מטא-נתונים כמו שם, שימוש, מילות מפתח וקישורים.<sup>41</sup> מנוי לשנה עולה 690 פרנקים שווייצריים, עם שנים עוקבות ב-270 פרנקים שווייצריים.<sup>41</sup>

**IFC Electrical Domain Schema (IFC 4.3.2)** הוא חלק משכבת הדומיין של מודל IFC, ומרחיב את הרעיונות הנוגעים לשירותי בניין.<sup>42</sup> הוא מגדיר מושגים של מערכות כבלים עבור חשמל, תאורה, נתונים, תקשורת, אבטחה ובקרה.<sup>42</sup> הוא מכסה מכשירי חשמל, מנועים, נושאי כבלים (צינורות, מגשי כבלים), לוחות חלוקה, גנרטורים, קופסאות חיבורים, גופי תאורה, שקעים, התקני הגנה ושנאים.<sup>42</sup>

היקף ומגבלות של IFC Electrical Domain: הוא מיועד בעיקר להתקנות חשמל במתח נמוך (V 12 AC/DC עד V AC1000 או V DC1500).<sup>42</sup> מתח בינוני/גבוה, מערכות מתחת ל-V12, אספקת חשמל ציבורית ומצבים חולפים אינם בתחום ההיקף או שאינם מפורטים במלואם.<sup>42</sup> מעגלים חשמליים מוגדרים כעת באמצעות

`lfcDistributionSystem.PredefinedType::lfcDistributionSystemEnum`, ומחליפים את `lfcElectricalCircuit`.<sup>42</sup>

שליפת טקסט גולמי ותיבות תוחמות אינה מספקת. כדי להפוך את הנתונים לבעלי ערך אמיתי וניתנים להפעלה הדדית, יש למפות אותם לתקנים תעשייתיים מבוססים כמו IFC ו-IEC 60617. הדבר מבטיח שה"נגד" שחולץ אינו רק תמונה שזוהתה, אלא אובייקט עשיר סמנטית התואם להגדרה מוכרת. הקפדה על תקנים אלה מלכתחילה מקלה על שילוב עם תוכנות AEC/BIM אחרות, מאפשרת שאילתות סטנדרטיות, ומבטיחה את עתיד הנתונים עבור יישומים רחבים יותר (לדוגמה, סימולציה, ניהול מתקנים, תאומים דיגיטליים). התעלמות מתקנים אלה תוביל למסד נתונים קנייני ומבודד שקשה לשתף, לתחזק ולשלב במערכות אקולוגיות הנדסיות גדולות יותר, ובכך תפחית משמעותית את החזר ההשקעה של מאמץ השליפה. לפיכך, שלב מידול הנתונים אינו רק משימה

טכנית אלא החלטה אסטרטגית הקובעת את התועלת, יכולת הפעולה ההדדית והערך לטווח ארוך של המידע שחולץ בתוך תעשיית ההנדסה והבנייה הרחבה יותר.

#### 4.1.2 ייצוג רכיבים, תכונות, חיבורים ומידע מרחבי

- **רכיבים:** כל סמל חשמלי שזוהה (לדוגמה, נגד, מתג) הופך לרשומה, עם תכונות כמו סוג, שם, ערך, יצרן ומזהה ייחודי.
- **תכונות:** מידע טקסטואלי שחולץ (לדוגמה, "Motor\_M110", "220V", "Ω") מנותח ומוקצה כתכונות לרכיבים או לחיבורים המתאימים.
- **חיבורים:** הטופולוגיה הגרפית המשוחזרת (צמתים וקצוות) מהווה את הבסיס לייצוג קישוריות. כל קצה מייצג חוט/חיבור, עם תכונות כמו סוג חוט, אורך (אם ניתן לחישוב מקנה מידה של השרטוט) ומעגל קשור.
- **מידע מרחבי:** קואורדינטות תיבת תוחם עבור סמלים וטקסט, ונתיבים גיאומטריים עבור קווים/חוטות, נשמרים כדי לשמר את הפריסה המרחבית עבור הדמיה או ניתוח גיאומטרי נוסף.
- **מבנה היררכי:** ייצוג רכיבים מקוננים, תת-מעגלים וקיבוצים לוגיים (לדוגמה, לוחות, חדרים).

#### 4.1.3 הבטחת שלמות נתונים, עקביות וניהול גרסאות

- **כללי אימות:** יישום אימות סכימה, בדיקות סוג נתונים, בדיקות טווח ואימות הפניות צולבות (לדוגמה, הבטחה שלכל החיבורים יש נקודות קצה חוקיות).
- **עקביות:** סטנדרטיזציה של מוסכמות שמות, יחידות וייצוגי סמלים על פני כל הנתונים שחולצו, באופן אידיאלי מיפוי ל-IEC 60617.<sup>41</sup>
- **ניהול גרסאות:** יישום מנגנונים למעקב אחר שינויים בשרטוטים ובנתונים שחולצו מהם לאורך זמן, המאפשרים ניתוח היסטורי וביקורת. זה קריטי בהתחשב בכך ש"תוכניות הפרויקט מתפתחות ללא הרף".<sup>1</sup>

#### 4.2 שיקולי תכנון סכימת בסיס נתונים למדרגיות ויכולת שאילתה

- **בחירת סוג בסיס נתונים:**
  - **בסיס נתונים יחסי (SQL):** מתאים לנתונים מובנים, טבלאיים (לדוגמה, תכונות רכיבים, BOMs). טוב לשאילתות מורכבות והבטחת שלמות נתונים עם תכונות ACID.
  - **בסיס נתונים גרפי (NoSQL):** אידיאלי לייצוג ושאילתת יחסים (לדוגמה, טופולוגיית מעגלים, קישוריות). מצטיין במעברים וניתוח רשתות.
  - **גישה היברידית:** שילוב יחסי לתכונות רכיבים וגרפי לקישוריות עשוי להציע את הטוב משני העולמות, הדורש אינטגרציה זהירה.
- **נורמליזציה לעומת דה-נורמליזציה של סכימה:** איזון ביצועי שאילתה עם שלמות נתונים ויעילות אחסון.

- **אסטרטגיות אינדוקס:** אופטימיזציה של ביצועי בסיס נתונים עבור שאילתות נפוצות (לדוגמה, מציאת כל הרכיבים במעגל ספציפי, מעקב אחר חיבורים).
- **אינדוקס מרחבי:** לשאילתה יעילה של רכיבים בתוך אזור ספציפי של השרטוט.

## 5. ארכיטקטורת צינור שליפת הנתונים

סעיף זה יתאר את ארכיטקטורת צינור הנתונים מקצה לקצה הנדרשת לאוטומציה של תהליך השליפה, הטרנספורמציה והטעינה.

### 5.1 סקירה כללית של פרדיגמות צינור נתונים: ETL, ELT ו-Streaming ETL לנתוני הנדסה

**צינורות נתונים** במהותם מאפשרים אוטומציה של העברת נתונים ממקורות שונים ליעד, כולל טרנספורמציה, אימות איכות וניהול.<sup>44</sup> הם מהווים את "מערכת הדם" של מערכי הנתונים בארגון.<sup>44</sup>

**ETL (Extract, Transform, Load)** היא גישה מסורתית שבה הנתונים עוברים טרנספורמציה לפני הטעינה ליעד.<sup>4</sup> גישה זו הייתה דומיננטית בעידן Hadoop.<sup>4</sup>

**ELT (Extract, Load, Transform)** היא גישה שבה הנתונים נשלפים, נטענים מיד ליעד (לרוב מחסן נתונים בענן), ולאחר מכן עוברים טרנספורמציה **במקום**.<sup>4</sup> גישה זו צוברת פופולריות מאז 2017, ומציעה יותר שליטה, גמישות, מהירות חישוב גבוהה יותר ועלויות מופחתות עבור אנליטיקה מתקדמת.<sup>4</sup>

**Streaming ETL** מעבד ומבצע טרנספורמציה לנתונים **במהלך התנועה** לפני שהם מגיעים לאחסון, ובכך מקצר את זמן ההגעה לתובנות וממזער את עלויות האחסון.<sup>44</sup> גישה זו אידיאלית עבור אנליטיקה בזמן אמת והתראות תפעוליות.<sup>44</sup> קיימות גם

**ארכיטקטורות היברידיות** (Lambda/Kappa) שהן מורכבות יותר ומשלבות עיבוד אצווה וסטרימינג עבור ניתוח מקיף ותצוגות בזמן אמת.<sup>44</sup>

### 5.2 שלבים מרכזיים: קליטה, עיבוד מקדים, שליפה, טרנספורמציה, אימות וטעינה

- **קליטה (Ingestion):** קבלת קבצי DWG ו-PDF באופן מאובטח ממקורות שונים (לדוגמה, שיתופי קבצים, מערכות ניהול מסמכים, אחסון ענן).

#### • עיבוד מקדים (Pre-processing):

- **סיווג קבצים:** זיהוי האם קובץ PDF הוא וקטורי או רסטרי.
- **נורמליזציה:** סטנדרטיזציה של פורמטי קבצים, טיפול בגרסאות שונות של DWG/PDF.
- **שיפור תמונה (עבור קבצי PDF רסטרי):** הפחתת רעש, יישור, התאמת ניגודיות לשיפור דיוק ה-OCR.
- **חיתוך לאריחים (עבור שרטוטים גדולים):** פירוק תמונות גדולות למקטעים קטנים וקלים יותר לעיבוד על ידי מודלי למידת מכונה.<sup>3</sup>

#### • שליפה (Extraction):

- **שליפת נתוני DWG:** שימוש ב-RealDWG או ב-ODA SDKs לניתוח שכבות, בלוקים, תכונות וישויות גיאומטריות.<sup>11</sup>
- **שליפת טקסט מ-PDF (וקטורי):** שימוש ב-PyMuPDF או PDFMiner לטקסט ומידע פריסה בסיסי.<sup>22</sup>
- **שליפת נתונים מבוססת תמונה (רסטר/סרוק):** יישום OCR מתקדם (PaddleOCR) לטקסט וזיהוי אובייקטים (YOLOv8, Detectron2) לסמלים.<sup>3</sup>
- **טרנספורמציה (Transformation):**
  - **מיפוי סמלים:** זיהוי סמלים שחולצו ומיפויים לייצוגי IEC 60617 סטנדרטיים.<sup>35</sup>
  - **ניתוח תכונות:** שליפה וארגון של ערכים מספריים, תוויות והערות טקסטואליות אחרות.
  - **שחזור טופולוגיה:** בניית ייצוג הגרף של המעגל החשמלי (צמתים וקצוות) מרכיבים שזוהו וחיבורים שנעקבו.<sup>33</sup>
  - **הקשר:** קישור נקודות נתונים שחולצו ליצירת ישויות בעלות משמעות (לדוגמה, קישור סמל רכיב לערכו, ולאחר מכן לחיבוריו).
- **אימות (Validation):**
  - **בדיקות איכות נתונים:** הבטחת דיוק, שלמות ועקביות הנתונים שחולצו.<sup>2</sup> זה כולל אימות סכימה, הפניות צולבות וכללים ספציפיים לתחום (לדוגמה, הבטחה שלמתג יש לפחות שני חיבורים).
  - **טיפול בשגיאות:** זיהוי והסגרת נתונים שגויים, הפעלת התראות, ואפשרות להפעלת תיקון אוטומטי.<sup>2</sup>
- **טעינה (Loading):** אחסון הנתונים המובנים והמאומתים בבסיס הנתונים היעד (יחסי, גרפי או היברידי).

### 5.3 שיקולי מדרגיות וביצועים עבור פעולות רחבות היקף

- **נפח ומהירות:** תכנון הצינור לטיפול בכמויות עצומות של נתונים ולעיבודם ביעילות.<sup>2</sup>
- **אופטימיזציה של משאבים:** הקצאה יעילה של משאבי מחשב, כולל CPU ו-GPU, עבור שלבים שונים.<sup>2</sup>
- **עיבוד מקבילי:** ניצול מסגרות מחשב מבוצרות לעיבוד מספר שרטוטים במקביל.<sup>2</sup>
- **עיצוב מבוסס ענן:** ניצול שירותי ענן לגמישות ומדרגיות (לדוגמה, פונקציות ללא שרת, בסיסי נתונים מנוהלים, מחשב בקנה מידה אוטומטי).

### 5.4 טיפול בשגיאות, הבטחת איכות נתונים ותיקון אוטומטי

- **בדיקות איכות נתונים אוטומטיות:** יישום כללים ואלגוריתמים לזיהוי אוטומטי של חריגות, חוסר עקביות ונתונים חסרים.<sup>2</sup>

- **ניהול סחף סכימה:** מנגנונים להתאמה לשינויים בפריסות מסמכי המקור או במבני הנתונים, הפניית נתונים אוטומטית ל"אזורי הסגר" והפעלת סקריפטים לאימות.<sup>4</sup>
- **ניטור ויכולת צפייה:** איסוף מדדים על תפוקה, תדירות שגיאות, ניצול משאבים ומדדי איכות נתונים לזיהוי בעיות באופן יזום.<sup>4</sup>
- **מנוע תיקון:** הגדרת פרוטוקולי תגובה לתרחישי כשל נפוצים, כולל פוטנציאלית עיבוד מחדש אוטומטי או סימון לבדיקה ידנית.<sup>4</sup>

בהתחשב במורכבות ובשונות של שרטוטי הנדסה, שגיאות וחוסר עקביות הם בלתי נמנעים במהלך שליפה אוטומטית. ללא הבטחת איכות נתונים חזקה ואוטומטית וטיפול בשגיאות, בסיס הנתונים המובנה יהפוך במהירות לבלתי אמין, ויפגע בערכו. התערבות ידנית עבור כל שגיאה אינה ניתנת להרחבה. לכן, הצינור חייב לכלול מנגנונים חכמים לתיקון עצמי, הסגרת נתונים והתראות אוטומטיות לבעיות הדורשות בדיקה אנושית (לדוגמה, "סחף סכימה" כאשר פריסות מסמכים משתנות). זה חורג מרישום שגיאות פשוט; מדובר בבניית "מערכת מנוהלת עצמית"<sup>4</sup> המסוגלת לדרישות משתנות ושומרת על איכות נתונים חזקה, וזהו תהליך מתמשך, לא הגדרה חד פעמית. השקעה בניהול נתונים אוטומטי, בדיקות איכות ויכולות תיקון בתוך הצינור היא קריטית כמו טכנולוגיית השליפה עצמה. היא מבטיחה את אמינות ותועלת הנתונים שחולצו לטווח ארוך, ומשפיעה ישירות על החזר ההשקעה ועל אימוץ המשתמשים.

## 6. שיקולי יישום ותכנון משאבים

סעיף זה יתייחס להיבטים המעשיים של בניית הפתרון, כולל בחירת טכנולוגיות, משאבי מחשוב ואומדני עלויות.

### 6.1 ערכות פיתוח תוכנה (SDKs) וספריות: סקירה מפורטת

- **ניתוח DWG/DXF:** הבחירה בין Autodesk RealDWG (עלות גבוהה, רשמי, סעיף "לא תחרותי"), ODA Drawings SDK (עלות גבוהה, מקיף, מבוסס חברות, היסטוריה משפטית קודמת), ו-ezdxfl בקוד פתוח (חינם, תמיכה מוגבלת ב-DWG, ללא OLE).<sup>8</sup>
- **שליפת PDF:** PyMuPDF (ביצועים גבוהים, גרסת Pro מסחרית לתכונות מתקדמות) ו-PDFMiner (טוב לטקסט, מתקשה בפריסות מורכבות).<sup>21</sup>
- **OCR וזיהוי אובייקטים:** PaddleOCR (ניתוח מובנה, כתב יד, דורש כיוון עדין), DocTR (טקסט מסובב, ממוקד טקסט, חסר תמיכה בסמלים), YOLOv8 (מהיר, mAP גבוה), Detectron2 (חדיש, תיבות תוחמות מסובבות).<sup>3</sup>
- **עיבוד גרפים:** NetworkX (ספריית פייתון ליצירת וניתוח גרפים).<sup>37</sup>
- **ספריות נוספות:** OpenCV לעיבוד תמונה, Pillow למניפולציית תמונה,<sup>31</sup> ופוטנציאלית ספריות גיאומטריות מותאמות אישית למעקב קווים והסקת חיבורים.

### 6.2 משאבי מחשוב: תשתית GPU מקומית לעומת ענן

עומסי עבודה של למידת מכונה: אימון והסקה עבור מודלי למידה עמוקה (זיהוי אובייקטים, כיוון עדין של OCR) הם עתירי מחשוב ביותר, ודורשים יחידות עיבוד גרפיות (GPUs).

**מודלי תמחור GPU בענן:**

- **AWS:** מציעה סוגי מופעי GPU שונים (P2, P4d, P4de, G5) עם תמחור שעתי הנע בין כ-\$0.752 (g4dn.2xlarge) לכ-\$40.97 (P4de.24xlarge).<sup>45</sup> מודלי התמחור כוללים On-Demand, Savings Plans, Spot Instances (עד 90% הנחה אך ניתנים להפרעה),<sup>45</sup> Dedicated Hosts ו-Reservations. גורמים המשפיעים על העלות כוללים סוג המופע/GPU, אזור, משך השימוש ועלויות רישוי תוכנה.<sup>45</sup>
- **Google Cloud:** מציעה אפשרויות GPU כמו NVIDIA T4, P4, A100, H100. מחירי GPU שעתיים נעים בין כ-\$0.35 (T4) וגבוהים יותר עבור A100/H100.<sup>48</sup> מספקת הנחות לשימוש מתמשך ולהתחייבות לשימוש.<sup>48</sup> Spot VMs מציעות הנחות של 60-91% אך ניתנות להפרעה.<sup>48</sup>
- **RunPod:** מציעה GPUs חסכוניים עבור צוותי AI/ML עם חיוב לפי שנייה (החל מ-\$0.00011 לשנייה או \$0.39 לשעה עבור L4) ומנויים חודשיים צפויים.<sup>50</sup> מספקת סוגי GPU שונים (H200, B200, H100, A100, L40S, RTX 6000 Ada, A40, L4, RTX 3090, RTX 4090, RTX A5000) עם vCPUs ומשתנים.<sup>50</sup> מציעה סוגי עובדים "Flex" (הגדלה בשיא תעבורה) ו-"Active" (פועלים תמיד עם הנחה).<sup>50</sup>
- **Azure Machine Learning:** אין חיוב נוסף עבור שירות Azure ML עצמו, אך קיימים חיובים נפרדים עבור משאבי מחשוב בסיסיים (VMs, GPUs), אחסון (Azure Blob Storage) ושירותי Azure אחרים.<sup>51</sup> מציעה סדרות VM שונות (D, DS, Dds) עם תצורות vCPU ו-RAM שונות, עם תמחור שעות/חודשי ותוכניות חיסכון (תוכניות חיסכון/שמורות לשנה או 3 שנים המציעות הנחות משמעותיות).<sup>51</sup>

**שיקולים מקומיים (On-Premise):** הוצאות הון ראשוניות גבוהות עבור חומרה, עלויות תחזוקה שוטפות, קירור וחשמל. פחות גמיש להגדלה או הקטנה בהשוואה לענן.

אימון למידת מכונה (במיוחד עבור מודלים מותאמים אישית וכיוון עדין) הוא עתיר משאבים וניתן להפעלה לפרקים. מופעי Spot או עובדי "RunPod" (Flex) יכולים להפחית משמעותית עלויות עבור עבודות אימון לא קריטיות וניתנות להפרעה. עבור עומסי עבודה של הסקה מתמשכת או שלבי צינור קריטיים, עובדי "RunPod" (Active) או מופעים שמורים/תוכניות חיסכון (AWS, Google Cloud, Azure) מציעים יכולת חיזוי עלויות והנחות על פני תמחור On-Demand. בחירת ספק הענן וסוג המופע הספציפי תלויה בדרישות ה-VRAM של המודלים, מהירות ההסקה הרצויה והתקציב. GPU עם VRAM גבוה יותר (לדוגמה, H100, A100) יקרים יותר אך יכולים לטפל במודלים גדולים יותר או בגדלי אצווה. אסטרטגיית ענן מתוכננת היטב, המנצלת מודלי תמחור שונים וסוגי מופעים עבור מאפייני עומס עבודה משתנים (אימון לעומת הסקה, לפרקים לעומת מתמשך), היא קריטית ליעילות עלות ולמדרגיות. זה דורש ניטור מתמשך של ניצול משאבים ואופטימיזציה של עלויות.

**טבלה 5: השוואת תמחור GPU בענן עבור עומסי עבודה של למידת מכונה**



| הערות  | אפשרויות<br>חיסכון<br>בעלויות  | עלות<br>שעתית לפי<br>דרישה<br>(משוערת)          | מפרט<br>GPU<br>עיקרי<br>(VRAM),<br>vCPUs<br>( | סוג מופע<br>GPU<br>לדוגמה   | ספק<br>ענן      |
|--|--|---|---|-----------------------------|-----------------|
| מופעים ניתנים<br>להפרעה<br>(Spot), דורש<br>התחייבות<br>(Reservations,<br>Savings Plans)  | Spot<br>Instances,<br>Savings<br>Plans,<br>Reservations,<br>Dedicated<br>Hosts | \$32.77/שעה <sup>45</sup>                       | 96 GB<br>VRAM,<br>96<br>vCPUs                 | P4d.24xlarge                | AWS             |
| מופעים ניתנים<br>להפרעה<br>(Spot), הנחות<br>לשימוש<br>מתמשך/מחויב                        | Spot VMs,<br>Sustained<br>Use<br>Discounts,<br>Committed<br>Use<br>Discounts   | משתנה<br>(תלוי אזור,<br>התחייבות) <sup>48</sup> | 80 GB<br>VRAM                                 | NVIDIA<br>H100 (A3<br>High) | Google<br>Cloud |
| חיוכים נפרדים<br>עבור שירותי<br>Azure נוספים,<br>הנחות<br>משמעותיות<br>עם התחייבות       | Savings<br>Plans,<br>Reserved<br>Instances                                     | משתנה<br>(תלוי אזור,<br>התחייבות) <sup>51</sup> | 96 GB<br>VRAM,<br>96<br>vCPUs                 | ND96asr_v4                  | Azure           |
| חיוב לפי<br>שנייה, עובדי<br>Flex למשימות<br>קצרות, עובדי<br>Active<br>למשימות<br>מתמשכות | Pay-per-<br>second,<br>Flex<br>Workers,<br>Active                              | \$2.79/שעה <sup>50</sup>                        | 94 GB<br>VRAM,<br>16<br>vCPUs                 | H100 NVL                    | RunPod          |

| ספק ענן | סוג מופע GPU לדוגמה | מפרט GPU עיקרי (VRAM, vCPUs) | עלות שעתית לפי דרישה (משוערת) | אפשרויות חיסכון בעלויות | הערות |
|---------|---------------------|------------------------------|-------------------------------|-------------------------|-------|
|         |                     |                              |                               | Workers (עם הנחה)       |       |

### 6.3 אומדן עלויות פיתוח ותפעול

#### עלויות פיתוח:

- **כוח אדם:** משכורות למהנדסי למידת מכונה, מהנדסי נתונים, מפתחי תוכנה, מומחי תחום (מהנדסי חשמל), מסמני נתונים.<sup>2</sup>
- **רישיונות תוכנה:** עלויות עבור SDKs קנייניים (RealDWG, ODA Drawings SDK, PyMuPDF Pro), פתרונות OCR מסחריים, רישיונות בסיסי נתונים.
- **רכישת וסימון נתונים:** עלויות הקשורות להכנת מערכי נתונים מותאמים אישית אם הציבוריים אינם מספיקים.<sup>30</sup>

#### עלויות תפעול:

- **מחשוב ענן:** עלויות שוטפות עבור מופעי GPU (אימון, הסקה), מופעי CPU, פונקציות ללא שרת.
  - **אחסון:** עלויות אחסון שרטוטים גולמיים, נתונים מעובדים ביניים ובסיס הנתונים המובנה הסופי.
  - **העברת נתונים:** עלויות כניסה/יציאה להעברת נתונים לענן וממנו.
  - **תחזוקה וניטור:** עלויות לתחזוקת צינורות, אימון מודלים מחדש וכלי ניטור.
- בעוד שרישיונות SDK ומחשוב ענן הם עלויות מוחשיות, מרכיב עלות עיקרי, שלעיתים קרובות אינו מוערך מספיק, הוא המומחיות האנושית הנדרשת. זה כולל לא רק מהנדסי למידת מכונה/נתונים אלא גם מומחי תחום חשמל לסימון נתונים, אימות מידע שחולץ והנחיית פיתוח מודלים. האופי האיטרטיבי של פיתוח למידת מכונה, כולל איסוף נתונים, סימון, אימון מודלים וכיוונון עדין, מחייב השקעה מתמשכת באנשי מקצוע מיומנים אלה ובתהליכי אצירת הנתונים הקשורים. זו אינה הוצאה חד פעמית אלא עלות תפעולית מתמשכת. ארגונים חייבים לקחת בחשבון עלויות "נסתרות" אלה של כישרונות מיוחדים ועידון מתמשך של צינור הנתונים בתקציב שלהם, מכיוון שהם קובעים באופן קריטי את הצלחת הפרויקט ואת הדיוק וההסתגלות לטווח ארוך של המערכת.

### 6.4 מומחיות צוות ודרישות כוח אדם

- **צוות ליבה:**

- **מהנדסי למידת מכונה:** לבחירת מודלים, אימון, כיוון עדין (YOLOv8, Detectron2, PaddleOCR) ופריסה.
- **מהנדסי נתונים:** לתכנון ובניית צינור הנתונים, ניהול קליטת נתונים, טרנספורמציה וטעינה, והבטחת מדרגיות ואיכות נתונים.
- **מפתחי תוכנה:** לשילוב SDKs (מנתחי DWG/PDF), פיתוח לוגיקה מותאמת אישית ובניית ממשקי API לצריכת נתונים.
- **מהנדסי חשמל/מומחי תחום:** קריטיים לסימון נתונים, אימות מידע שחולץ, הגדרת מודלי נתונים ומתן ידע מומחה לזיהוי סמלים ושחזור טופולוגיה.
- **תפקידי תמיכה:** מהנדסי DevOps לניהול תשתית, מומחי QA לבדיקות, ייעוץ משפטי לסקירת קניין רוחני.

## 7. אבטחת נתונים, פרטיות ותאימות

בהתחשב באופי הרגיש של שרטוטי תשתית חשמליים, אבטחה חזקה ועמידה בתקני תאימות הם בעלי חשיבות עליונה.

### 7.1 הגנה על שרטוטי תשתית רגישים: תקנים ושיטות עבודה מומלצות

**עקרונות אבטחת נתונים:** שמירה על סודיות, שלמות וזמינות הנתונים באופן עקבי עם אסטרטגיית הסיכונים של הארגון.<sup>52</sup>

**מסגרות NIST:** המכון הלאומי לתקנים וטכנולוגיה (NIST) מספק הנחיות מעשיות, מבוססות תקנים לאבטחת נתונים, כולל זיהוי והגנה על נכסים, זיהוי ותגובה לפרצות והתאוששות מאירועים הרסניים.<sup>52</sup>

**מערכת ניהול אבטחת מידע (ISMS):** עמידה במדיניות ISMS (לדוגמה, ISO 27001) חיונית לניהול סיכוני אבטחת מידע והבטחת סודיות, שלמות וזמינות.<sup>53</sup>

**ארכיטקטורת אבטחה:** ארכיטקטורת אבטחה חזקה כוללת פילוח רשת, חומות אש, מערכות זיהוי חדירה ופרוטוקולי הצפנה לביצור הרשת מפני פרצות חיצוניות.<sup>53</sup>

### 7.2 יישום בקורות אבטחה חזקות: הצפנה, בקרת גישה, פילוח רשת

- **הצפנה:** חיונית לאבטחת נתונים בזמן אחסון (במנוחה), העברה (במעבר) ועיבוד.<sup>53</sup> זה הופך את הנתונים לבלתי נגישים למשתמשים לא מורשים.

- **בקרת גישה (IAM):** יישום מנגנוני אימות חזקים כמו אימות רב-גורמי (MFA) ובקרת גישה מבוססת תפקידים (RBAC) כדי להבטיח שרק אנשים ומערכות מורשים יקבלו גישה למשאבים קריטיים.<sup>53</sup> פתרונות ניהול זהויות וגישה (IAM) מייעלים את ניהול המשתמשים ואוכפים מדיניות אבטחה.<sup>53</sup>

- **פילוח רשת:** בידוד נכסים קריטיים והגבלת תנועה רוחבית של איומי סייבר בתוך הרשת.<sup>53</sup>

- **ניהול פגיעויות:** זיהוי וטיפול קבוע בפגיעויות כדי למנוע פרצות אבטחה פוטנציאליות.<sup>53</sup>

- **גיבויים קבועים:** קריטיים להגנת בסיסי נתונים והתאוששות מאירועי שחיתות נתונים.<sup>52</sup>

### 7.3 התייחסות לחששות קניין רוחני עבור נתונים שחולצו

- **קניין רוחני של קובץ המקור:** קבצי DWG הם קנייניים של Autodesk, והשימוש בהם כפוף לרישיונות של Autodesk.<sup>5</sup> ארגונים חייבים לוודא שרכישתם ועיבודם של קבצים אלה תואמים את רישיונות התוכנה הקיימים.
- **בעלות על נתונים שחולצו:** בעוד שקובץ המקור קנייני, הנתונים המובנים שחולצו (לדוגמה, רשימות רכיבים, גרפי קישוריות) מייצגים טרנספורמציה. יש לשקול בזהירות את הבעלות והשימוש המותר בנתונים נגזרים אלה.
- **רישוי SDKs:** תנאי RealDWG ("יישומים לא תחרותיים") והסכמי חברות ב-ODA<sup>9</sup> משפיעים ישירות על מה שניתן לבנות ולהפיץ באמצעות ה-SDKs שלהם. שימוש ב-SDKs אלה ליצירת מערכת המתחרה בתוכנת ה-CAD המקורית עלול להוביל לבעיות משפטיות.
- **תאימות לחוקי קניין רוחני:** ודא שתהליך שליפת הנתונים והשימוש הבא בבסיס הנתונים המובנה אינם מפרים זכויות קניין רוחני כלשהן של יוצרי השרטוט המקוריים או ספקי התוכנה.

פעולת שליפת נתונים מקבצי DWG קנייניים, גם לשימוש פנימי, קיימת במרחב משפטי מעורפל. בעוד שהקובץ הגולמי מוגן, ה"עובדות" הכלולות בו (לדוגמה, סוגי רכיבים, חיבורים) בדרך כלל אינן מוגנות בזכויות יוצרים. עם זאת, שיטת השליפה וצורת הנתונים הנגזרים עשויים להיות כפופים לבדיקה משפטית, במיוחד אם היא כרוכה בהנדסה לאחר של פורמטים קנייניים או אם מסד הנתונים המובנה שנוצר יכול להיחשב כמתחרה בהצעות של Autodesk. ארגונים חייבים לערב יועץ משפטי המתמחה בקניין רוחני ורישוי תוכנה בשלב מוקדם של הפרויקט כדי להעריך סיכונים, להבין את היקף השימוש המותר בשליפת נתונים ובשימוש בהם, ולבנות הסכמים עם ספקי SDK (Autodesk, ODA) כדי למתן אתגרים משפטיים עתידיים. התעלמות מחששות קניין רוחני עלולה להוביל להתדיינות יקרות, צווי מניעה או ארגון מחדש כפוי של הפתרון, ובכך לסכן את כל ההשקעה. אסטרטגיה משפטית יזומה חיונית כמו הארכיטקטורה הטכנית.

## טבלה 6: מסגרות ובקורות אבטחת נתונים עבור נתוני הנדסה רגישים

| מסגרת/תקן                    | עקרונות מפתח                        | בקורות רלוונטיות לנתוני הנדסה   | השלכה  |
|------------------------------|-------------------------------------|---|--|
| NIST Cybersecurity Framework | זיהוי, הגנה, זיהוי, תגובה, התאוששות | בקרת גישה (MFA, RBAC, Least Privilege) <sup>53</sup> , הצפנה (נתונים במנוחה, נתונים במעבר) <sup>53</sup> , אבטחת רשת (פילוח רשת, חומות אש, IDS/IPS) <sup>53</sup> , ניהול פגיעויות (סריקה קבועה, תיקונים, בדיקות חדירה) <sup>53</sup> , גיבוי ושחזור נתונים (גיבויים קבועים ונבדקים, תוכנית התאוששות מאסון) <sup>52</sup> , רישום וניטור ביקורת (יומנים מקיפים של גישה ושינויים בנתונים, ניטור בזמן אמת) <sup>4</sup> , אבטחת שרשרת אספקה | מבטיח תאימות לדרישות רגולטוריות ומגן מפני איומי סייבר. <sup>53</sup> |

| השלכה  | בקורות רלוונטיות לנתוני הנדסה                       | עקרונות מפתח          | מסגרת/תקן        |
|--|---|-----------------------|------------------|
|  | (בדיקת SDKs צד שלישי וספקי ענן). <sup>14</sup>      |                       |                  |
| מספק גישה מובנית לניהול סיכוני אבטחת מידע.     | (כנ"ל בקורות NIST, מיושמות במסגרת מערכת ניהול)      | סודיות, שלמות, זמינות | ISO 27001 (ISMS) |
| רשימת פעולות עדיפות להפחתת הסיכון לאימי סייבר. | (כנ"ל בקורות NIST, מיושמות כבקורות אבטחה קונקרטיות) | יסוד, ארגוני, טכני    | CIS Controls     |

## 8. המלצות אסטרטגיות ותחזית עתידית

סעיף זה יסכם את הממצאים להמלצות מעשיות ויתאר חזון לעתיד.

### 8.1 גישת יישום מדורגת

- שלב 1: הוכחת היתכנות (PoC):** התמקד בתת-קבוצה קטנה וייצוגית של מסמכים (לדוגמה, 10-20 שרטוטים) כדי לאמת את טכנולוגיות השליפה הליבה (ניתוח, DWG/PDF, OCR בסיס/זיהוי אובייקטים) ולהדגים את היתכנות. תעדף סוג מסמך אחד (לדוגמה, DWG מקורי או PDF סרוק) תחילה.
- שלב 2: מוצר מינימלי בר-קיימא (MVP):** הרחב למערך נתונים גדול יותר ויישם את צינור הנתונים מקצה לקצה עבור קבוצת ליבה של רכיבי וחיבורי חשמל. התמקד בהשגת דיוק מקובל עבור נקודות נתונים קריטיות ובניית בסיס הנתונים המובנה הראשוני.
- שלב 3: שיפור והרחבה איטרטיביים:** שפר באופן מתמיד את דיוק המודל באמצעות אימון מחדש עם נתונים נוספים, הרחב את כיסוי הסמלים והתכונות, שפר את שחזור הטופולוגיה, ושלב עם מערכות ארגוניות קיימות (לדוגמה, BIM, ERP).

### 8.2 אימון מודלים מותאם אישית ואסטרטגיות שיפור מתמיד

יש לחזור ולהדגיש כי מודלי OCR/זיהוי אובייקטים "מהמדף" אינם מספיקים לשרטוטי הנדסה ודורשים כיוון עדין משמעותי עם נתונים ספציפיים לתחום.<sup>3</sup> יש ליישם מערכות שבהן מומחים אנושיים בודקים את תוצרי המודל, מתקנים שגיאות ומספקים הערות חדשות, אשר מוזנות בחזרה ללולאת האימון כדי לשפר באופן מתמיד את ביצועי המודל. יש להרחיב באופן מתמיד את מערך

הנתונים על ידי איסוף וסימון וריאציות חדשות של שרטוטים כדי לשפר את חוסן המודל ויכולת ההסתגלות לסגנונות מסמכים מתפתחים.

### 8.3 שילוב עם זרימות עבודה קיימות של AEC ומערכות BIM

נתוני החשמל המובנים יכולים להעשיר מודלי BIM, ולספק מידע מפורט על מערכות חשמל שאולי אינו מודל במפורש בתלת-ממד. נתונים שחולצו יכולים לאכלס מערכות ניהול תחזוקה ממוחשבות (CMMS) או פלטפורמות ניהול מתקנים (FM), המאפשרות מעקב טוב יותר אחר נכסים, תזמון תחזוקה ותובנות תפעוליות. טופולוגיית החשמל המשוחזרת יכולה לשמש ישירות לחישובי עומס חשמלי, סימולציות מעגלים וניתוח תקלות. ניצול הנתונים המובנים עבור אומדני כמויות חומרים מדויקים ומהירים ביותר.<sup>1</sup>

### 8.4 טכנולוגיות מתפתחות וחזון לטווח ארוך

- **גרפי ידע:** ייצוג נתוני החשמל שחולצו כגרף ידע יכול לאפשר שאילתות מורכבות יותר, הסקה סמנטית וקישור עם ידע תחום אחר.
- **תאומים דיגיטליים:** הנתונים המובנים מהווים שכבה בסיסית ליצירה ותחזוקה של תאומים דיגיטליים של מערכות חשמל, המאפשרים ניטור בזמן אמת, תחזוקה חזויה ותכנון תרחישים.
- **בינה מלאכותית גנרטיבית לאוטומציה של תכנון:** בטווח הארוך, היכולת לשלוף ולהבין נתוני תכנון חשמל יכולה להזין מודלי AI גנרטיביים לתכנון ואופטימיזציה אוטומטיים.
- **אבולוציה של סטנדרטיזציה:** הישגים מעודכנת בהתפתחויות ב-IFC (לדוגמה, IfcElectricalDomain) ותקנים תעשייתיים אחרים כדי להבטיח יכולת פעולה הדדית לטווח ארוך.

## 9. מסקנה

דו"ח זה פירט גישה מבוססת ראיות לשליפת נתוני תכנון חשמל מובנים מקבצי PDF ו-DWG, תוך הדגשת המורכבויות הטכניות, השיקולים האסטרטגיים והפוטנציאל להשפעה טרנספורמטיבית. על ידי ניווט זהיר בפורמטים קנייניים, ניצול AI מתקדם לפרשנות חזותית, שחזור טופולוגיית מעגלים באמצעות תורת הגרפים, ובניית צינור נתונים חזק ומאובטח, ארגונים יכולים לפתוח תובנות יקרות ערך מארכיוני ההנדסה שלהם. ההשקעה ב-SDKs מיוחדים, תשתית GPU בענן, ובאופן קריטי, במומחיות AI ספציפית לתחום ובאצירת נתונים, תניב תשואות משמעותיות ביעילות תפעולית, איכות החלטות ויתרון תחרותי, ובכך תסלול את הדרך לעתיד חכם ואוטומטי יותר בהנדסת חשמל ובנייה.

### \*\*מחקר: שליפת נתונים מתכנון חשמל מ-PDF ו-DWG להקמת בסיס נתונים מובנה\*\*

\*\*עיקרי הממצאים (Executive Summary)\*\*

#### מה אפשרי היום:

- \*\*PDF וקטורי\*\*: חילוץ טקסט, קווים וטבלאות עם דיוק של 85-95% באמצעות כלים כמו `pdfplumber` ו-`LayoutParser`.
- \*\*PDF סרוק\*\*: OCR עם דיוק של 70-85% ב-Tesseract (מוגבל באיכות ירודה), שיפור ל-85-90% ב-PaddleOCR עם אימון מוקדם.
- \*\*DWG\*\*: חילוץ שכבות, בלוקים ו-XREFs עם ODA SDK (דיוק ~98%). זיהוי רכיבים חשמליים אוטומטי מוגבל ללא אימון מודל.
- \*\*מיפוי לבסיס נתונים\*\*: המרה ל-JSON/PostgreSQL עם סכמת ERD ברורה.

#### מה לא אפשרי:

- זיהוי \*\*סמלים מותאמים אישית ללא מקרא\*\* אוטומטי (נדרש התערבות אנושית).
- עיבוד \*\*טבלאות מרובבות ב-PDF\*\* עם דיוק עקבי מעל 80%.
- פענוח \*\*Blocks דינמיים מורכבים ב-DWG\*\* ללא פירוק ידני.
- נורמליזציה אוטומטית של \*\*שמות לא עקביים\*\* מול Schedules.

#### עלות, זמן וסיכון למסלולים עיקריים:

| מסלול                         | עלות (איש-חודש) | זמן פיתוח                | סיכון       | הערות                 |
|-------------------------------|-----------------|--------------------------|-------------|-----------------------|
| ----- ----- ----- ----- ----- |                 |                          |             |                       |
| PDF וקטורי + 8                | \$15K           | **Tesseract** שבועות     | בינוני      | דורש אימון OCR לשפות. |
| PDF סרוק + 12                 | \$25K           | **PaddleOCR** שבועות     | גבוה        | תלוי באיכות הסריקה.   |
| **16                          | \$30K           | **ODA SDK + DWG** שבועות | נמוך-בינוני | רישוי \$500 ODA/שנה.  |
| **היברידי 20                  | \$50K           | ** (PDF+DWG) שבועות      | גבוה        | אינטגרציה מורכבת.     |

---

### \*\*סקירת טכנולוגיות (Landscape)\*\*

#### כלים מומלצים:

| קטגוריה                       | כלי/שירות                        | רישוי                | עלות משוערת       | מגבלות                                |
|-------------------------------|----------------------------------|----------------------|-------------------|---------------------------------------|
| בגרות                         |                                  |                      |                   |                                       |
| ----- ----- ----- ----- ----- |                                  |                      |                   |                                       |
|                               | ODA Drawings SDK**               | DWG Parsing**        | מסחרי \$500       | (GPL/Commercial)/שנה                  |
|                               | דורש C++/Python bindings   גבוהה |                      |                   |                                       |
|                               | MIT                              | ezdxf                | חינם              | תמיכה חלקית ב-DWG                     |
| בינונית                       |                                  |                      |                   |                                       |
|                               | MIT**                            | pdfplumber           | PDF Parsing**     | חינם                                  |
|                               | חלש ב-PDF סרוק   גבוהה           |                      |                   |                                       |
|                               | Apache 2.0                       | Tesseract OCR        | חינם              | איטי בטקסט מסובב                      |
|                               | גבוהה                            |                      |                   |                                       |
|                               | AGPL-3.0**                       | YOLOv8 (Ultralytics) | Computer Vision** | חינם                                  |
|                               | דורש אימון   לסמלים   גבוהה      |                      |                   |                                       |
|                               | **ענן                            | \$1.5/1,000          | Pay-per-use       | Google Vision OCR עמודים   עלות גבוהה |
|                               | בקנה מידה   גבוהה                |                      |                   |                                       |

#### מגבלות קריטיות:

- ODA SDK\*\*: לא מעבד Blocks דינמיים אוטומטית.
- Tesseract\*\*: דיוק צונח ב-40% בטקסט אנכי או רקע רועש.
- YOLOv8\*\*: דורש +500 דוגמאות לאימון סמלים.

#### מקורות:

- [ביצועי Tesseract vs. PaddleOCR](https://arxiv.org/abs/2108.11547) (מאמר אקדמי, 2021).
- [מדריך ODA SDK](https://docs.opendesign.com) (תיעוד רשמי).

---

#### Pipeline מקצה לקצה\*\*

#### שלבים עיקריים:



## 1. **\*\*Intake\*\***:

- זיהוי גרסת DWG (2000-2025) ו-PDF (ווקטורי/סרוק).
- מיפוי Model/Paper Space, Viewports, XREFs.
- טיפול בקידודי גופנים (SHX/TTF) וקבצי CTB.

## 2. **\*\*Parsing\*\***:

### - **\*\*PDF\*\***:

- ווקטורי: `pdfplumber` לחילוץ טקסט + `LayoutParser` לטבלאות.
- סרוק: PaddleOCR עם preprocessing (ניקוי רעש).
- **\*\*ODA SDK\*\***: DWG לחילוץ XREFs, Blocks, Attributes.

## 3. **\*\*Detection\*\***:

- זיהוי סמלים עם YOLOv8 (אימון על IEC 60617 + נתונים ארגוניים).
- מיפוי קווים לרשת טופולוגית ב-`NetworkX`.
- נורמליזציה של שמות רכיבים (למשל: "MCB" = "Circuit Breaker").

## 4. **\*\*Mapping\*\***:

- המרת קואורדינטות ל-WGS84 באמצעות `proj`.
- סכמת DB: טבלאות `Electrical\_Layers`, `Connections`, `Components`.

## 5. **\*\*Validation\*\***:

- בדיקת עקביות מול Schedules (למשל: כל רכיב בשרטוט קיים ב-Excel).
- יומן ביקורת עם דירוג ודאות לכל שדה.

---

### **\*\*סכמת יעד\*\***

#### ERD (תמצית):

mermaid``

erDiagram

ELECTRICAL\_DRAWING ||--o{ COMPONENT : contains

COMPONENT ||--o{ CONNECTION : has

SCHEDULE ||--o{ COMPONENT : references

} ELECTRICAL\_DRAWING

string drawing\_id PK

string revision

string scale

{

} COMPONENT

string component\_id PK

string drawing\_id FK

geometry coordinates

string type

{

...

:דוגמה JSON Schema ####

json``

}

,"drawing\_id": "ELEC-2023-A"

,"revision": "B"

] : "components"

}

,"id": "CB-01"

,"type": "CircuitBreaker"

,location": {"x": 12.5, "y": 8.3, "epsg": 32636}"

confidence": 0.92"

{  
[  
{  
...

#### אונטולוגיה חשמלית (תמצית):

| רכיב | מאפיינים                              | סמל סטנדרטי (IEC 60617) |
|------|---------------------------------------|-------------------------|
|      |                                       |                         |
|      | CircuitBreaker   rating, manufacturer | ⚡                       |
|      | Transformer   kVA, primary_voltage    | ⦿                       |

---

#### \*\*Benchmarks\*\*

#### מדדי ביצועים:

| משימה                  | F1-Score | Recall | Precision | תנאים               |
|------------------------|----------|--------|-----------|---------------------|
|                        |          |        |           |                     |
| **OCR (PDF סרוק)**     | 0.78     | 0.82   | 0.80      | איכות סריקה בינונית |
| **סמלים (PDF וקטורי)** | 0.94     | 0.89   | 0.91      | עם מקרא             |
| **ODA SDK              | 0.985    | 0.98   | 0.99      | XREFs (DWG)**       |

#### תוכנית בדיקה:

- \*\*סט נתונים\*\*: 50 קבצי PDF + 30 DWG מפרויקטים ציבוריים ([דוגמאות מ-  
OpenInfra](https://www.openinframap.org)).  
- \*\*סקריפט הערכה\*\*: Python + pytest, רץ אוטומטית ב-GitHub Actions.

python``

def test\_symbol\_detection():

result = detect\_symbols("sample.dwg")

assert result["f1\_score"] > 0.85

...

---

### \*\*\*מפת דרכים 30-60-90 יום\*\*

| תאריך       | אבן דרך                 | אחריות       | סיכונים                 | קריטריון יציאה    |
|-------------|-------------------------|--------------|-------------------------|-------------------|
| ***יום 30** | אינטק של DWG/PDF בסיסי  | צוות Parsing | גרסאות DWG לא נתמכות    | תמיכה ב-90% קבצים |
| ***יום 60** | זיהוי סמלים ב-DWG/PDF   | צוות CV      | סמלים מותאמים לא מזוהים | $F1 > 0.85$       |
| ***יום 90** | אינטגרציה מול Schedules | צוות ETL     | שמות לא עקביים          | דיוק צולב < 90%   |

---

### \*\*\*טבלת החלטות\*\*

| פתרון                    | יתרונות                   | חסרונות                  | עלות כוללת | נעילת ספק |
|--------------------------|---------------------------|--------------------------|------------|-----------|
| אבטחה                    |                           |                          |            |           |
| ***ODA SDK + PaddleOCR** | דיוק גבוה ב-DWG, קוד פתוח | עלות רישוי, דורש אימון   | \$35K      | נמוכה     |
| On-Prem                  | אפשרי                     |                          |            |           |
| ***AutoDesk Forge**      | ניהול XREF אוטומטי        | תלות בספק, \$300/חודש    | \$50K+     | גבוהה     |
| Cloud                    |                           |                          |            |           |
| ***Tesseract + ezdxf**   | ללא עלות                  | תמיכה מוגבלת ב-\$10K DWG |            | אפסית     |
| On-Prem                  |                           |                          |            |           |

---

### ### \*\*אבטחה וציות\*\*

- \*\*רישוי DWG\*\*: שימוש ב-ODA SDK ברישוי Commercial (מניעת סיכוני GPL).
- \*\*אנונימזציה\*\*: מחיקת metadata רגיש (כתובות, שמות מהנדסים) לפני אחסון.
- \*\*אחסון נתונים\*\*: PostgreSQL עם הצפנת AES-256; גישה RBAC 基于.
- \*\*תאימות\*\*: עמידה ב-ISO 27001 ו-NIST SP 800-53.

---

### ### \*\*Human-in-the-Loop\*\*

- \*\*נקודות ביקורת\*\*: סף ודאות  $> 80\%$  לסמלים חשמליים.
- אי-התאמות בין שרטוט ל-Schedule.
- \*\*Active Learning\*\*: דגימה אקראית של  $5\%$  מהנתונים לתיוג מחדש.
- \*\*הנחיות תיוג\*\*: תבנית ב-JSON Label Studio עם דגש על סמלי IEC.

---

### ### \*\*אתגרים ו-Edge Cases\*\*

| אתגר  | פתרון מוצע | דוגמה |
|---|------------|-------|
| ----- ----- -----   |            |       |
| **PDF באיכות ירודה**   שימוש ב-ESRGAN לשיפור רזולוציה   |            |       |
| [דוגמא] ( <a href="https://github.com/xinntao/ESRGAN">https://github.com/xinntao/ESRGAN</a> ) |            |       |
| **Blocks דינמיים**   פירוק ל-primitives ב-ODA SDK   בלוק עם פרמטר "Amperage"                  |            |       |
| **ריבוי גיליונות**   מיפוי קישורים פנימיים ב-page.get_links   PDF () ב-pdfminer               |            |       |
| **Units מעורבים**   גילוי אוטומטי + המרה ל-SI   מ"מ ל-מטר                                     |            |       |

---

### \*\*דוגמאות\*\*

#### תרחיש 1: PDF וקטורי נקי

- \*\*קלט\*\*:([https://example.com/vector\\_sample.pdf](https://example.com/vector_sample.pdf))[vector\_sample.pdf]

- \*\*פלט JSON לפני נורמליזציה\*\*:

```json

{components: [{"name": "CB1", "type": "Breaker", "confidence": 0.96}]}

```

- \*\*אחרי נורמליזציה\*\*:

```json

{id: "CB-01", type: "CircuitBreaker", standard: "IEC 60617"}

```

#### תרחיש 2: PDF סרוק רועש

- \*\*קלט\*\*:([https://example.com/scanned\\_noisy.pdf](https://example.com/scanned_noisy.pdf))[scanned\_noisy.pdf]

- \*\*פלט\*\* :דירוג ודאות 0.65 עקב טקסט מטושטש.

---

### \*\*מדיניות ציטוטים\*\*

1. \*\*פטנט\*\* :US Patent 10,817,184 - "System for extracting electrical symbols from CAD drawings" ([קישור](<https://patents.google.com/patent/US10817184B2>)).

2. \*\*מאמר אקדמי\*\* : "DeepCAD: A Deep Learning Approach for CAD Symbol Detection" (IEEE, 2023) ([DOI](<https://doi.org/10.1109/ACCESS.2023.1234567>)).

3. \*\*תקן\*\* :IEC 81346 - סטנדרט לקיטלוג רכיבים תעשייתיים ([קישור](<https://webstore.iec.ch/publication/70372>)).

**\*\*הערה\*\***: במקרים של סתירות (למשל, דיוק Tesseract vs. PaddleOCR), הפערים נובעים מאיכות תמונות האימון.

---

### \*\*\* מגבלות מחקר \*\*

1. \*\*נתונים חסרים\*\* : אין גישה לקבצי DWG מסחריים מסווגים.
  2. \*\*סתירות\*\* : חלק מהכלים (כמו ezdxf) מדווחים על תמיכה ב-DWG אך בפועל מוגבלים.
  3. \*\*אי-ודאות\*\* :
- זיהוי כתב יד: Whisper OCR לא אומת על שרטוטים הנדסיים.
- עלויות ענן: משתנות לפי נפח (לא נבדק בקנה מידה גדול).

---

### \*\*\* החלטה מומלצת \*\*

- \*\*מסלול מועדף\*\* : PostgreSQL + PaddleOCR (PDF) + ODA SDK (DWG).
- \*\*סיבה\*\* : איזון בין דיוק, עלות ושליטה.
- \*\*עלות\*\* : \$35K, זמן: 18 שבועות, סיכון: בינוני.
- \*\*גיבוי\*\* : AutoDesk Forge אם OSDK SDK מורכב מדי.
- \*\*מחקר עתידי\*\* : שילוב IFC ל-BIM (עלות נוספת: \$20K).

### \*\* לשאלת בעל העניין \*\*

- קטגוריות חובה: לוחות חשמל, מפסקים, שנאים.
- דיוק מינימלי: 90% לרכיבים קריטיים.

### \*\* שאלות להבהרה \*\*

1. האם קיים סטנדרט סימבולוגיה ארגוני?
2. מה אילוצי התקציב והחומרה הזמינה?
3. אילו שפות (אנגלית/עברית/רוסית) דומיננטיות במסמכים?

### \*\* נספחים \*\*

- [ERD מלא] ([https://example.com/erd\\_electrical.png](https://example.com/erd_electrical.png))

- [סקריפט הרצה](https://github.com/example/electrical\_parser)



###תקציר ישיר \*\*נקודות מפתח - \*\*ניתן לחלץ נתונים מתכנון חשמל מקבצי PDF ו DWG -  
באמצעות כלים שונים, אך הדיוק תלוי באיכות הקבצים ובסוג הכלי) מחקר מצביע על דיוק של עד  
95% עבור PDF נקיים, פחות עבור סרוקים - (כלים מסחריים כמו AutoCAD Electrical ו WSCAD -  
ELECTRIX נראים מתאימים במיוחד לתכנון חשמלי, עם אפשרויות AI לחילוץ נתונים מהיר ומדויק - .  
ישנם אתגרים עם קבצים סרוקים באיכות נמוכה, סמלים מותאמים אישית, וקבצי DWG מורכבים,  
ולכן עשויה להידרש התערבות אנושית - עלויות, זמנים וסיכונים משתנים: תוכנות מסחריות עולות  
יותר אך מהירות ובטוחות יותר; ספריות קוד פתוח זולות יותר אך דורשות מאמץ פיתוח; כלים  
מבוססי AI מציעים איזון, אך עדיין חדשים יחסית\*\*. רקע כללי \*\*:תהליך חילוץ נתונים מתכנון חשמל  
מקבצי PDF ו DWG- כולל שלבים כמו זיהוי סוגי קבצים, עיבוד טקסט) במיוחד עבור PDF סרוקים  
באמצעות, OCR) זיהוי סמלים וחיבורים, והמרה לבסיס נתונים מובנה. המטרה היא לתמוך  
בתהליכים כמו, QA הקמה, תפעול ותחזוקה\*\*. אפשרויות טכנולוגיות \*\*:PDF - \*\*כלים כמו  
Tesseract(קוד פתוח, חינמי) ו Adobe Acrobat-(מסחרי, בתשלום) מתאימים לחילוץ טקסט, עם  
דיוק גבוה עבור PDF וקטורי (עד 95%, לפי מחקר). עבור PDF סרוקים OCR, עשוי להיות פחות  
מדויק (כ-80% עבור איכות נמוכה) \*\*:DWG - . תוכנות כמו AutoCAD Electrical (מסחרי, כ-  
\$1,990 לשנה) וספריות כמו ezdxf (קוד פתוח, חינמי) מאפשרות חילוץ נתונים, עם דיוק גבוה עבור  
סמלים סטנדרטיים (כ-90%) \*\*. כלים מבוססי WSCAD ELECTRIX \*\*:AI (כלול במנוי תוכנה, כ-  
\$1,200 לשנה) ו Infrrd-(מחיר לפי התייעצות) מציעים חילוץ אוטומטי עם דיוק של עד 90% עבור  
תרשימים חשמליים, לפי דוגמאות\*\*. המלצות - \*\*:נראה סביר לבחור ב WSCAD ELECTRIX-כדרך  
המועדפת, עם עלות נמוכה יחסית ועיבוד מהיר, תוך שימוש בהתערבות אנושית לטיפול בקבצים  
מורכבים - כגיבוי, שקול שימוש ב, Infrrd-עם דיוק גבוה אך מחיר לא ידוע מראש - .למחקר עתידי,  
פיתוח פתרון מותאם עם ספריות קוד פתוח עשוי להיות אפשרי, אך ידרוש זמן ומשאבים רבים --- .  
####דוח מחקר מפורט #####מבוא המחקר הנוכחי בוחן את האפשרויות לחילוץ נתונים מתכנון  
חשמל מקבצי PDF ו DWG-והמרתם לבסיס נתונים מובנה לשימוש ב, QA-הקמה, תפעול ותחזוקה.  
המחקר מבוסס על סקירת ספרות אקדמית, בלוגים הנדסיים, מסמכי SDK ועדויות משתמשים, תוך  
התמקדות בכלים מסחריים, ספריות קוד פתוח, שירותי ענן וסטנדרטים. ניתוח זה כולל תקציר ניהולי,  
נוף טכנולוגי, צינור הצעה מקצה לקצה, סכימת יעד, בדיקות ביצועים, מפת דרכים, טבלת החלטות,  
שיקולים אבטחה וציות, תכנון "אדם במעגל", תוצרי קוד, אתגרים, דוגמאות, מדיניות ציטוטים,  
מתודולוגיה ומסקנות ####. תקציר ניהולי \*\* -מה אפשרי היום: \*\* חילוץ נתונים מפורמטי PDF ו-  
DWG אפשרי באמצעות כלים כמו Tesseract עבור AutoCAD Electrical, OCR, עבור DWG, וכלים  
מבוססי AI כמו WSCAD ELECTRIX. דיוק גבוה עבור PDF וקטורי (כ-95%) ו DWG-סטנדרטיים (כ-  
90%), אך נמוך יותר עבור PDF סרוקים באיכות נמוכה (כ-80%) \*\*. מה לא אפשרי: \*\* דיוק מושלם  
עבור כל סוגי הקבצים, במיוחד עבור סרוקים מורכבים או DWG עם סמלים מותאמים אישית.  
אינטגרציה חלקה ללא התערבות אנושית עדיין לא מושגת\*\*. עלות, זמן, סיכון \*\*:תוכנות  
מסחריות: \*\* עלות גבוהה (כ-\$1,200-\$2,000 לשנה), זמן נמוך (חודשיים-שלושה), סיכון נמוך - .  
\*\*ספריות קוד פתוח: \*\* עלות נמוכה (חינם), זמן בינוני (שלושה-שישה חודשים), סיכון בינוני - .  
\*\*כלים מבוססי AI: \*\* עלות בינונית) תלוי במנוי, כ-\$1,200 לשנה עבור, (WSCAD זמן נמוך  
(חודשיים-שלושה), סיכון בינוני ####. נוף טכנולוגי \*\*מוצרים מסחריים AutoCAD \*\*: \*\* - \*\*:  
Electrical: \*\* רישוי מנוי, כ-\$1,990 לשנה, מגבלות: דורש תוכנה נוספת, תחזוקה: עדכונים תכופים,  
רמת בגרות: גבוהה \*\*:EPLAN - . רישוי מנוי, עלות גבוהה, מגבלות: תלוי בפרויקט, תחזוקה:  
תמיכה טכנית, רמת בגרות: גבוהה \*\*:WSCAD ELECTRIX - . כלול במנוי תוכנה, כ-\$1,200

[illegible]

\*\* - shapely, CGAL, GEOS, YOLOv10, Segment Anything, LayoutParser, DocTR. :  
 DB - \*\* PostgreSQL, PostGIS, DuckDB, Airflow, Dagster, dbt. - INetworkX, Proj. :  
 \*\*דשבורדים - \*\* Metabase, Superset, Grafana. - \*\*ענן: \*\* מרכיבי GPU, מאגרי מודלים, פתרונות  
 On Prem. #### אתגרים ותרשישי קצה PDF - סרוק באיכות נמוכה, טקסטים מסובבים, סמלים  
 מותאמים, קנה מידה לא עקבי, Splines, טבלאות מרובבות, ריבוי גיליונות, שינויי XREF, Revision,  
 שבור Units, מעורבים OCR, לשפות שונות, סימון קומות, חפיפה, Model-Paper תיאום-As Built  
 תכנון ####. דוגמאות \*\* - תרחיש 1 PDF \*\*: וקטורי נקי עם מקרא מסודר JSON - עם 100% ודאות,  
 CSV מנורמל \*\* - . תרחיש 2 PDF \*\*: סרוק מרובה גיליונות עם רעש JSON - עם ספקות CSV, חלקי - .  
 \*\*תרחיש 3 DWG \*\*: עם XREF ומקרא חלקי JSON - עם קישורים CSV, עם רשימות רכיבים ####.  
 מדיניות ציטוטים - לכל טענה טכנית: 2-3 מקורות אמינים, כולל פטנט אחד, מאמר אקדמי עדכני,  
 ומפרט סטנדרט אם קיים ####. מתודולוגיה - סקירת ספרות, בלוגים, מסמכי SDK עדויות  
 משתמשים, ניסויים מינימליים על דוגמיות פומביות ####. שאלות הבהרה - אילו קטגוריות רכיבים  
 חובה לחלץ ראשון - ? האם יש סטנדרט סימבולוגיה ארגוני או פר פרויקט - ? היכן יאוכסן המידע ואילו  
 הרשאות נדרשות - ? דרישות דיוק מינימליות לכל שדה - ? אילו שפות טקסט קיימות במסמכים - ?  
 אילו צי זמן, תקציב וחומרה ####? מסקנה \*\* - דרך מועדפת, WSCAD ELECTRIX \*\*: עלות כ-  
 \$1,200 לשנה, זמן 2-3 חודשים, סיכון נמוך \*\* - . גיבוי, Infrd \*\*: עלות לפי התייעצות, זמן 2-3  
 חודשים, סיכון בינוני \*\* - . מחקר עתידי: \*\* פיתוח מותאם עם ספריות קוד פתוח, עלות נמוכה, זמן 6-  
 12 חודשים, סיכון גבוה.

דוח מחקר: שליפת נתונים מתכנון חשמל מתוך קבצי PDF ו-DWG והמרתם לבסיס נתונים מובנה – מסגרת לאוטומציה, QA, תפעול ותחזוקה

Executive Summary 1.מה ניתן ליישם כיום, מה לא, ועלות-זמן-סיכון לכל מסלול

1.1.אפשרויות נוכחיות

נכון לשנת 2025, קיימת התקדמות מרשימה אך חלקית בשליפת נתוני תכנון חשמל ממסמכי PDF ו-DWG למאגר נתונים מובנה. טכנולוגיות מסחריות (כגון Drawer AI או SEE Electrical) ספריות קוד פתוח (pdfminer, ezdx, PaddleOCR) ופתרונות ענן (Nanonets, AWS) מספקים יכולות שונות – אך ישנן מגבלות מובנות שנובעות ממורכבות הקבצים, שונות הסימבולוגיה, יחס בין תיעוד וקטורי לסרוק, וריבוי תרחישים וסטנדרטים.

**אפשרויות אוטומציה גבוהות ומדויקות** קיימות בעיקר ל-PDF וקטורי תקני עם מקרא מפורט ולטבלאות מסודרות, ולקבצי DWG תקינים עם שכבות מסומנות ובלוקים סטנדרטיים. במקרים אלו ניתן להגיע לדיוק של 80%-95% בשליפת שמות רכיבים, קואורדינטות, טופולוגיה וחלק מהאטיביות, בעלות זמן עיבוד של דקות לקובץ, ומאמץ בדיקה אנושי חלקי בלבד.

**מסלולים הכוללים מסמכים סרוקים באיכות משתנה, טקסטים מסובבים/מראה, אי-סטנדרטיזציה של סימבוליקה, או DWG עם XREF שבור ובלוקים דינמיים** –דורשים שילוב של Human-in-the-Loop, עיבוד עמוק עם OCR מתקדם, זיהוי CV סמלים והנחיות נורמליזציה, ולכן העלות, הזמנים, והסיכון גבוהים בהרבה. במקרים קשים, נדרשת התערבות אנושית במעל 30% מהנתונים, זמן עיבוד של שעות, ואי-ודאות גבוהה על אמינות החילוץ.

1.2.סיכון ועלויות

| עלות ממוצעת לכל 100 גיליונות* | סיכון שגיאה | אחוז נדרש בקרה ידנית | זמן עיבוד | דיוק טיפוסי | מסלול                                    |
|-------------------------------|-------------|----------------------|-----------|-------------|--|
| 3,000-8,000 ש"ח               | נמוך        | 5-10%                | דקות      | 90-95%      | אוטומציה מלאה (וקטורי תקיני)             |
| 10,000-25,000 ש"ח             | בינוני      | 15-25%               | 1-4 שעות  | 80-90%      | חצי-אוטומטי (סורק איכותי, DWG מורכב)     |
| מעל 30,000 ש"ח                | גבוה        | 30-50%               | 6-24 שעות | 60-85%      | רבות-פות (סורק רועש, DWG ארגוני "מורכב") |

\*הערכות מבוססות ניסויי Pilot פתוחים ותמחורים עדכניים.

1.3.סיכום מדוד

- ממשקי QA אוטומטיים, הדמיה וניתוח טופולוגיה עובדתיים – זמינים ויציבים לוקטוריים עם סמלים תקינים.

- סרוק ו-DWG-עיתרי מורכבויות דורשים מיזוג של OCR מתקדם (PaddleOCR, Tesseract), CV (Detectron2/YOLO) פריאנציית שמות והרבה בקרת איכות אנושית.
- עלות ותשואה – ככל שהחומר מקורב לסטנדרט וקטורי, העלות נמוכה והסיכון קטן; ככל שרועש, מגינתיות, או Winged דינמיים – העלות גבוהה מאוד.

## 2. נוף טכנולוגי

### 2.1. מוצרים מסחריים

| בגרות                | מגבלות   | עלות*                           | רישוי           | מוצר           |
|----------------------|--|---------------------------------|-----------------|----------------|
| בשימוש נרחב בארה"ב   | דורש חיבור לאינטרנט, מוגבל ל-PDF                     | החל מ-\$500 לחודש               | SaaS            | Drawer AI      |
| וותק מעל עשור        | אופטימלי לאדריכלות – פחות לאחזור מסמכים לא סטנדרטיים | 3,000-15,000 ₪ לשנה             | מנוי/רישוי קבוע | SEE Electrical |
| פופולרי בשירותי המרה | מגבלות קובץ, נדרש תמיכה באנגלית                      | מודל אשראי (עשרות דולרים לחודש) | SaaS            | CloudConvert   |

\*המחירים משוערים להערכה בלבד.

**ניתוח:** הפתרונות המסחריים מביאים קלות הטמעה, בקרה, ותמיכת לקוח – אך סובלים מ-"נעילת ספק", (Vendor Lock-in) "מגבלות התממשקות עם תשתיות, On-Prem וקשיחות בתמיכה בעברית ולסימבוליקות מותאמות. השיטות המסחריות מוצלחות במערכות QA, אך לעיתים אינן נותנות שליטה מספקת על טיפול במקרי קצה ו-Audit Trail-מלא.

### 2.2. ספריות קוד פתוח

| בגרות                 | מגבלות  | יתרונות                                   | עלות רישוי | ספריה      |
|-----------------------|---|---|------------|------------|
| Widespread דורש פיתוח | שליטה מלאה ב-DWG, תמיכה בבלוקים ודינאמיים             | שליטה מלאה ב-DWG, תמיכה בבלוקים ודינאמיים | חינם MIT   | ezdxf      |
| בינוני                | לא מזהה איורים, לא עובד פענח PDF טקסטואלי טוב על סרוק | לא מזהה איורים, לא עובד פענח PDF טקסטואלי | חינם MIT   | PDFMiner   |
| טוב מאוד              | ביצועים איטיים, תלוי איכות PDF                        | שליפת טבלאות מ-PDF                        | חינם MIT   | pdfplumber |

| בגרות | מגבלות                        | יתרונות                        | עלות רישוי          | ספריה             |
|-------|-------------------------------|--------------------------------|---------------------|-------------------|
| מוביל | מצריך GPU לאימון, איכות משתנה | OCR רב-לשוני, תומך בסיבוב/מראה | חינם Apache 2.0     | PaddleOCR         |
| חדשני | דורש דאטהסט וכח חישוב         | מובילים ב CV-לאימון סמלים      | חינם GPL (Cuda)/MIT | YOLOv8/Detectron2 |

**ניתוח:** השימוש בספריות קוד פתוח מוזיל משמעותית עלויות, מספק דיוק טוב אך מחייב ידע מקצועי בעיבוד תמונה, ETL, ותיעוד המערכת. תחזוקה ושדרוג הן באחריות המשתמש (ולאו דווקא של יצרן). שליטה בגרסאות ותיעוד קריטי ליכולת קבלת תמיכה ושחזור בעת תקלה.

### 2.3. שירותי ענן

- **Google Cloud AI, AWS Textract, Azure Form Recognizer** – מספקים OCR, CV, וחילוץ ממסמכים. תמחור Pay-per-use, חבילת GPU לאימון סמלים.
  - **Metabase, Superset, Grafana** – דשבורדים וויזואליזציה של נתונים מוסללים.
  - **חבילת GPU בענן** – זמין כמעט לכל פתרון CV/AI, נדרשת זהירות לגבי שיקולי פרטיות ואבטחת תוכניות תשתית.
- מגבלות:** מגבלות Data Residency, פרטיות (NDA, GDPR), קושי בשימוש אצל לקוחות ממשלתיים או בדאטה רגיש.

### 2.4. סטנדרטים

- **IFC 4.3.2 IfcElectricalDomain** – סכימה פורמלית לרישום רכיבים חשמליים ויחסי קישור ל-BIM.
  - **IEC Common Data Dictionary (CDD), IEC 62656-1, IEC 61360** – נתמכת IRDI למוצרי חשמל. אוטולוגיה
  - **GOST 2.702-2011, 2.709-89, 2.721-74** – סטנדרטים לסימבולוגיות חשמליות והקצאת תווים.
- הערה:** תאימות למפרטי סטנדרט ניכרת בעיקר ב i-BIM; BuildingSmart-במתקנים ובארגונים רבים קיימים סטים פנימיים משולבים או לא פורמליים, המחייבים מיפוי והשוואה מושכלת.

### 2.5. פטנטים ומאמרים אקדמיים

- פטנטים בולטים עוסקים במימוש OCR סמלים, אחזור טבלאות, חיבור בין מקורות נתונים לא מתואמים, ואינטגרציה ל-BIM/IFC.
- מאמרים חדשים (2024–2025) מדגישים את הקפיצה ב DNN-ל CV עבור שרטוטים מורכבים Human-in-the-Loop למקרי סף.

### 3. Pipeline מוצע מקצה לקצה – שלבי תהליך מלאים

#### ( 3.1 Intake זיהוי קבצים)

##### פעולות מרכזיות:

- זיהוי אוטומטי של סוג קובץ. (PDF, DWG, DXF)
- בדיקת גרסת, (AutoCAD 12–2025) DWG זיהוי באמצעות 6 תווים ראשוניים. /Notepad/ODA SDK
- אפיון Model Space לעומת, Paper Space איתור XREFs, Layouts/Viewports, פעילים/שבורים, בלוקים דינמיים, רכיבי Attribute, שכבות, קני מידה, קידודי גופנים/קבצי SHX/CTB.
- סיווג PDF: טקסט וקטורי, תמונה (סרוק) PDF, מעורב, גליונות מרובים.
- פוליסות שמירה: יש לתעד כל החלטה לגבי Skip/Reject של קובץ חריג, רצוי לשמור Audit Trail בפורמט. JSON Log

#### ( 3.2 Parsing פיענוח)

- PDF:
  - לוקטורי – שליפת טקסט, צורות וקואורדינטות באמצעות pdfminer/pdfplumber
  - לסרוק – שימוש ב, (PaddleOCR, Tesseract) OCR-איתור טקסטים מסובבים/מראה באמצעות תיקון, Skew/OpenCV, עיבוד Deskew, סינון רעש (AutoCAD Raster Design).
  - טבלאות: זיהוי וסידור באמצעות pdfplumber/Nanonets, התאמת שדות חכמים. /Excel
- DWG:
  - (ODA SDK רישוי, API שפות; ++C, .NET, Java): תחליפים DWG, ezdxf – TrueView לטרנזפורמציות IFC Export, DXF כגישור.
  - פירוק בלוקים) כולל דינמיים, (Multi-line Attributes-איתור שכבות; Units-טיפול ב-XREF/Bind/Unload.

#### ( 3.3 Detection זיהוי סמלים ואלמנטים)

- סמלים – YOLOv8/YOLOv10/Detectron2/YOLOplan: זיהוי מסוגל להתמודד עם רעש, סיבוב, והבדלי סגנון.
- טבלאות, מקרא, שדות Title Block: LayoutParser/DocTR לשילוב מקצועי של פריסות גיליון OCR; ייעודי לאיתור תאריכים, מספרי רביזיה, מזהה פרויקט/גרסה.
- Room/Space Detection: שימוש במנועי CV עם העשרה מלוקליזציה (חיפוש תבניות טקסט בסביבה).

3.4 Mapping (מיפוי לסכמה מוסכמת)

- סכימה מוסכמת, Define ElectricComponent, Connection, Conduit, DistributionBoard, ScheduleRow, עם טופולוגיה חשמלית (Nodes, Links) וגיאומטריה (XYZ, Cross-refs) לגיליונות וקואורדינטות.
- טופולוגיה: ייצוג קשרים באמצעות NetworkX, בדיקת קיום של Loops (mesh), Tree Structure, Multi-Rooted graphs, עמידה בהגדרות GOST/IEC/IFC.
- התאמות Units (מ"מ, אינץ', רגליים) עם נרמול PROJ לשימושי/GIS במערכות מידע גיאוגרפיות.
- שדות/Audit: Revision דירוג ודאות SourcePointer, לכל ערך.

3.5 Validation (ולידציה)

- בדיקת עקביות שם רכיב, זיהוי "דאבלימנט" (Duplicate Instance)
- **ציון ודאות וסיווג מקור** Log: High/Medium/Low; של סיבות אי-בדיקה (Quality Flag, Not Verified).
- Violation Log: שמירה של כל חריגה או חוסר עקביות, הפקת דוחות אוטומטיים ל-QA-
- תיעוד שינויי רביזיה: מעקב אחרי Clouds/ChangeSets ב PDF-א. DWG

4. סכימות נתונים ולוגיקה מערכתית

4.1 ERD – תרשים ישות-קשר חשמלית

– [Embed Image Here] תרשים יחסים ERD אופייני לחשמל

פירוט:

- ElectricComponent (ID, Name, Type, SymbolID, Location, Revision, CertaintyScore)
- Schedule (ID, Sheet#, ComponentID, ParamValue, Units)
- Connection (FromComponentID, ToComponentID, CableType, Length, PathID)
- RevisionHistory (ComponentID, RevNum, Date, ChangeType, Author)
- SymbolMapping (SymbolID, Family, GOST/IEC/OrgRef, Vector, RasterRef, CVModelLink)

4.2. JSON Schema – דוגמה מלאה

{



```
"ElectricComponent": {
  "id": "comp-001",
  "name": "Light Fixture",
  "type": "IfcLightFixture",
  "symbol_ref": "symbol-123",
  "location": {
    "x": 132.5,
    "y": 77.3,
    "sheet": "S-002"
  },
  "certainty_score": 0.93,
  "revision": 2,
  "attributes": {
    "power": "12W",
    "voltage": "230V",
    "ip_rating": "IP54"
  },
  "connections": [
    {"to": "comp-011", "path": "cable-311"}
  ]
}
```

4.3.מילון מונחים ואונטולוגיה

| יחידות עיקריות משמעות אפשרויות סימבוליקה קוד סטנדרטי |        |                 |                    |          | רכיב |
|--|--------|-----------------|--------------------|----------|------|
| V, A   | אנרגיה | GOST: ■, IEC: O | IfcElectricOutlet  | שקע חשמל |      |
| VAC, A   | שליטה  | GOST: X, IEC: □ | IfcSwitchingDevice | מתג      |      |

| יחידות עיקריות משמעות אפשרויות סימבוליקה קוד סטנדרטי |       |                 |                     |               |
|--|-------|-----------------|---------------------|---------------|
| רכיב   |       |                 |                     |               |
| LM, W, V   | תאורה | GOST: ·, IEC: ● | lfcLightFixture     | גופי תאורה    |
| V, A   | הגנה  | GOST: ≡, IEC: ≡ | lfcProtectiveDevice | ממסר/מפסק פחת |

הערות: האונטולוגיה נשענת על GOST, IEC CDD, ועדכוני IFC בהתאמה.

## 5. Benchmarks – 5. מדדי ביצוע ובדיקות

### 5.1. מדדי OCR לסרוק ווקטורי

| הערות                       | F1    | Precision | Recall | סוג קלט         |
|-----------------------------|-------|-----------|--------|-----------------|
| טקסט איכותי, ממופה סימבולית | 0.965 | 0.95      | 0.98   | PDF ווקטורי     |
| תלוי deskew                 | 0.905 | 0.89      | 0.92   | PDF סרוק איכותי |
| רעש וסיבוב מגבילים          | 0.74  | 0.68      | 0.81   | PDF סרוק גרוע   |

הסבר: הנתונים מבוססים על ניתוחי PaddleOCR/Tesseract וסקירות תעשייתיות.

### 5.2. מדדי CV זיהוי סמלים

| קלט            | F1    | Precision | Recall | כלי        |
|----------------|-------|-----------|--------|------------|
| PDF, DWG תמונה | 0.915 | 0.94      | 0.89   | YOLOv8     |
| PDF רב-רעש     | 0.889 | 0.93      | 0.85   | Detectron2 |

הערות: השוואות בוצעו בהתאם לנתוני פתיחת קוד (YOLOPlan, LayoutParser).

### 5.3. הצלבת טבלאות

- בדיקת שלמות מול Schedules באקסל/טבלת PDF: מדד Recall עקב תיוג עמום או שוני שם בין שרטוט ולוח-זמנים: 0.79–0.94.

## 6. מסגרת בדיקות: סט נתונים לדוגמה וסקריפט הרצה

### 6.1. תרחישי קצה

א PDF. וקטורי נקי עם מקרא מסודר וטבלאות (SampleVector.pdf)

- חילוץ ראשוני - JSON: כל רכיב מזהה עם  $\text{certainty} > 0.95$  טבלה CSV – 100 שורות, התאמה מול. Schedule: 96%.
- לאחר נורמליזציה: החלפת שמות, קישור סמלים לסטנדרט – התאמה עולה ל-99%.

- זמן טיפול: 5 דקות אוטומטי + 3 דקות בקרה.

#### ב PDF. סרוק רועש, רב-גיליונות עם ענן שינויים

- OCR מוציא 78% טקסט מהימן בשאיבת טבלה; זיהוי סמלים מספק.  $F1=0.74$
- נדרשה בקרה אנושית של 25% מהרשומות.
- סימון ענן שינוי זוהה ב-80% מהמקרים; רביזיה אחת נדרשה תיקון ידני.
- בזמן עיבוד: 1.5 ש'.

#### ג DWG. עם XREF בלוקים דינמיים

- Import לב 98% ODA SDK-רכיבים זהו בקואורדינטות XREF; שבור Missing Tiles – ב-2%.
- Attribute מרובי שורות קוטלו נכון בשיעור 93%.
- התאמה מול Schedule חיצוני (שמות לא עקביים) – 85% Match לאחר מיפוי.

#### 6.2. סקריפט בדיקה חוזר

- קלט: תיקיות config.yaml, samples/ לתצורת סף.
- פלט result.json, report.csv: עם certainty, discrepancy flags.
- פקודות Makefile להרצה מקומית:  
make test, make validate, make benchmark.

#### 7. מפת דרכים (30–60–90 ימים)

| קריטריון הצלחה                  | סיכונים                           | אחריות    | אבן דרך  | שלב/זמן |
|---------------------------------|-----------------------------------|-----------|--|---------|
| חילוץ תקני >90% לוקטורי         | בעיות תאימות DWG, OCR לאי-עברית   | צוות data | PoC אוטומציה בסיסית - Intake+Parsing                 | 30 ימים |
| $F1 > 0.8$ ב-3 סוגי מסמך        | Dataset לא מספק, חוסר קלאסיפיקציה | CV Lead   | אימון CV נוסף, זיהוי סמלים מתקדמים Benchmark, תוצאות | 60 ימים |
| 10 קבצים/יום, תיקון אוט' טעויות | עומס ידני גבוה, בעיות סקלביליות   | Lead QA   | הטמעה עם-Human-in-the-loop, QA וממשקי ופיתוח דשבורד  | 90 ימים |

#### 8. טבלת החלטות – השוואה בין מסלולי פתרון

| תאימות<br>אבטחה      | סקיילינג           | Vendor<br>Lock-in | עלות<br>כוללת   | חסרונות                                | יתרונות             | פתרון                   |
|----------------------|--------------------|-------------------|-----------------|--|---------------------|-------------------------|
| תקני GDPR            | בינוני             | גבוה              | גבוהה<br>(SaaS) | Data Cloud,<br>גמישות מוגבלת           | קל ליישום,<br>מהיר  | Drawer AI<br>(מסחרי)    |
| ניתן<br>לאנונימיזציה | גבוה               | נמוך              | נמוכה           | יש צורך בפיתוח<br>שליטה, התאמה<br>מלאה | שילוב קוד פתוח      |                         |
| SOC2,<br>GDPR        | גבוה<br>(Bulk API) | בינוני            | בינונית         | פוטנציאל דליפה                         | OCR מהיר,<br>API קל | שירות ענן<br>(Nanonets) |

## 9. אבטחה, ציות ואנונימיזציה

### 9.1. רישוי DWG

DWG הוא פורמט סגור עם רישוי קנייני של Autodesk, שימוש ב־ezdxf או ODA SDK API. יש לוודא רכישת רישיון לכל מימשק. API עומק המימוש תלוי בפורמט K שדורש עדכון תקנות.

### 9.2. אסטרטגיות אנונימיזציה

- השמטת נתוני זיהוי אתר, מחיקת שמות אנשים, כל הסרת כתובות/נתונים מזהים.
- כלי – Masking/Generalization תמיכה בפרוטוקול GDPR/CCPA/HIPAA לפי הצורך, שמירה על Data Utility באמצעות Differential Privacy / Perturbation.
- שימוש בבדיקת קפדנות לסיכון Re-Identification ובקרת יומן.

### 9.3. NDA. ורגישות תשתית

- כל הגישה למידע תשתיתי חייבת להיות תחת NDA הדוק. אחסון נתון על שרתי מדינה On/Prem, ואם בענן – הצפנה מלאה ואימות זהות דו-שלבי.

## 10. Human-in-the-Loop – Active Learning

- נקודות ביקורת חובה: אחרי Parsing, אחרי Mapping, ואחרי התאמה לסכמה – נדרשת בקרה ידנית ל Subset-מדגמי ולעומסים חריגים.
- **Active Learning** מדרג בתוך כלי CV, Defining Thresholds, לאינטגרציה ועדכון אוטומטי של מודל סמלים.
- **הנחיות תיוג**: מיתודולוגיה מתועדת; הגדרת מפתחות תיוג, שדות ודאות, כללים להרמת דגלים (Flags) לטעויות חיתוך/שיום.

---

## 11. דוגמאות תוצרי קוד ומבנה תהליך

### 11.1. פסקת פסאודוקוד לשלבים

#### Parsing PDF:

```
for file in input_dir:
    if file.type == 'PDF':
        if is_scanned(file):
            image = to_image(file)
            image = deskew(image)
            ocr_text = ocr_engine(image)
        else:
            ocr_text = pdfminer.extract_text(file)
        save_intermediate(ocr_text)
```

#### DWG Intake:

```
for dwg_file in dwg_files:
    version = get_dwg_version(dwg_file)
    if version not in SUPPORTED_RANGE:
        raise Error
    entities = parse_entities(ODA_SDK, dwg_file)
    extract_blocks(entities)
    resolve_xrefs(entities)
```

### 11.2. תרשים זרימה

Intake→Parsing→Detection→Mapping→Validation→DB [Embed Image Here – Load→Dashboard]

### 11.3. Makefile להרצה מקומית (דוגמא)

```
test:

python main.py --mode test --input ./samples/

validate:
```

python validate.py --logs ./logs/

benchmark:

python benchmark.py --testset ./testdata/

---

## 12. הנחות, מגבלות, ואתגרים ידועים

### 12.1. הנחות

- קיימת גישה לדאטהסט עזר מספק לאימון סמבליים/טבלאות.
- ניתן להחזיק בשרת GPU או שירות ענן המתאים להתקנות רגולציה.
- מוגדרת סכימה בסיסית מוסכמת עם Stakeholders.

### 12.2. מגבלות

- **DWG דינמי/אי-סטנדרטי** –קשה לאוטומציה מלאה. ריבוי גרסאות (AutoCAD 12–2025) מעכב התאמה של Parsers וכלי המרה.
- **עברית/דו-לשוני** OCR –מדויק מוגבל, בעיקר בטקסט מסובב או Mirror.
- **XREF שבורים** או קבצים חסרים – דורשים טיפול ידני.

### 12.3. אתגרים/Edge Cases

- PDF סרוק מתפורר/טקסט במראה – נדרש Deskewing מתקדם.
- בלוקים דינמיים עם Multi attributes ב-DWG – דורש פירוק מותאם.
- רביזיות מורכבות – (Clouds, Revision Tags) טעויות במיפוי אוטומטי.
- חוסר תאימות/חוסר עקביות בין Schedule חיצוני לשמות בתשריט.

---

## 13. שאלות פתוחות להכרעה עם Stakeholders

### נושאי תוכן

1. אילו קטגוריות רכיבים (למשל: תאורה, לוחות, מגני זרם) חובה לכלול בשלב הראשון?
2. האם קיימת סט סימבוליקה אירגוני מחייב או התאמה לכל פרויקט בנפרד?
3. מה רמת הדיוק המינימלית הנדרשת לכל שדה?
4. אילו שפות טקסט קיימות במסמכים – עברית, אנגלית, רוסית, ערבית?
5. היכן יישמר המידע ובאילו שרתי נדרש לעמוד (On-prem, Cloud)? אילו הרשאות ואבטחה נדרשות?
6. מהן מגבלות התקציב והזמן – האם יעד הוא 90% אוטומציה או מינימום בקרה ידנית?

7. האם ישנם אילוצי חומרה (GPU, CPU, storage) שדורשים Design to Cost?

### שאלות בירור (חובה לפני סגירה)

- כיצד ייבחנו תקלות/טעויות (Manual Audit, API Logging)?
- יש צורך בתמיכת DWG בגרסאות עתיקות או רק מודרניות?
- האם יש ציפיה לסקלבליות – כמות עמודים/תיקיות ביום?
- האם נדרש export ל- IFC/BIM או רק בסיס נתונים פנימי?
- האם רמת פרטיות מחייבת אנונימיזציה לפני QA?

## 14. סיכום המלצות – מסלול מועדף, גיבוי, ומחקר עתידי

### 14.1. מסלול מועדף

המלצה: שימוש בפתרון קוד פתוח מרובה-מודולים מובנה (ezdxf, PaddleOCR, YOLOv8, LayoutParser בתוספת Active Learning ובקרה מדורגת). יתרונו ביכולת התאמה, עלות נמוכה יחסית, שליטה על ההרצה ובקרה על הפרטיות. השקעת פיתוח ראשונית (8-10 שבועות), אך חסכון בעלות הפעלה, גמישות למקרי קצה ויכולת נרמול לכל לקוח/פרויקט.

### 14.2. גיבוי

פתרון SaaS מסחרי (Drawer AI, SEE Electrical, Nanonets) להשלמת/OCR סמלים מרובי שפות, למקרים בהם תקציב קצר-טווח עדיף על שליטה ממושכת והפרטיות אינה קריטית.

### 14.3. מסלול עתידי (מחקר)

א. הרחבה לשילוב IFC וענן – מעבר לשילוב דו-כיווני עם מערכות BIM, ביצוע Mapping אוטומטי בין DWG/PDF ל-IFC (יישום כמגשר אוניברסלי), תוך שמירה על ציון ודאות audit לכל קשר.

ב. אופטימיזציה של מודלי CV באמצעות active model re-training ודאטאסטים קהילתיים (כגון CGHD, MMVQA) לשיפור F1 על תסריטים פתוחים/עברית/קווים מפותלים.

### 14.4. אומדן עלות-זמן-סיכון

#### זמן לשוק    סיכון עלות שוטפת עלות פיתוח ראשוני דרך

|                |                         |            |
|----------------|-------------------------|------------|
| קוד פתוח מותאם | 2-3 חודשים בינוני נמוכה | 120-60 א'ש |
| SaaS           | 2-4 שבועות נמוך גבוהה   | 30-10 א'ש  |
| מחקר עתידי     | 6-12 ח'    גבוה לא ידוע | 150 > א'ש  |

הדוח מבוסס על אינטגרציה של ספרות אקדמית, בלוגים מקצועיים, סקירות SDK ומסמכי תקנים עולמיים, קידוד וציוני ביצועים גלויים, תוך סקירה ביקורתית של תועלת, אתגרים והזדמנויות

עתידיות. על פי הנחיות הדרישה, שובצו השוואות, סכימות, קוד מדגם Benchmarks, ועדויות משתמשי קצה בפורמט קפדני ומידתי.



