

מחקר מבוסס ראיות: שליפת נתונים מתכון شامل מ- PDF ו-DWG והמרתם לבסיס נתונים מבנה

דרישה: להיות קונקרטי, ביקורתי, ולצין במפורש היכן יש ספק או חוסר מידע. לכל טענה מוצמדים מקורות. כשם קורות חלקים מציגים את שני הצדדים. אין הסתמכות על ניסויים פנימיים כאן לא בוצעו ניסוי OCR או CV כמותיים במהלך המענה והיכן שנדרשו מדדים מצויינית תכנית מדידה ולא תוצאות.

1. Executive Summary

מה אפשרי היום

- קריאה סיסטמית של PDF וקטורי:** חילוץ טקסט, פסקאות, טבלאות ומיקום גאומטרי של תווים אפשרי באמצעות [six.pdfminer](#).pdfminer.six עם רישוי MIT. מוגבלה: ביצועים נמוכים יחסית וモרכבות בעיבוד גרפייה וקטוריית מורכבת. ([pdfminer.six](#), [GitHub](#), [PyPI](#))
- OCR OCR סרוק:** מנועים עדכניים כמו PaddleOCR docTR או מספקים זיהוי רב לשוני וביצועים טובים יותר מטsparket בסביבות מסוימות אך התוצאות תלויות מאוד באיכות הסריקה והסתה הארגוני. יש חוסר קונצנזוס פתוח לגבי יתרון עיקרי של מנוע אחד בכל תחום תכון, [paddlepaddle.github.io](#), [mindee.github.io](#), [digitalcommons.usm.maine.edu](#), [Medium](#))
- גישה ל-DWG:** הפתרונות המובילים תעשייתית הם Autodesk ODA Drawings SDK או RealDWG. ODA RealDWG מספק SDK רחב עם מודול רישוי חברות RealDWG. ODA לקריאה וכתיבה "מהמקור". שני הנתיבים בתשלום ובעלי תנאי רישוי שונים ezdxf. מספק קריאה וכתיבה לDXF בלבד לא ([opendesign.com](#), [Autodesk Platform Services](#), [Autodesk](#), [ezdxf.readthedocs.io](#))
- סטנדרטיזציה מבנית:** שימוש IFC בקשר אפשרי לחלק מהישיות החשמליות אך הritisoi של פרט סכומות סימבוליות ושדות תפעוליים מוגבל ויהיו פורי מייפוי ([buildingSMART](#), [Technical](#), [standards.buildingsmart.org](#), [ifc43-docs.standards.buildingsmart.org](#))
- זיהוי סמלים קריין טופולוגיה:** מודלים כללים Detectron2 YOLOv8 YOLOv10 (Detectron2) ו-SAM מספקים זיהוי אובייקטים/סגמנטציה ברמת State of the Art אך לא קיימת מערכת נתונים ציבורית סטנדרטית לשרטוטי חשמל המאפשרת מסקנות ביצועים חד משמעיות. נדרש אנטציה ייעודית. ([detectron2.readthedocs.io](#), [Ultralytics Docs](#), [arXiv](#)) .

מה לא אפשרי היום ללא השקעה ייעודית

- אחסון טופולוגיה חשמלית מלאה "מהkopfse":** אין מוצר מדף שממיר באופן אמין DWG/PDF של חשמל לטופולוגיה חיבורים ומספר מעגלים כולל ודאי. יידרש צינור מותאם עם מודלים מאומנים-בארגן HITL. ראיות: פערים בסטנדרטים מגבלות OCR ו-Layout העדר קורפוס פתוח([Frontiers](#), [MDPI](#))

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

זמן עלות סיכון בשלושה מסלולים

- **מסלול + ODA SDK A צינור קוד פתוח ל+ PDF/OCR מודל סמלים ייעודי**
זמן 10 שבועות ל 6 POC חדשים להקשה. עלות ישירה מנוי + ODA כ 300,000 דולר/חדש GPU בענן לפרק אימון קצרים תלוי בענן/כרטיס. סיכון בגין תלות בא ODA ובמודול CV מותאמים([opendesign.com](#), [Vantage](#), [Google Cloud](#)) .
- **מסלול + RealDWG B שירותינו ענן מנהליםOCR**
זמן 12 שבועות ל POC. עלות רישיון RealDWG לפי התקשרות עם Autodesk לא מפורסם פומביות דרוש פנינה. סיכון משפטי נמוך בהיבטי תאימות DWG אך נעילת ספק גבוהה יותר . ([Autodesk Platform Services](#), [Autodesk](#))
- **מסלול DXF C בלבד עפודה ezdxf + OCR**
זמן קצר ל POC אך מוגבל לשעון ממיר DWG TrueView עם דוגמה עם DWG שלב פרה פלייט. סיכון גבוה לאבד פרמטרים וסמנטים "יחודיים" לדגש DWG ולבlokים דינמיים . ([ezdxf.readthedocs.io](#), [Autodesk](#))

פסק דין תמצית: למסגרת הנדסית אמינה מסלול A הוא המועד מסלול B לגבי תוכן ומסלול מחקר עתידי ליצירת קורפו אנט齊יה והערכת ביצועים סטנדרטיבית.

2.טכנולוגיות Landscape

DWG

- **OLA Drawings SDK**agineה ל DGN/DWG רישיונות חברות עלויות לפי מדרגות. יתרונות בגנות גבואה ותמכה בבלוקים דינמיים/XREF/শকبات. מגבלות מנוי בתשלום קוד סגור. תחזקה נדרשת הידוק גרסאות. בגרות תעשייתית([opendesign.com](#)) .
- **Autodesk RealDWG SDK** לקריאה/ כתיבה DWG/DXF מהמקור "רישיון" דרך Autodesk. בגרות תאימות מקסימלית ל DWG מגבלות תלות חוזית פרטיה מחיר לא גבוהים. בגרות תעשייתית([Autodesk Platform Services](#), [Autodesk](#)) .
- **ezdxf**קוד פתוח ל DXF לא תומך ב DWG רמת בגרות גבואה ל . ([ezdxf.readthedocs.io](#), [PyPI](#))
- **DWG TrueView** צפייה/המרה DWG לארסאות אחרות כל פרה פלייט חינמי. מגבלה ללא SDK. ([Autodesk](#))
- **סוגיות IP ורישיון DWG** סימני המסחר והתוכן בבעלות Autodesk והлицנסורים שלה שימוש כפוף לרישיון. מסקנה לעיבוד DWG מסחרי יש להסתמך על SDK מורשה([Autodesk](#)) .

PDF

- **pdfminer.six**pdfMinerpdfMinerMIT (pdfminer.six, PyPI)pdfMinerpdfMinerMIT (GitHub)
- **PyMuPDF/MuPDF**AGPL (PyMuPDF, PyPI)PyMuPDF/MuPDF
- **PDFium**BSD (GitHub, Hacker News, Nutrient)PDFium

Layout&OCR

- **Tesseract**חופשי ותיק מגבלות שימושיות על איזות נמוכה וטקסטים הנדסיים. (digitalcommons.usm.maine.edu)
- **PaddleOCR**מודרני רב לשוני פרקטיקות פרישה קלות עדינות שטח לשיפור על פני חלופות אך לא סט ציון ייחוס אחד(paddlepaddle.github.io, Medium)
- **odooocr**OCR عمוקה Docker ותמיית GPU רישוי Apache 2.0. (GitHub, mindee.com)
- **LayoutParser**זיהוי מבנה מסמך משלב Detectron2/ocr מותאים לטפסים וטבלאות. (layout-parser.github.io, GitHub)

CVלזיהו סמלים וקווים

- **Detectron2**Apache 2.0 (detectron2.readthedocs.io, GitHub)Object Detection/Segmentation פלטפורמה ל.
- **YOLOv8**מודרני ביצועים בזמן אמת תיעוד נרחב(Ultralytics Docs)
- **YOLOv10**עובדת מחקרית NMS-free training שיפורים בלטנס ויעילות. רמת בגרות גבוהה סביבת קהילה לא "ሞוצר". (arXiv)
- **Segment Anything (SAM)**יכולה zero-shot לא תמיד יציב בסביבות תעשייתיות. (arXiv)

גרפיקה וטופולוגיה

- **Shapely/GEOS**טופולוגיה חישובי פוליגונים קווים תוארי.
 - **CGAL / GEOS**אלגוריתמיקה גאומטרית.
 - **NetworkX**גרפים לחיבורים חשמליים.
 - **Proj**טרנספורמציות קווארדינטות.
- (מקורות ייודים לכל ספרייה קיימים לא צורפו כאן בשל מגבלות אורן אך כוללים בהמלצות כלים מוכרים בסביבת Python)

ETL&DB

- **PostGIS + PostgreSQL** תמיכה גאותרת מתקדמת (PostGIS).
- **DuckDB OLAP** Prototype (DuckDB) בתחילת דרכו.
- **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** מודול ציבורי סביר ~\$1.006/large ~\$1.624/large עד 4XLarge.
- **GCP** עמוד תומך רשמי ל-A100/A100L4/H100 Machine Type ל-פוי אימוט בתחריב בפועל (Vantage).
- **Cloud** סקרי שוק בלתי רשמי מציגים פערים משמעותיים בין ספקים וירידות מחירים אין לראות בהם התחריבות (DataCrunch, Thunder Compute).

3. מוצע מקצה לenza Pipeline

Intake

- **דיאוי סוג קובץ וגרסאות:** חתימות PDF מול DWG בדיקת גרסה XREF איתור CTB/ODA/RealDWG APIs לבולוקים דינמיים Attributes ושכבות ;
- **מגבלה DXF בלבד:** משפטית לשימוש SDK מושתת (ezdxf). (opendesign.com, Autodesk Platform Services, ezdxf.readthedocs.io, Autodesk)
- **Model vs Paper Space Viewports Clips:** קריאה מ-Layouts ו-Viewports ב-DWG דרך SDK; נדרש שיחזור לוגי של אזורי גילון. **סיכון:** קליפים מסתירים פרטם יטופל ביוםן זהירות (sourcecad.com).
- **Preflight:** המרת DWG לארטfactות נתמכות ובדיקה תלויות עם DWG TrueView. ([Autodesk](#))

Parsing

- **PDF אקטורי sixpdfplumber:** מיקום טקסט KlipTables. **מגבלה:** איטיות יחסית . ([pdfminer.six](#), [GitHub](#))
- **PDF OCR PaddleOCR:** עם docTR או PaddleOCR טקסט מסובב/מראה עם פינה עיבוד. **אנו** וDAOOT: שפות רבות פונטיים טכניים אינטלקט נומקה יוריד. Recall. ([paddlepaddle.github.io](#), [mindee.github.io](#))

- **Block References Dynamic Params Attributes**: ODA
שכבות וקווארדינטות. **חלופה**: יצא LF DXF ואז ezdxf אך קיימים אובדן מידע .
(opendesign.com, Autodesk Platform Services, ezdx.readthedocs.io)
- **IFC** אגש: בחלק מהמרקם יתכן מיפוי IFC IfcElectricalDomain אך לא מכסה מקרה וشرطוטן .
לכן רק עזר-ifc43-docs.standards.buildingsmart.org

Detection

- **סמלים**: אימון 10/YOLOv8 או Detectron2 על קורפוס ארגוני עם נתוני סינטזה מסימבולוגיה כולל (IEC 60617 IEC Augmentations לסייע/קנה מידע). פטנטים תעשייתיים מציגים גישות דומות לייצור דאטה סינטטי לSIMBULOGIC הנדסיות. **הבדלי דעות**: יש מאמרם שמציגים יתרון לשילוב DL Heuristics עם DL בטכני מול גישות DL טהורות, (Google Patents, Ultraalytics Docs, arXiv, MDPI)
- **קוויים וחיבורים**: Skeletonization + I: Vectorization כללים לטופולוגיה שימוש שוננות NetworkX/GEOS Shapely/Ishapely/Splines: וקשנות עובי קוו משנתה שכבות שונות.
- **הסמלים Room/Space**: אם קיימים חדרים בIFC/DWG לקשר לפי נקודת מרכז Bounding Box של Room/Space: הסמל עם טולרנס Z. **אי ודאות**: כאשר אין גבולות חדר נדרשת האנשה.

Mapping

- **סכמה מוסכמת**: Device Connection Room DrawingSheet: קווארדינטות world Schedules מקומיות לגילוין יחידות עם המפה. **LINKS** בין שרטוטים בין מקרה לשימוש ובין חיצוניים.
- **נורמליזציה שלשמות וסוגים**: מילון מונחים ארגוני מיפוי IEC 60617 למונחי הארגון. **מגבלה** IEC 60617: הוא מאגר בתשלום (IEC Webstore).

Validation

- **בדיקות עקבות**: סמל חייב להיות על שכבה תואמת חיבור סגור לארף התאמת חדר וצלבים מול Schedule.
- **ציון ודאות**: הרכיב של detection score מרחק לחדר התאמת שם למקרה ועקבות טופולוגית.
- **Audit Log**: כל חריגה נרשמת עם קוד סיבה.

4. סכימת יעד JSON Schema/ERD

קובצי 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 ודותות:
 - [sample_pre_normalization.csv](#) •
 - [sample_post_normalization.csv](#) •

: DrawingSheet 1 n ElectricalElement; ElectricalElement 1 n ERD (from_element_id to_element_id); RoomSpace 1 n ElectricalConnection דרך קצחות ElectricalConnection אופציונלי כ קישור. ElectricalElement. IFCRoom

5. מיון מונחים אונטולוגיה 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 סמליים**: מיפוי פנימי 60617 EC אבקורלציה נדרש מנוי ל DB IEC לצפיה בציורים
 - **רשומים. סיכון**: שנות ארגונית גובה HITechובה ([IEC Webstore](#)) .
-

6. והערכת ביצועים תכנית Benchmarks

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

מדד OCR

- **ל PDF סרוק** Precision/Recall/F1: על מילים ושורות חלקה לפי איקות סריקה שפה סיבוב. השוואת. Paddle/OCR docTR Tesseract. עדיפות בסקרים קשים אך תלויים בתצורה. (. [paddlepaddle.github.io](#), [mindee.github.io](#), [digitalcommons.usm.maine.edu](#))
- **ל PDF וקטורי**: דיווק קריאת טקסט וטבלאות מול אמת מידת ידנית עם pdfminer.six/pdfplumber. ([pdfminer.six](#), [GitHub](#))

דיהוי סמלים

- הפרדה בין PDFækטורי ל PDFסורך. מודל YOLOv8/10 מודול Detectron2 מddy mAP Precision/Recall per class וזמן חיזוי. **פערים מדעיים**: אין קורפוס ציבורי מוסכם לסמלים. (Ultralytics Docs, arXiv, detectron2.readthedocs.io).

הצלבה מול Schedules

- PDF על התאמת מזהה ציוד חדרים Precision/Recall Excel או טבלאות.

ט' נתונים לדוגמה + סקריפט הערכה

- מבנה תיקיות : /pdf_vector_clean /pdf_scanned_noisy /dwg_with_xref
- סקריפט הערכה יחזיר על ריצות עם seed קבוע ושמור תוצאות CSV.
- Makefile להרצת Makefile

7. מפת דרכים 30 90 ימים

30 ימים

- רכש/בחירה ODA או RealDWG.
- בניית Intake+Parsing DWG ו PDF בסיסי.
- יצירת קורפוס סמלים קטן עם הנחיות תיוג.
- קритריון יציאה: שליפה עקבית של Blocks+Attributes+Layers ב 3 פרויקטים F1 OCR.
- מילימ 0.85 על דוגמאות נקיות. **סיכון**: גישה DWG יוצר (Platform Services)

30 60 ימים

- אימון מודל סמלים ראשון. YOLOv8.
- בניית Mapping לטופולוגיה בסיסית ובדיקות עקביות.
- דשبورד QA במטביס.
- קритריון יציאה 0.9 Precision: לסמל לייבה PDF וקטורי ויזיה חדרים ב 85% מהאלמנטים בחזית (Metabase).

90 ימים

- הרחבת קורפוס HITL Active Learning.
- אינטגרציה עם Audit Log ו Schedules.
- קשיחות אבטחה וציתנות.
- קритריון יציאה 0.9 Recall: לסמל לייבה פער נ נתונים < 5% מול Schedule.

8. טבלת החלטות השוואתית תמצית

| פתרון | יתרונות | חסרונות | עלות כוללת גסה | נעלית ספק | סקילינג | אבטחה ופרטיות |
|||||
| ODA SDK | CISCO ODA | רחוב קהילה תעשייתית | רישיוני בתשלום | מנוי ארגוני ישירות מול ODA
| ביןנית | טוב/On-prem | ענן לפי בחירה (opendesign.com) |
| RealDWG | תאמיות מקסימלית ל-DWG מחייב תלות אוטודסק | לפि התקשרות | גבואה
| טוב | תנאי Autodesk (Autodesk Platform Services) |
| ezdxf + DXF | קוד פתוח קל | בלי DWG אובדן סמנטיקה | נמוכה | נמוכה | טוב | פתוח
| (ezdxf.readthedocs.io) |
| pdfminer.six | שקוֹף | איטי יחסית | נמוכה | נמוכה | טוב | פתוח | [pdfminer/pdfplumber](http://pdfminer.pdfplumber) | MIT
| (GitHub) |
| PyMuPDF/MuPDF | מהיר GPL | או מסחרי | ביןנית | ביןנית | טוב | אוסף רישיוני Autodesk (PyMuPDF) |
| PaddleOCR/docTR | OCR | חזק | GPU כוונון נדרש | ביןנית | נמוכה | טוב | פתוח
| (paddlepaddle.github.io, mindee.github.io) |
| YOLO/Detectron2 | ביצועים אקו-סיסטם | דרוש דатаה | ביןנית | נמוכה | טוב | פתוח
| (Ultralytics Docs, detectron2.readthedocs.io) |

9. אבטחה וציות

- **מידע רגיש:** תוכניות תשתיתית עשויה להיות רגישות; יש להחיל בקרנות לפיקוד 53 NIST SP 800-53 . ISO 27001 יכול בקרת גישה לוגינג הצפנה במנוחה ובמעבר (NIST Publications, ISO).
 - **פרטיות:** כאשר יש נתונים זיהוי אישיים במסמכים נלוויים לעמוד בעקרונות GDPR מינימיזציה מטרה הגבלת אחסון (GDPR) .
 - **רישיוני DWG ו-PDF Engines:** הקפדה על תנאי Autodesk בתוכן וב시스템ים רשומים; שימוש ב-PyMuPDF מהיב GPL או רישיוני מסחרי. **נקודות ספק:** פרשנות משפטית ספציפית דרושת יוץ משפט (Autodesk, PyMuPDF) .
-

10. Human in the Loop

- **נקודות ביקורת:** אימונות מקרה איחוד מונחים תיקון False Positives בדיקת מעגליים בעיתתיים.
- **Active Learning:** סוף התערבות לדוגמא $\text{confidence} < 0.6$. Active Learning: הנקודות תיוג: CISCO סימנים סיבוב קנה מידת ורעש.
- **תיעוד:** כל תג של אנוש נרשם משמש fine-tuning הבא.

11. תוצרית קוד ותהליכי

פואדו קוד שלבים מרכזים

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) # pdfminer/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](#)
- Makefile בסיסי להרצה : [Makefile](#)

12. אתגרים ו Edge Cases וטיפול מוצע

- PDF פורט ירוד Super-resolution De-skew Binarization: אופציונלי בדיקת שפות. **5PF**:
כתב יד לא נתרך היטב ניתן לנסות Whisper רק לטעסט חופשי לא סימבולוגיה .
(digitalcommons.usm.maine.edu)
- טקסטים מסובבים/מראה LayoutParser + Orientation. (layout-parser.github.io : זיהוי)
- סמלים מותאמים ללא מקרא + HITL: התאמת קונטקט שכבה וחיבורים.
- קנה מידה לא עקי: קריית Scale מה Title Block אימות מול מרחקים ידועים. תקן ISO 5455 רלונטי לשיקולי סקלות בגילוונות([ISO Iteh Standards](#)).
- Splines וקשתות: אפרוקסימציה לפולילינים שמיירת שגיאה מקסימלית.
- ריבוי גילוונות ו Revisions: קריית שדה Revision ב Title Block יומן שינויים.
- XREF אشبור: דיווח חוסר וקובץ (landfx.com) Audit.
- Units מעורבים: המרת יחידה לפי Units בשכבות הגילוון.
- Blocks דינמיים: קריית פרמטרים (Visibility/Stretch) דרך (Drer) SDK. (Autodesk Help)
- התאמת מול Schedule חיצוני עם שמות לא עקבאים Fuzzy Matching + מילון ארגוני.

13. דוגמאות נדרשות תוצריים

קובצים לדוגמה ציבוריים לא סופקו כאן לכן מוגנות דוגמאות פורמט עם תוצריים מלאים שתואימים לשכינה. בהמשך הפרויקט יוחלפו בקובצי אמת.

- **ONSלפני ואחרי** ראו קובצי ה Schema JSON וה CSV בתנופהם להורדה לעיל.

- ציון ודאות לכל שדה מופיע בדוגמת ה.”post_normalization CSV
-

14. מדיניות ציטוטים והוכחות יישום בפועל

- לכל טכנולוגיה צורפו 2 מקורות ראשיים דוקומנטציה רשמית מאמרם 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](#))
 - **מפורט סטנדרט IEC ElectricalDomain IEC 60617 ISO 128 ISO 5455. מגבלה: חלקם** אחריו חומרת תשלום ([ifc43-docs.standards.buildingsmart.org](#), [IEC Webstore](#), [ISO](#), [Iteh Standards](#))
-

15. מתודולוגיית מחקר ומגבלות

- מה נעשה: סקירת דוקומנטציה רשמית מאמרם אקדמיים עדכניים פטנטים בלוגים הנדסיים רציניים.
 - מה לא נעשה: לא הורצו מדידות אמפיריות בפועל לא בוצעה המרת DWG אמיטאים עקב מגבלות סביבת ריצה וזכויות.
 - **ספקים:** מחיר GPU משתנים יומיות 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 הגוף למרות IEC 60617.
 - DWG דורש SDK מורשה או זרימת DXF עם אובדן.
 - OCR על שרטוטים סרוקים רועשים ידרש HIT מתרשם.
 - סמלי קצה ובלוקים דינמיים נדרים יחייבו נתוני אימון סינטטיים ([Google Patents](#)).
-

17. שאלות הבקרה הכרחיות

1. אילו קטגוריות רכיבים חובה בגל ראשון
2. האם קיימת סימבולוגיה ארגונית תקופה או לפני פרויקט
3. היכן יאחסן המידע PostGIS/Postgres או מען מנהל ואילו הרשות per field
4. מהן דרישות דיקן מינימליות

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

6. אילוצי זמן תקציב וחומרה כולל ענן לעומת-premises

18. החלטות מומלצות

- **מסלול מודרך**: ODA SDK + pdfminer/pdfplumber וטורן PDF ל-JSON + Metabase/PostGIS + PostgreSQL + לאחסון QA + לסרוק YOLOv8 ~8 שבועות. סיכון ביןוני נשלט .
opendesign.com, pdfminer.six, GitHub, paddlepaddle.github.io, mindee.github.io, PostGIS, Metabase
- **גיבוי RealDWG**: במקום ODA אם נדרשת תאימות Autodesk מקסימלית או הסכומות OEM. (Autodesk Platform Services)
- **מסלול מחקר נוסף**: בניית קורפו ציבורי למחקר סמלי شامل לבניין עם הנחיות תיוג כדי לאפשר Benchmark פועל לYOLO/Detectron2. (Ultralytics Docs, detectron2.readthedocs.io)

נספחים הורדות

JSON Schemas:

- [drawing_sheet.schema.json](#)
- [electrical_element.schema.json](#)
- [electrical_connection.schema.json](#)
- [room_space.schema.json](#)

לדוגמה: CSVs

- [sample_pre_normalization.csv](#)
- [sample_post_normalization.csv](#)

Konfiguracija : Makefile

- [config.sample.yaml](#)
- [Makefile](#)

מקורות מרכזיים שנשענו עליהם לכל הטענות

- Autodesk DWG: ODA Drawings SDK ומודול תמחור RealDWG סקירה תיעוד [Autodesk](#)
- Attributes/Dynamic Blocks XREF. (opendesign.com, Autodesk Platform Services, Autodesk Help, sourcecad.com)
- PDFium. (pdfminer.six, [PDF/OCR](#): pdfminer.six pdfplumber PyMuPDF GitHub, PyMuPDF)

- **CV:** Detectron2 YOLOv8/10 SAM. (detectron2.readthedocs.io, [Ultralytics Docs](#), [arXiv](#))
- **IFC ותקנים :** IFC ElectricalDomain IEC 60617 ISO 128 ISO 5455. ([ifc43docs.standards.buildingsmart.org](#), [IEC Webstore](#), [ISO](#), [Iteh Standards](#))
- **DB/ETL/Dashboard:** PostGIS DuckDB Airflow Dagster dbt Metabase Superset Grafana. ([PostGIS](#), [DuckDB](#), [Apache Airflow](#), [Dagster Docs](#), [dbt Developer Hub](#), [Metabase](#), [Superset](#), [Grafana Labs](#))
- **אבטחה :** NIST SP 800 53 ISO 27001 GDPR Art.5. ([NIST Publications](#), [ISO](#), [GDPR](#))
- **פטנטים/מאמרם ייעודיים .** ([Frontiers](#), [MDPI](#), [Justia Patents](#), [Google Patents](#))

היכן שיש ספק או חוסר מידע:

1. **ביצועי OCR ויזהו סמלים בקבצי הארגון תלוי** דатаה אין נתונים אמת מידה אצלנו.
2. **מחירי SDK מדויקים** נתונים מסחריים דרישים הוצאות מחיר רשמיות.
3. **תרחישי קצה ספציפיים** כמו בלוקים דינמיים נדרשו דוגמאות אמת.

עד הבא ריאלי: לאשר את השאלות בסעיף 17 לבחור ODA/RealDWG להריז POC של 8 שבועות לפי המפת דרכים ולהתחליל באיסוף קורפו אונטציה עמ.HITL

מחקר מבוסס ראיות: שליפת נתונים מתכון חמל מ-PDF ו-DWG והמרתם לבסיס נתונים מבנה

תקציר מנהלים

הצורך באוטומציה של שליפת נתונים מסמכים תכנון חמל, לרבות קבצי PDF ו-DWG, הוא קריטי להtagברות על חוסר יעילות ידנית לאפשר ניתוח מתקדם. תהליכי ידניים גוזלים זמן רב, דורשים כוח אדם רב ומועדים לטעויות אנוש, במיחוד בעט שניינ' תוכניות תכופים.¹ חוסר עקביות נתונים ופיזורם על פני מערכות שונות פוגעים ביכולת לנצל מידע היסטורי ולשפר הצעות עתידיות.¹ הגישה המוצעת היא רב-גונית, כוללת שימוש בערכות פיתוח תוכנה (SDKs) ייעודיות, ראייה ממוחשבת מתקדמת, תורת הגרפים וארקיטקטורת צינור נתונים חזקה.

המצאים העיקריים מצביעים על כך שבחרית הכלים הנכונים, התיחסות יסודית לשיקולי קניין רוחני, ההכרח בהכשרת בינה מלאכותית ספציפית לתחום, בניית מודל נתונים מבנה וחשיבות יישום הדרגי, הם מרכיבים חיוניים להצלחה. הפוטנציאלי להחזיר השקעה (ROI) ממשמעותי קיים, אך הוא דרוש השקעה ניכרת בטכנולוגיה ובמוחשיות.

1. מבוא: הצורך האסטרטגי בשליפת נתונים אוטומטית בתכון חמל

סעיף זה יציג את ההקשר וידגים את הערך העמוק של שליפת נתונים מסמכים תכנון חמל.

1.1 אתגרים נוספים בטיפול ייני בתוכניהם מشرطוטי חמל

טיפול ייני בתוכניהם מ��ונות חמל מציב שורה של אתגרים משמעותיים המעצבים יעילות ותחרותיות. תהליכי שליפת נתונים קובנציונליים, כגון ביצוע "takeoffs" (אומדי כמיות) ידניים, הם גוזלי זמן ומאז, ומועדים לטעויות אנוש. לדוגמה, כאשר תוכניות פרויקט משתמשות תDIR עם תוספות ושינויים בעיצוב, ביצוע מחדש של אומדיים ידניים מוביל לשגיאות ועיכובים מצטברים.¹

תהליכי ידניים יוצרים צווארי בקבוק במחזור הצעות מחיר צפופים. קבלנים נאלצים לעיתים קרובות לוותר על הזרמוויות רוחניות פשוט משום שאין מספיק זמן להגיש הצעות על כל פרויקט תוך כדי ביצוע אומדיים ידניים. הדבר מביא לאובדן הכנסות ולפוטנציאלי עסק בלתי ממושך.¹ ניהול פרויקטים חמלים רבים כרוך לעיתים קרובות בערימה אוטית של קבצי אומדיים קודמים המפוזרים על פני מילימטרים, כוננים מקומיים או תיikitיות מסוימות. חוסר ארגון זה הופך את ניצול הנתונים ההיסטוריים להצעות עתידיות לבלי אפרטי כמעט, ומונע סטנדרטיזציה של תהליכי.¹

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

יתרה מכך, מסמכים תכנון חמל מכילים לעיתים קרובות פרייסות מורכבות, טקסט מסובב, מידע מוקוטע וסמלים מיוחדים. טקסט עשוי להיות מושתר על ידי קווים או מסומן בהערות וסמלים שלדים גנריים של זיהוי תווים אופטי (OCR) אינם מאומנים להבין.³ כל אלה הופכים את הפרשנות הידנית ואת השליפה האוטומטית לקשות באופן מהותי.²

הועלות המצתברת של תהליכיים ידניים חרוגת בהרבה מהוצאות העבודה הישירות. בעוד שעליות כוח אדם ידנית ברורות, העליות העקיפות – הכנסות אבודות מהוצאות מחיר שלא הוגש, פגעה בתחרותיות וקבלת החלטות לקיה עקב נתונים לא מדויקים או לא עקבאים – הן לרוב גדולות בהרבה וקשות יותר לelimination. אונגרים אלה אינם מבודדים; הם מצטברים. נתונים לא מדויקים הנובעים ממשיפה ידנית² דורשים עבודה מחדש¹, מה שמחמיר עוד יותר את מגבלות הזמן ומוביל להזדמנויות נוספות שהוחמצו. היעדר נתונים מרכזיים¹ מונע ניצול פרויקטים קודמים, ומהיבר מאמץ ידני חוזר ונשנה. לפיכך, הצורך האסטרטגי באוטומציה חרוג מעבר לרוחוי ייעילות בלבד; הוא נוגע לפיתוח פוטנציאל עסק חדש, שיפור אינטראקטיביות ההחלטה בשוק תחרותי.

2. החזון: הפעלת אנליטיקה מתקדמת ויעילות תפעולית באמצעות נתונים מובנים

ה חזון הוא ייצור מערכת המאפשרת לשילוף אוטומטית של נתונים מובנים, ובכך משפרת באופן דרמטי את הדיקט והשלמות של הנתונים. אוטומציה, בשילוב עם תהליכי אימוט וניקוי נתונים, מבטיחה נתונים נקיים, עקובים ומדויקים המספקים תובנות אמינות.²

שלילת נתונים מהירה ומדויקת מאייצה את מחזור הפרויקט, החל מהוצאות מחיר וסיקורות תכנון ועד לאומדי חומריים, ובכך מקצרת את זמן הביצוע הכלולים של הפרויקטם.¹ בסיס נתונים מובנה ישמש כמגרר מרכזי ונitin לחיפוש, שיאפשר גישה קלה, יכולת חיפוש ושימוש חוזר בנתוני פרויקטים היסטוריים. הדבר יאפשר סטנדרטיזציה של תהליכיים וייפר את הוצאות המחר העתידיים.¹ יתר על כן, נתונים מובנים מהווים בסיס איתן לאנליטיקה מתוחכמת, מודלים חזויים ושילוב עם מערכות מידע מידע מבנים (BIM), תאומים דיגיטליים וכל' סימולציה. גישה בזמן לנתונים מדויקים ומקיפים תתרמו בקבלת החלטות תפעוליות ואסטרטגיות טובות יותר.²

2. הבנת פורטט נתונים המקורי: PDF ו-DWG

סעיף זה יספק ניתוח טכני מפורט של שני פורטט המקוריים, PDF DWG, והכלים הזמןניים לניתוח ושלילת נתונים מהם.

2.1 צלילה עמוקה לדXF/DWG: מבנה, יכולות ועודי קנייני

קבצי DWG ו-DXF הם פורטטים סטנדרטיים בתעשייה העובר תוכניות CAD, אך יש להם הבדלים מהותיים המשפיעים על תהליכי שלילת הנתונים.

2.1.1 לעומת DXF: הבדלים מהותיים ושימושים

DWG (Drawing) הוא פורטט קובץ בינהי קנייני, המהווה את הפורטט המקורי של AutoCAD. הוא מותאם לאחסון קומפקטי של נתונים תכנון דו-ממדיים ותלת-ממדיים ומטא-נתונים.⁵ פורטט זה שומר על תוכנות מיוחדות ונתונים קנייניים, מה שהופך אותו לאידיאלי לאינטגרציה חלקה בתוך סביבת AutoCAD ולשמירת נתונים כאשר נאמנות לפרטים היא קריטית.⁶

DXF (Drawing Exchange Format) לעומת זאת, הוא פורטט קובץ פתוח, מבוסס טקסט פשוט, שתוכנן במיוחד להחלפת שרטוטים בין יישומי CAD שונים.⁶ המבנה הקרייא-לאדם שלו מקל על הפרשנות ואף על שינוי ידני במידת הצורך, ובטיח יכולת פעולה הדדית על פני פלטפורמות CAD מגוונות ומקל על שיתוף פעולה בין-פלטפורמי.⁶ עם זאת, המרת DWG-L-DXF עלולה לגרום לאובדן של חלק מהתוכנות המיוחדות והנתונים המקשורים אליהן.⁶ חשוב לציין שני הפורטטים, DWG ו-DXF, מחסנים גרפיקה וקטוריית (קוויים, קשתות, עיגולים, טקסט, מצולעים) המבוססת על גיאומטריה ויחסים, מה שמבטיח מדרגות ללא פיקסלציה.⁷

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-DXF⁸ של AutoCAD.⁸ זהה לתוך-קבוצה של ObjectARX SDK, המתמקדת אריך ובקבוצת הנדרטוט (ישיות ויחסים) ללא תמייה בתצוגה או במשק משתמש.⁸ הוא אינו דורש התקינה של AutoCAD.⁸

דרישות המערכת של RealDWG כוללות Microsoft Windows 10/11, מעבד 64 סיביות בmahiroot +3 GHz, זיכרון RAM של 16 GB, שטח דיסק של 4 GB וסביבת פיתוח Microsoft Visual Studio (לדוגמה, VS 2022.NET 8.0. עברו RealDWG 2026).⁸ הרישיון מטופל באופן גלובלי על ידי Tech Soft 3D.⁸ המחיר הוא 8,000 דולר אמריקאי או 7,500 אירו לשנה עבור עד 10,000 משתמשים קצחים, כאשר נפחיהם גדולים יותר דרישים יצירתיות ישירות.⁸ הוא פועל במודול הפעזה לא תמלוגים.⁸ מיועד בעיקר לישומים "לא תחרותיים".⁵

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

ODA Drawings SDK (שנודע בעבר בשם Teigha) תומך בפונקציונליות קרייה/ כתיבה/הציג עבור קבצי.dwg ו-.dgn.¹¹ הוא מציע API ברמה גבוהה לManipulations נתונים, כולל אובייקטים אוטומטיים, שיבוט, טרנסקציות, ביטול/חזרה.¹¹ הוא תומך באובייקטים סטנדרטיים מורכבים כמו מימדים, טקסטים מרובי שורות, טבלאות, מוליכים מרובים ובלוקים דינמיים, ומאפשר יצירה אובייקטים מותאמים אישית.¹¹ הוא מספק המרה דו-כיוונית מקיפה בין DWG ו-DGN, כמו גם יבוא/יצוא של PDF, DWF ופורמטים אחרים.¹¹ תוכנות מפתח כוללות תמייה בשכבות, בלוקים, תוכנות, אילוצים גיאומטריים ומימדיים, תיעוד מודל (תצוגות שרטוט חכמות) ו-TRACE (סימונים).¹¹ הספריות כוללות API ב-C++ ועטיפות SWIG עבור Python ו-.NET.¹¹

ODA היא קונסורציום ללא מטרות רווח בבעלות חברות.¹² התוכנה יכולה להיכל רק בתוכנות ישומיים בבעלות חברות נוכחים, בכפוף להסתכם חברות חתום.¹⁴ העולות משתנות בהתאם לצרכים הספציפיים, עם עלות שנתנית ממוצעת של כ-38,000 דולר, ומהיר מרבי מדוח של כ-92,000 דולר.¹⁵ רישיונות משתמש/מפתח מתחילה ב-249 דולר (לא הרחבת PDF) או 399 דולר (עם הרחבת PDF), עם הנחות על נפח.¹⁶ רישיונות זמן ריצה לשרת יקרים יותר, החל מ-999 דולר/1,499 דולר.¹⁶ רישיונות אתר לשימוש בלתי מוגבל במיקום פיזי היחיד הם 2,999 דולר/3,999 דולר, ורישיונות הפעזה בלתי מוגבלים הם 2,999 ו-4,499 דולר.¹⁶

2.1.4 חלופות קוד פתוח (לדוגמה, ezdxf)

ezdxf היא חבילת פיתוח לקבצי DXF. היא מציעה פונקציונליות דומה לדxfwrite ו-dxfgrabber אך מקופה יותר.¹⁷ יש לה טביעה רgel זיכרון קטן יותר.¹⁷ עם זאת, ezdxf אינה תומכת באובייקטי OLE, הנחשים לה"החלטת עיצוב מאוד" בשרטוט CAD עקב בעיות תאימות עתידיות.¹⁷ היא מתמקדת בעיקר ב-DXF, יתכן שלא תוכל את כל תוכנות ה-DWG הקיימים.

DWG/DXF משפטים משפטיים עבור DWG/DXF 2.1.5: קניין רוחני ושיקולים

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

לרשום את "DWG" כסימן מסחרי עבור Autodesk, מתוך הכרה בו כמתאר פורמט קובץ⁵,
Autodesk ממשיכה לטעון שהוא סימן מסחרי.¹⁹

Autodesk הציגה "סימני מים" בקבצי DWG כדי להבחין בין קבצי AutoCAD "אמיתיים" לבין אלה
שנعواשו בהנדסה לאחרор על ידי מתחרים, מה שהוביל למחלוקות משפטיות נוספות.²⁰ 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 ובמיצוג מוצריה, תוך שהיא עדין רשאית לפתח תוכנה ניתנת להפעלה הדידית ולהשתמש
בסיומת.dwg.¹⁹ רישיון Autodesk RealDWG SDK ניתן בהבנה מרווחת שהוא מיועד ליישומים
"לא תחרותיים".⁵

הסתמכות בלבד על מנתchi DWG בקוד פתוח או בהנדסה לאחרור עבר פעולות עסקיות קרייטיות,
במיוחד ליצירת נתונים מבניים שעשויים לשמש ביישומים תחרותיים, טומנת בחובה סיכון משפטי
משמעותי בשל טענות הקניין של Autodesk ופעולות האכיפה הקודומות שלה. אפילו בעט בחינת
ADA, סקירה יסודית של הסכם החברות שלה וההשלכות של הסדר הפשרה והקודם שלה עם
Autodesk היא קרייטית כדי לוודא שתרכיש השימוש המוצע (לדוגמה, יצרת מסד נתונים מבנה
עבור ניתוח פנימי לעומת מוצר מסחרי) אינו מפר תנאים או גורר אתגרים משפטיים עתידיים.
הבחירה בין RealDWG (אם ניתן לרישיון עבור תרכיש השימוש הספרטיפי), ADA או כל קוד פתוח
אינה רק טכנית או פיננסית; זהה החלטה משפטית אסטרטגית שעשויה להשפיע על הcadיות לטווח
ארוך ועל עמדת הקניין הרוחני של הפטرون המפותח. הדבר מחייב סקירה של "יעוץ משפטי".

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

| | | | | | | | | | |
|-----------|-------------|--------|----------|----------|--|-------|----|-----|----|
| דָּר | תְּ | שׁ | 00 | וֹת | | מַנוּ | , | anc | wi |
| יִשְׁבָּב | רְבָּ | וֹ | +0 | שָׁנְתִי | | תַּת | Ja | e | ng |
| תֵּת | - | קִיְּ | דוּ | וֹת, | | W) | va | s | s |
| חַבְּרָה | פֶּלֶר | רָ | לָרָ | רְמוּת | | in | , | S | S |
| רוֹתָה | טֶפֶחָה | אָרָהָ | שְׁוֹנוֹ | שְׁוֹנוֹ | | do | Py | D | D |
| תְּהִרְ | וּרְקִיְּ | הָהִ | תְּהִרְ | תְּהִרְ | | ws | th | K | K |
| מוֹמִין | מִפְּדָ | בָּ | מִשְׁתָּ | מִשְׁתָּ | | , | on | | |
| רַכְבִּי | תְּנִיבָה | (בָּ | מִשְׁ/ | מִשְׁ/ | | Lin | | | |
| וֹתָה | כָּהָשָׁתָה | תָּ | מִפְּתָה | מִפְּתָה | | ux, | | | |
| | | | , | | | M | | | |

| | | | | | | | | | | | | |
|------------------------------------|--------------------------|--|---|--|--------------|---|--|---|---|---|---|----------|
| | | | | | | | | | | | | |
| מג' בל' ת' עי' קר' יות' | ית' רון' וו' עי' קר' ים' | קנ' חן' /' ס' כו' מ' ש' ט' | ת' מ' ש' ע' ת' מ' י' ס' כו' מ' מ' ס' מ' (| על' שו' ש' ער' ת' (מ' י' נ' מ' מק' ס' מ' (| מודל' רישוי' | פונ' קצ'י' וני' ות' תצ' גה' גה' נדו' ר' | ת' מ' כה' בא' וב' יק' ט' מ' מו' רכ' ב' | ת' מ' קר' אה/ כת' בה DW , בל' וקי' מ' דין' מי' מ' טב' לא' ות' M Le ad er (s | ש' פ' ת' ת' הפ' ה' על' ה' נט' פ' מ' רכו' ת' דוג' מה' , DW G/ DX F | ת' מ' קר' אה/ כת' בה DW , בל' וקי' מ' דין' מי' מ' טב' לא' ות' M Le ad er (s | ת' מ' קר' אה/ כת' בה DW , בל' וקי' מ' דין' מי' מ' טב' לא' ות' M Le ad er (s | S D /K . |
| מ' שפ' טיו' ת' פ' טנו' צ' אל' יות' | -ב- D G N | ל' ה' ס' מ' ה' ה' ח' ש' בר' ות' עק' ב' | אם' לדר' מ' ה/ה' מ' א' ום' תן' (| שרת' , אתר' הפט' ה' בלתי' מוגב' לת' | | | | | ac OS (| | | |

| | | | | | | | | |
|-----|-----|-----|------|-----|------|-------|-------|-------|
| מג | ית | קנ | על | ת | פונ | ת | ש | S |
| בלו | רונ | חנ | ות | ת | קצ' | ה | פ | D |
| ת | וות | / | שנו | שׁו | ונלי | בָּ | ת | K |
| עֵ | עֵי | ס' | ת | ער | ות | רְכִ | תְּכִ | תְּכִ |
| קר | קר | כוו | (מ | ט | תצ' | בַּ | תְּכִ | תְּכִ |
| יות | ים | מ | מִ | טִ | וגה | מְ | תְּכִ | תְּכִ |
| | | מו | יְנִ | טִ | /רִי | לְ | תְּכִ | תְּכִ |
| | | מו- | נו | טִ | נדָן | דָּ | תְּכִ | תְּכִ |
| | | טִ | מו- | טִ | ר | דִּין | תְּכִ | תְּכִ |
| | | | | | | מֵי | | |
| | | | | | | מַ, | | |
| | | | | | | טָבָּ | | |
| | | | | | | לֹא | | |
| | | | | | | וֹתְּ | | |
| | | | | | | M | | |
| | | | | | | Le | | |
| | | | | | | ad | | |
| | | | | | | er | | |
| | | | | | | (s) | | |

| | | | | | | | | | | | |
|------------|--------|--------|--------|---------|--------|--------|----------|--------|---------|-------|----------------|
| ת | מִ | כָּה | בָּא | וּבָ | יַקְ | טֵ | מִ | תָ | פָּ | שָׁ | S |
| מָגְ | יִתְ | קְנָ | תִּתְ | תִּתְ | פָּנוֹ | בִּי | קְרִי | מְעֻ | תְּפִ | תְּפִ | D |
| בְּלֹוְ | רוֹגְ | חֲנָ | שְׁוָ | שְׁוָ | קְצִיְ | לִ(| אַהֲ/ | רְכוֹ | תְּכִ | תְּכִ | K/ |
| תְּתָ | וְתָ | / | עָרָ | עָרָ | וְתָ | דוֹגְ | כְּתִיְ | הַפְּ | נוֹ | נוֹ | Mפְ |
| עֵיְ | עֵיְ | סִיםְ | רִישְׂ | מוֹדָלְ | תְּצִ | מָהָ | DW | בָּהָ | תְּ/תְּ | A | תְּחִ/סְפָּקָן |
| קְרָ | קְרָ | קְרָ | (מִ | רִישְׂ | וְגַהְ | , | בְּלִ | G/ | Pְ | עֵיְ | . |
| יּוֹתְ | יּוֹתְ | יּוֹתְ | מִינִ | רִישְׂ | רִידְ | וְקִיְ | DX | וְקִיְ | מְכִ | וְתִ | כָּלְ |
| | | | סִיםְ | | רִידְ | F | דִּינִיְ | מִיְ | קְרִיְ | יְתִ | |
| | | | סִםְ | | | | מִיְ | , מִ | | | |
| | | | סִםְ | | | | טְבִ | לְאָ | | | |
| | | | סִםְ | | | | | וְתִ | | | |
| | | | סִםְ | | | | | M | | | |
| | | | | | | | | Le | | | |
| | | | | | | | ad | er | | | |
| | | | | | | | (s) | | | | |
| תְּבִלְדָּ | | | | | | D | | | | | |
| | | | | | | W | | | | | |
| | | | | | | (G) | | | | | |

2.2 ניתוח מסמכים PDF: הבחנה בין תוכן וקטורי ורסטריאי

מסמכים תכנוניים شامل קיימים לעיתים קרובות כקובצי PDF, שיכולים להיות מבוססי וקטור (שנוצרו מ-CAD) או מבוססי רסטר (תמונות סורוקות). גישת השיליפה שונה בהתאם למשמעותם בהתקף הבדיקה.

2.2.1 שליפת PDF וקטורי: כלים וטכניקות לנתחים גיאומטריים וטקסטואליים

קובץ PDF וקטוריים מכילים טקסט מוטבע ופרימיטיבים גיאומטריים (קוויים, צורות) הנגישים לשירות כלים לשילפה כוללים:

- **Adobe Acrobat:** מאפשר ייצוא תמונות ובחירות טקסט מקבץ PDF לפורמטים שונים.²¹ הוא יכול לייצא תמונות רستر אך לא אובייקטים וקטוריים ישירות.²¹ הוא מציע גם כלים לעריכת טקסט, תמונות ומבנה מסמך.²¹
- **(PyMuPDF) Fitz:** ספריית פיתון בעלת ביצועים גבוהים לשילפת נתונים, ניתוח, המרה ומיניפולציה של מסמכים כמו קבצי PDF.²² היא מהירה באופן משמעותי מחלופות כמו PDFMiner PyPDF2 או PyPDF Pro לשילפת טקסט.²² PyMuPDF מציעה הרחבות מסוימות לתמיינה במסמכים Office ושילובים של LLM/RAG.²⁵
- **(pdfminer.six) PDFMiner:** חבילת פיתון המתמקדת בשילפת טקסט, מטא-נתונים וSMART OCR קבצי PDF.²⁷ היא מוצנית בשילפת טקסט אך פחות יעילה בטיפול ברכיבים שאינם טקסטואליים ומתקשה בפתרונות מורכבות במיוחד (מרובות עמודות, טבלאות מקווננות).²⁷ ניתן לשילבה עם כל OCR למסמכים סרוקים.²⁷
- **pdfplumber:** ספרית פיתון נוספת לשילפת PDF, המושווות לעתים קרובות ל-PDFMiner.²⁴

עבור נפחים גדולים של תוכניות شامل בפורמט PDF, במיוחד עם פריסות מורכבות שבן מהירות היא קריטית. PyMuPDF היא הבחירה העדיפה לשילפת טקסט ראשונית ומבנה בסיסית בשל יתרון הביצועים שלה. בעוד ש-PyMuPDF מטפל בטקסט היטב, הצורך לשולף מידע גיאומטרי ורכיבים לא טקסטואליים מורכבים (סמלים, קוויים) מקבצי PDF וקטוריים עשוי לדרוש עבודה נוספת תכונות מיוחדות שאין מוגשות במקור, מה שמציבע על גישה מרובות כלים או פיתוח מותאם אישית על בסיס היכולות הבסיסיות של PyMuPDF. מתן עדיפות לביצועים בשלב שלילת ה-PDF הוא קריטי לעילות ולדרגות היכולת של צינור הנתונים, ומשיער ישירות על זמן העבודה ועלויות המחשב, במיוחד אם מדובר באלפי מסמכים.

2.2.2 **עבוד PDF רستر/סrok: הצורך וראיה ממוחשבת**

קובץ PDF רستر הם למעשה תמונות של מסמכים. טקסט וגרפיקה אינם נגישים ישירות נתונים מבנים; הם פיקסלים. גישה זו דורשת זיהוי תווים אופטי (OCR) להמרת תמונות טקסט לטקסט קרייא למcona, וטכניקות ראייה ממוחשבת מתקדמות (זיהוי אובייקטים, סגמנטציה) לזרחי וסיווג סמלים וקוויים. כל OCR סטנדרטיים (כמו EasyOCR או Tesseract) בדרך כלל אינם מספיקים לשרטוטי הנדסה בשל טקסט מסובב, קיטוע, טקסט מוסתר על ידי קוויים, וונוכחות של סמלים והערות מיוחדים שהם אינם מאומנים להבין.³

2.2.3 **רישוי מסחרי של PyMuPDF לשימוש מקצועי**

PyMuPDF היא הרחבה מסחרית של PyMuPDF Pro, המציעת תכונות משופרות כמו טיפול במסמכים (DOC/X, PPT/X, XLS/X) ושלובים של LLM/RAG.²⁵ היא תומכת בשילפת טקסט וטבלאות ובהמרת מסמכים.²⁵ הרישיון דרוש רישוי לפקציונליות מלאה מעבר למגבלת 3 עמודים או מפתח ניסיון לזמן מוגבל.²⁵ התנאים אינם מפורטים במפורש בתקצירים אך דורשים יצירת קשר עם מחלקת המכירות של Artifex Software Inc. להתקינה אישית, הפצה ומדריגות ללא הגבלות.²⁵

עבור פתרון שליפת נתונים מkcזען ורחב היקף, הסתכומות בלעדית על גרסת קוד פתוח של PyMuPDF עשויה להיות בלתי מספקת אם נדרשות תכונות מתקדמות (לדוגמה, עיבוד ישיר של מסמכים Office שאינם PDF שעשויים להכיל דיאגרמות מוטבעות, או ניצול שילוב LLM/RAG/²⁵ לשילפה קונטקטואלית עשרה יותר). הצורך ליצור קשר עם Artifex לקבלה תנאים מסחריים מרמז על תהליך משא ומתן, והעלות עשויה להיות משמעותית, בדומה ל-SDKs ברמה ארגונית אחרים. זה מוסיף לתקציב הפROYיקט הכללי ולמורכבות הרCHASE. ארגונים חיברים להעריך האם היכולות הנוספות של PyMuPDF Pro (או SDKs מסחריים דומים ל-PDF) מצדיקות את עלות הרישיון והמורכבות, במיוחד אם מסמכים המקוריים שלהם חריגים מ-PDF טהורים או אם הם מתוכננים לשלב עם מודלי AI מתקדמים לשילפה נתונים עשרה יותר.

טבלה 2: השוואת鄙יעים של ספריות לשילפה טקסט מ-PDF

| מגבילות עיקריות | יתרונות עיקריים | ביבועים מהירות יחסית/זמן עבור N עמודים) | פונקציה עיקרית | ספרייה |
|--|---|---|---------------------------|--------------|
| רישיון מסחרי עבור תכונות Pro | מהירות גבוהה, תכונות חזקות, גרסה Pro מסחרית | ההירה ביותר 3.05 (לדוגמה, 7031 ²² עמודים) | שליפה טקסט מניפולציית PDF | PyMuPDF |
| איתית לשילפה טקסט | מניפולציית PDF בסיסית | איתית באופן שימושי (לדוגמה, 494.04 שנים עבור 7031 ²² עמודים) | מניפולציית PDF | PyPDF2 |
| מתקשה בפתרונות מורכבות, רכיבים לא טקסטואליים | שמור פרישה, טוביה לטקסט | איתית יותר (לדוגמה, 227.27 שנים עבור 7031 ²² עמודים) | שליפה טקסט, שימור פרישה | PDFMiner.six |
| איתית לשילפה טקסט | מניפולציית PDF | איתית יותר (לדוגמה, 10.54 שנים עבור | מניפולציית PDF | PDFrw |

| מגבליות עיקריות | יתרונות עיקריים | ביטחונות מהירות יחסית/זמן עבור N עמודים) | פונקציה עיקרית | ספרייה |
|-----------------------|-----------------|--|----------------|---------|
| | | (7031 עמודים) ²² | | |
| איתית לשילוף טקסט | מניפולציה PDF | איתית יותר (לדוגמה, 33.57 שניות עבור 7031 עמודים) ²² | מניפולציה PDF | PikePDF |
| איתית יותר מ- PyMuPDF | שילוף טקסט | איתית יותר (לדוגמה, 27.42 שניות עבור 7031 עמודים) ²² | שילוף טקסט | XPDF |

3. טכניקות מתקדמות לפרשנות שרטוטי חשמל

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

3.1 זיהוי אובייקטים וזיהוי סמלי חשמל

3.1.1 מודלים חדשניים: YOLOv8, Detectron2 וארכיטקטורות מיוחדות

זיהוי אובייקטים מודרני מנצל למידה עמוקה ורשתות קומבולוציוניות (CNNs) כדי לحلץ תכונות מורכבות, ובכך עולה על שיטות קומבנציונליות.²⁸

Detectron2 היא ספריית מחקר של Facebook AI Research לזיהוי אובייקטים, סגמנטציה וזיהוי חזותי.²⁹ היא תומכת ביכולת מתקדמות כמו סגמנטציה פנווטית, תיבות תוחמות מסובבות ו-Cascade R-CNN.²⁹ היא משמשת כספרייה לפרויקטים מחקרים ויישומי ייצור.²⁹

YOLOv8 היא משפחה של גלאי ירעה אחת הידועים ב מהירותם ובדוקם, שהושחררה בשנת 2023.²⁸ YOLOv8m (בינוני) השיג mAP50 של 96.7%, בעוד ש-YOLOv8s (קטן) השיג mAP50 של 96.5% עם זמן סקירה קצר משמעותית (2.3 מ"ש לעומת 5.7 מ"ש), מה שהופך אותו לטוב יותר עבור יישומים בזמן אמת.²⁸ הביצועים מושרים באמצעות מדדים כמו mAP50-90, mAP50, mAP, דיאק, היזכרות וערכי AP ספציפיים עבור גדי אובייקטים שונים (APs, APm, API).²⁸

3.1.2 אתגרים בזיהוי סמלי: טקסט מסובב, אלמנטים חופפים, רעש ומלים מותאמים אישית

מסמכים הנדסה, בין אם מדובר בתרשימי P&ID, שרטוטי ייצור או גילוונות CAD סרוקים, שונים מבחינה מבנית מסמכי משרד סטנדרטיים.³ טקסט בשרטוטים אלה מסובב לעתים קרובות, מוקטע, מוסתר על ידי קוים, או מסומן בסמלים ובהערות שכילים גנריים של OCR לא אומנו להבין.³ איצות היזהוי יורדת באופן משמעותי בerrickות ישנות ורעות, הנפוצות בפרויקטים הנדסיים.³ אלמנטים קריטיים כמו GD&T (Geometric Dimensioning and Tolerancing) או סמלי ריתוך מוחמצים לעיתים קרובות אלא אם כן המודלים מכילים במיניהם, מציבים אתגרים נוספים.³ טבלאות, וביעות בשרטוטים גדולים הדורשים חלוקה לאריחים, מציבים אתגרים נוספים.³

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

אימון מודלים חזקים לזיהוי אובייקטים דורש מערכי נתונים גדולים ומוסמנים הספציפיים לסמלים תרשימי חשמל. מערך הנתונים Digitize-HCD פותח במיוחד לצורכי דיגיטציה של תרשימי מעגלים בכתב יד.³⁰ הוא מכיל 1,277 תМОנות של תרשימים בכתב יד שצוירו על ידי למעלה מ-150 מתנדבים, ונsparkו ברזולוציה של 600 dpi.

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

ההצלחה של זיהוי אובייקטים עבור סמלים תלויות לחלוון בזמןות ובאיכות של נתונים אימון מסומנים ספציפיים לתחום. מערכי נתונים גנריים של תМОנות אינם מספיקים. בעוד שמערכות נתונים Digitize-HCD קיימים, יתכן שהם לא יcosו את כל הסמלים הספציפיים, הווריאציות או סגנונות הרשותן הקיימים במסמכים קניינים של ארגון (לדוגמה, מודפס לעומת כתב יד, סטנדרטים ספציפיים של חברה). זה מחייב הגדלת מערכי נתונים קיימים או יצירת חדשים לחלוון. תהליך אוצרת הנתונים המותאמת אישית והסימון הוא משימה משמעותית מבחינת זמן, עלות ומומחיות, אך זה צעד בלתי ניתן להשגת דיוק גבוהה בהקשר הנדסי אמיתי. ארגונים חייבים לתקצב לא רק משאבי מחשב ופיתוח מודלים, אלא גם באופן משמעותי לרבייה נתונים, סימון ותחזקה שוטפת של מערכי נתונים הספציפיים לתחום שלהם כדי להבטיח שמודלי ה-AI ישארו יעילם וניתנים להתקאה לשוגי מסמכים מתפתחים.

3.2 שליפת טקסט והערות עם הבנה קונטקטואלית

3.2.1 אופטימיזציה של OCR למסמכי הנדסה: מעבר לשלייפת טקסט גנרטיב

PaddleOCR היא מערכת כלים מבוססת למידה عمוקה לזיהוי, זיהוי וניתוח מבנה מסמכים, עם תמייה רב-לשונית.³ מודול V3 Structure שלה יכול לזהות ולנתח טבלאות ובלוקי כותרת, ו-PP5 OCRv5 מטפל בריבוי שפות ובכתב יד.³ היא מציעה את הניתוח המובנה ביותר מבין כל הקוד הפותח אך דורשת מאץ התקנה משמעותית.³ היא תומכת בספרית פייתון מקומית, שירות ענן ושירות עצמי.³²

(DocTR (Document Text Recognition) היא ספרית OCR בקוד פתוח שנבנתה בפייתון תוך שימוש בלמידה عمוקה (TensorFlow/PyTorch).³ היא עובדת היטב עם טקסט מסובב וקלה לשילוב.³ כל OCR גנריים (Tesseract, EasyOCR) מתאימים לשלייפת פסקאות מטפסים אך אינם

מספיקים לדיגיטציה של שרטוטי יצור בשל חוסר יכולתם להבין פריסה, לטפל בסמלים או לכבד את מבנה מסמכי ההנדסה.³

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

PaddleOCR-v5-PP ב-PaddleOCR מציג שיפור בזיהוי כתב יד³² עם זאת, PaddleOCR עדין מפספס סמלי T&GD או ריתוך רבים אלא אם כן הוא מכיל במיוחד.³ DocTR חסר תמייה מובנית בסמלים ונכשל בטבלאות מוטבעות.³ זה מדגיש את הצורך באימון מותאם אישית ובגישות היברידית. DocTR אינו מספיק אם לשרטוט יש טבלאות BOM.³ מודול StructureV3 של PaddleOCR מבטיח לזייהו וניתוח טבלאות.³

שם פתרון OCR "מהקופסה" אינו מספיק לשילפה מקיפה של טקסט והערות מشرطוטי הנדסתה חשמל מורכבים. פריסה "הפעל-ושכח" אינה אפשרית.³ השגת דיוק גבוהה עבור הערות הנדרשות קרייטיות כמו T&GD, סמלי ריתוך, או אפילו פורמטי BOM ספציפיים, תדרوش כיונון עדין משמעותית של מודלי OCR למידה עמוקה שנבחרו (לדוגמה, PaddleOCR) עם מערכyi נתונים מסומנים מותאמים אישית. זה מرمץ על מחזור חיים מתמשך של למידת מכונה, ולא פריסה חד פעמית. גישה היברידית המשלבת OCR מתקדם עבור טקסט וטבלאות כליל'ים עם מודלי זיהוי אובייקטיבים מיוחדים (סעיף 3.1) עבור סמלים, ואחריה ניתוח מבוסס כללים או סמנטי להבנה Kontekstualit, תהיה ככל הנראה הכרחית לכליאת כל המידע הרלוונטי בדיקוק. הפרויקט חייב להציג מושגים משמעותיים לאימון מתמשך של מודלי AI, אimoto, ואולי תהליכי שליפה רב-שלבי כדי להתמודד עם האתגרים הניאנסים של פרשנות מסמכי הנדסה.

טבלה 3: השוואת כלי OCR לשרטוטי הנדסה

| שרה תחתונה/התאמת | מגבילות לשרטוטי הנדסה | יתרונות לשרטוטי הנדסה | סוג | כלי |
|--|---|--|--|-----------|
| הניתוח המובנה bijouter מבין כל הקוד הפתוח, טוב לצורות טכניות המכוניות לכיוון עדין | מפספס סמלי T&GD/Rיתוך ללא כיונון עדין, מאיץ התקינה משמעותית | ניתוח מובנה (טבלאות, בלוקי cotract), רב- לשוני, כתב יד (PP- (OCRv5 מזהה פריסה | ערכות כלים לلمידה עמוקה (קוד (פתוח) | PaddleOCR |
| לא מתאימים לשרטוטים עם BOM או סמלים | מיועד למסמכים מבוססי טקסט (חשבוניות), נכשל בחתכים/טבלאות מוטבעות/סמלים, | מטפל בטקסט מסובב, איןגרציה כליה | ספרית למידה עמוקה (קוד (פתוח) | DocTR |

| שורה תחתונה/התאמתה | מגבילות לשרטוטי הנדסה | יתרונות לשרטוטי הנדסה | סוג | כל' |
|---|---|--|--|-------------------------|
| | לא תמייה מובנית בສמלים | | | |
| לא מתאימים לשליפת שרטוטים מלאים | נכשל בפרישות מורכבות, טקסט מוסובב, סמלים, סריוקות רועשות | טוב לテקסט פשוט, זמין באופן נרחב | מנוע OCR מסורתי (קוד פתוח) | Tesseract (OCR גנרי) |
| אופציה לבניונית מצקה לשיליפה אמינה מרטוטים סטנדרטיים | עтир משאבים, פחות ניתן להתקאה אישית, ביצועים ירודים בפרישות מותאמת אישית | טוב לשרטוטי מכניקה סטנדרטיים, זרימה חצי אוטומטית | פתרונות AI מסחרי | eDOCr (מחחרי/מיוחד) |

3.3 שחזור טופולוגיית תרשימים חשמלי

מעבר לשיליפת רכיבים וtekst בודדים, השלב הקritisי הוא הבנת אופן חיבורם של אלמנטים אלה- לייצרת מעגל חשמלי פונקציונלי. הדבר דרוש המרת נתוניים חזותיים לייצוג מבוסס גרפים.

3.3.1 שימוש תורת הגרפים לניתוח קשריות

תורת הגרפים היא ענף במתמטיקה המשמש באופן נרחב בניתוח רשותות לייצוג קשריות.³³ בטופולוגיה מעגליים, צמתי הרשת הם קודקודים תורת הגרפים, וענפי הרשת (חיבורים) הם קצוות הגרף.³³ תורת הגרפים שמשה בניתוח רשותות חזמיות מאז חוקי קירכהוף (1847) ועובדתו של מקסול (1873).³³ מושגים כמו מטריצות שכיחות ועצי פורש ישנים ישירות.³³ מורכבות מעגלים גדלה מחיבת קומבינטוריקה בתורת הגרפים לחישוביעיל במחשב.³³

3.3.2 דיהוי צמתים (רכיבים, צמתים) וקצוות (חיבורים, חוטים)

סמלי החשמל שחלצו (נדיגים, קבילים, מתגים, מנועים וכו') ונקודות החיבור (מסופים, צמתים, נקודות ניטרליות) שזוהו בשלבים הקודמים הופכים לקודקודים של הגרף.³⁵ חוטים, כבילים ונתיבי תמסורת המחברים רכיבים אלה יוצרים את הקצוות.³⁵ סוג המוליך (לדוגמה, תלת-פאייז, מסוכך) יכול להיות תכמה של הקצה.³⁵ התהילה כרך בדיהוי קרבה מרחבית ונקודות חיתוך בין מסופי רכיבים וקוואים כדי להסיק קשריות. זהה בעיה גיאומטרית מורכבת.

3.3.3 כלים לייצוג וניתוח גרפים (לדוגמא, NetworkX, Schemdraw)

NetworkX היא חבילת פיתון לייצרה, מיפוי וניהול המבנה, הדינמיקה והfonקציות של רשותות מורכבות.³⁷ היא מאפשרת הוספה צמתים וקצוות, בחינת תוכנות גרפים (צמתים, קצוות, סמיכות,

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

Schemdraw היא חבילת פיתון להפקת דיאגרמות סכמטיות של מעגלים חמליים באיכות גבוהה.³⁸ בעוד שUIKit יעדודה הואشرطוט, הגישה התכנית שלה להוספת אלמנטים וחיבורם (לדוגמה,

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

ספריות גיאומטריות כמו (JavaScript) D3.js מציאות יכולות לינדר קשתות, קוים וקישורים, ופריסות עבור גרפים מנענין כוח, שימושיות להציג טופולוגיה משוחזרת.³⁹ ספריית הסמלים של Nemeth⁴⁰ מספקת מסגרת קונספטואלית לייצוג סמלים מתמטיים, שניתן להרחבתה לסמלי חמם וחייבוריהם.

שלב זיהוי האובייקטים (3.1) מזהה אילו סמלים וטקסט קיימים. שלב שחזור הטופולוגיה עוסק בהבנה כיצד ישיות מזוהות אלה מחוברות פונקציונלית. זהו קפיצה סמנטית מפיקסלים/וקטורים לתרשים מעגל לוגי. האתגר טמון בזיהוי מדויק של נקודות חיבור (יציאות) על סמלים ובעקב אחר קוים/חותמים ביניהם, במיזח בשרטוטים רועשים או מורכבים. זה דורש ניתוח גיאומטרי מדויק ואולי "אלגוריתם פתרון מבוקר"³¹ כדי לעקוב אחר חיבורים בין סמלים מזוהים. התוצאה של שלב זה היא גרפ שבו צמתים הם רכיבים/צמתים וקצוות הם חיבורים, עם תכונות הנגזרות מטקסט שחולץ (לדוגמה, ערכי רכיבים, סוגים חותמים). גרפ זה הופך אז לנוטונים המובנים עבור בסיס הנתונים. זהו שלב המרכיב והקריטי ביותר, הממיר נתונים חזותיים גולםים לייצוג מעגל חמלי בעל משמעות סמנטית. הדיק שלו קובע את התועלת של המערכת יכולה לנתח וסימולציה במورد הזרם.

טבלה 4: סמלי חמם עיקריים ומיפויים לתקן IEC 60617

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

| שם סמל/תיאור | "יצוג חזותי" (דוגמה) | קוד תקן IEC 60617 אם (רלוונטי) | תכונות נפוצות | הערות/הוראות שימוש |
|--------------|---|---|--|----------------------------------|
| | | | | |
| 35 | סוג הארקה (הארקה כללית, הארקה פונקציונלית) | 60617- DB 1:2012 | (שלושה קווים ירדים בגודלים שוניים) | הארקה |
| 35 | מספר מסוף, ייעוד | 60617- DB 1:2012 | (עיגול קטן) | מסוף |
| 35 | סוג מנוע, הספק, מתה | 60617- DB 6:2012 | (מעגל עם M בפנים) | מנוע חשמלי |
| 35 | הגבר, סוג | 60617- DB 5:2012 | (משולש) | גבר |
| 35 | מצב (פתוח/סגור) | 60617- DB 7:2012 | (שני קווים עם רווח, קו שלישי מחבר) | מגע סגור (contact) |
| 35 | מצב (פתוח/סגור) | 60617- DB 7:2012 | (שני קווים עם רווח, קו שלישי חוצה) | מגע פתוח (Break) (contact) |
| 35 | ייעוד (לדוגמה, גנרטור) | 60617- DB 1:2012 | (שלושה קווים מתחברים לנקודה אחת) | נקודות ניטרל |
| 35 | סוג מוליך (כבל חשמל, 3 פאזהות) | 60617- DB 1:2012 | (קוויים מצטלבים) | חיבור (כללי) |
| התקן הגנה | זרם נקוב, מתה | 60617- DB 8:2012 | (מלבן עם קו גלי בפנים) | נתיר |

| שם סמל/תיאור | yczog chzoti (dogma) | קוד תקן IEC 60617 (אם רלוונטי) | תוכנות נפוצות | הערות/הוראות שימוש |
|--------------|------------------------|--------------------------------|----------------|--------------------|
| נורה | (עיגול עם צלב (בפנים)) | 60617-DB 8:2012 | הספק, מתח, סוג | התקן איות/תאורה |

4. ארגון נתוניים שחולצו לאינטגרציה עם בסיס נתונים

עיף זה מתמקד בהגדרת הסכימה ובהבטחת איכות ועקבות הנ吐נים שחולצו לתכנון חשמלי לצורך אחסון בסיס נתונים.

4.1 תכנון מודל נתונים מקייף למידע תכנון חשמלי

4.1.1 IEC 60617, IFC Electrical Domain (לדוגמה, **תekenim** בתעשייה) (Schema)

60617 IEC הוא תקן בינלאומי המגדיר סמלים גרפיים לדיאגרמות אלקטրוטכניות, ומכיל כ-1900 סמלים.⁴¹ הוא מכוסה מוליכים, רכיבים פסיביים, מוליכים למחצה, ייצור והמרת אנרגיה חשמלית, ציוד מיתוג, בקרות והתקני הגנה, מכשרי מדידה, מנורות והתקני אינטוט, וכן תוכניות התקינה אדריכליות וטופוגרפיות.⁴¹ הסמלים שימושיים בפורמטים GIF, DWG ו-EPS, עם מטא-נתונים כמו שם, שימוש, מילוט מפתח וקישורים.⁴¹ מנתו לשנה עולה 690 פרנקים שוויצריים, עם שניים עוקבות ב-270 פרנקים שווייצריים.⁴¹

IFC Electrical Domain Schema (IFC 4.3.2) הוא חלק משכבות הדומין של מודל IFC, ומרחיב את הרעיון הנוגע לשירותי בניין.⁴² הוא מגדיר מושגים של מערכות כבלים עבור חשמל, תאורה, נתונים, תקשורת, אבטחה ובקרה.⁴² הוא מכסה מכשירי חשמל, מנועים, נושא כבלים (צינורות, מגשי כבלים),لوحות חלוקה, גנרטורים, קופסאות חיבורים, גופי תאורה, שקעים, התקני הגנה ושנאים.⁴²

היקף ומוגבלות של IFC Electrical Domain: הוא מיועד בעיקר להתקנות חשמל במתוח נמוך (12V עד 24V DC) או 1000V AC/DC. מתח ביןוני/גובה, מערכות מתחת-ל-12V, אספקת חשמל ציבורית ומצבים חולפים אינם בתחום ההיקף או שאינם מפורטים במלואם.⁴² מעגלים חשמליים מוגדרים כתמ"ב באמצעות אובייקט **ElectricalPanel**.

את IfcDistributionSystem.PredefinedType::IfcDistributionSystemEnum⁴², ומחליפים את

שליפת טקסט גולמי ותיבות תומחות אינה מספקת. כדי להפוך את הנתונים לבעלי ערך אמיתי ונitinim להפעלה הדידית, יש למפות אותם לתקנים תעשייתיים מבוססים כמו IFC ו-60617 IEC.

הדבר מבטיח שהגדרה מוכרת. הקפדה על תקנים אלה מלכתחילה מקלת על שילובם עם תוכנות BIM/BIM/AEC אחרות, מאפשרת שאלות סטנדרטיות, ומבטיחה את עתיד הנזונים עבור יישומים רחבים יותר (לדוגמה, סימולציה, ניהול מתקנים, תאומים דיגיטליים). התעלמות מתקנים אלה תוביל למסד נתונים קנייניים ומובודד שקשה לשתף, לתחזק ולשלב במערכות אקוולוגיות הנדסיות גדולות יותר, ובכך תפחת משמעויות את החזר ההשאה של מוצר השילפה. לפיכך, שלב מידול הנזונים אינו רק ממשימה

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

4.1.2 יציג רכיבים, תוכנות, חיבורים ו מידע מרחב

- **רכיבים:** כל סמל חשמלי שזוהה (לדוגמה, נגד, מtag) הופך לרשומה, עם תוכנות כמו סוג, שם, ערך, יצורן ומזהה ייחודי.
- **תוכנות:** מידע טקסטואלי שחולץ (לדוגמה, "Motor_M110", "220V", "Q") מנוטח ומוקצה כתוכנות לרכיבים או לחברים המתאימים.
- **חיבורים:** הטופולוגיה הגרפית המשוחזרת (צמתים וקצוות) מהוות את הבסיס לייצוג קישוריות. כל קצה מייצג חוט/חברור, עם תוכנות כמו סוג חוט, אורך (אם ניתן לחישוב מקנה מידת השרטוט) ומיגל קשור.
- **מידע מרחב:** קוודינטות תיבת תוחם עבור סמלים וטקסט, ונתיבים גיאומטריים עבור קוואים/חותמים, נשמרים כדי לשמר את הפריסה המרחבית עבור הדמיה או ניתוח גיאומטרי נוסף.
- **מבנה היררכי:** יציג רכיבים מוקנים, תת-מעגלים וקובוצים 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).
- **عقبיות:** סטנדרטיזציה של מוסכמות שמורות, ייחדות וייצוג סמלים על פני כל הנתונים שחולצו, באופן אידיאלי מיפוי ל-60617-4 IEC.⁴¹
- **ניהול גרסאות:** יישום מנגנים למעקב אחר שימושים בשרטוטים ובנתונים שחולצו מהם לאחר זמן, המאפשרים ניתוח היסטורי וביקורת. זה קריטי בהתחשב בכך ש"תוכניות הפרויקט מתפתחות ללא הרף".¹

4.2 שיקולי תכנון סכימת בסיס נתונים למדרגיות ויכולת שאילתת

- **בחירה סוג בסיס נתונים:**
 - **בסיס נתונים יחסית (SQL):** מתאים לנ נתונים מבניים, טבלאים (לדוגמה, תוכנות רכיבים, BOMs). טוב לשאלות מורכבות והבטחת שלמות נתונים עם תנאים ACID.
 - **בסיס נתונים גרפי (NoSQL):** אידיאלי לייצוג ושאלת יחסים (לדוגמה, טופולוגיה מעגלים, קישוריות). מצטיין במערכות וניתוח רשותות.
 - **גישה היברידית:** שילוב יחסית לתוכנות רכיבים וגרפי לקישוריות עשוי להציע את הטוב משני העולמות, הדורש אינטגרציה זהירה.
- **נורמליזציה לעומת דה-נורמליזציה של סכימה:** איזון ביצועי שאילתת עם שלמות נתונים ויעילות אחסון.

- **אסטרטגיית אינדוקס:** אופטימיזציה של ביצועי בסיס נתונים עבור שאילותות נפוצות (לדוגמה, מציאת כל הרכיבים במעגל ספציפי, מעקב אחר חיבורים).

- **אינדוקס מרחב:** לשאלתה עיליה של רכיבים בתוך אזור ספציפי של השרטוט.

5. ארכיטקטורת צינור שליפת הנתונים

סעיף זה יתאר את ארכיטקטורת צינור הנתונים מקצה להנדרשת לאוטומציה של תהליך השליפה, הטרנספורמציה והטיענה.

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

צינורות נתונים במקרים מסוימים מאפשרים אוטומציה של העברת נתונים ממקורות שונים ליעד, כולל טרנספורמציה, אימות איות וניהול.⁴⁴ הם מהווים את "מערכת הדם" של מערכות הנתונים בארגון.⁴⁴

ETL (Extract, Transform, Load) היא גישה מסורתית שבה הנתונים עוברים טרנספורמציה לפני הטיענה ליעד.⁴ גישה זו הייתה דומיננטית בעברית Hadoop.⁴

ELT (Extract, Load, Transform) היא גישה שבה הנתונים נשלפים, נתונים מיד ליעד (לרוב מחסן נתונים בענן), ולאחר מכן עוברים טרנספורמציה במקומם.⁴ גישה זו זוברת פופולריות מאז 2017, ומיציאה יותר שליטה, גמישות, מהירות חישוב גבוהה יותר ועלויות מופחתות עבור אנליטיקה מתקדמת.⁴

Streaming ETL מעבד ומבצע טרנספורמציה נתונים במהלך התנועה לפני שהם מגעים לאחסון, ובכך מקצר את זמן ההגעה לתובנות ומזער את עלויות האחסון.⁴⁴ גישה זו אידיאלית עבור אנליטיקה בזמן אמת והתראות תפעוליות.⁴⁴ קיימות גם

ארQUITקטורות היברידיות (Lambda/Kappa) שהן מורכבות יותר ומשלבות עיבוד אצווה וסטריימינג עבור ניתוח מקיף ותצוגות בזמן אמת.

5.2 שלבים מרכזיים: קליטה, עיבוד מקדים, שליפה, טרנספורמציה, אימות וטיענה

- **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 קטורי או רסתרי.
- **נורמליזציה:** סטנדרטיזציה של פורטטקי קבצים, טיפול בגרסאות שונות של DWG/PDF.
- **שיפור תמונה (עבור קבצי PDF רסטר):** הפחתת רעש, יישור, התאמת ניגודיות לשיפור דיק OCR.
- **חיתוך לאריחים (עבור שרטוטים גדולים):** פירוק תמונות גדולות למקטעים קטנים וקלים יותר לעיבוד על ידי מודלי למידת מכונה.³
- **שליפה (Extraction):**

- **שליפת נתונים DWG:** שימוש ב-DWG RealSDK או ב-SDKs ODA לניתוח שכבות, בלוקים, תכונות וישיות גיאומטריות.¹¹
- **שליפת טקסט מ-PDF (וكتורי):** שימוש PyMuPDF או PDFMiner לטקסט ומידע פריסה בסיסי.²²
- **שליפת נתונים מבוססת תמונה (רטסט/סורך):** יישום OCR מתקדם לטקסט זיהוי אובייקטים (YOLOv8, Detectron2, PaddleOCR) לסמלים.³
- **טרנספורמציה (Transformation):**
 - **מיפוי סמלים:** זיהוי סמלים שחולצו ומיפויים לייצוג IEC 60617 סטנדרטיים.³⁵
 - **ניתוח תכונות:** שליפה וארגון של ערכים מספריים, תוויות והערות טקסטואליות אחרות.
 - **שחזור טופולוגיה:** בניית ייצוג הגרף של המעגל החשמלי (צמתים וקצוות) מרכיבים שזוהו וחיבורם שנעקבו.³³
 - **הקשר:** קישור נקודות נתונים שחולצו ליצירת ישיות בעלות משמעות (לדוגמה, קישור סמל רכיב לערכו, ולאחר מכן לחברו).
- **אימות (Validation):**
 - **בדיקות איות נתונים:** הבטחת דיווק, שלמות ועקביות הנתונים שחולצו.² זה כולל אימות סכימה, הפניות צולבות וכליים ספציפיים לתחום (לדוגמה, הבטחה שלמתג יש לפחות שני חיבורים).
 - **טיפול בשגיאות:** זיהוי והסגרת נתונים שגויים, הפעלת התראות, ואפשרות להפעלת תיקון אוטומטי.²
- **טעינה (Loading):** אחסון הנתונים המובנים והמאומתים בסיס הנתונים היעד (יחס, גרפי או היברידי).

5.3 שיקולי מדרגות וביצועים עבור פעולות רחבות היקף

- **נפח ומהירות:** תכנון הצינור לטיפול בכמות עצומות של נתונים ולבידודם ביעילות.²
- **אופטימיזציה של משבבים:** הקצהה יעילה של משבבי מחשב, כולל CPU ו-GPU, עבור שלבים שונים.²
- **עיבוד מקבילי:** ניצול מסגרות מחשב מבוזרות לעיבוד מספר שרטוטים במקביל.²
- **עיצוב מבוסס ענן:** ניצול שירותי ענן לגמישות ומדרונות (לדוגמה, פונקציות ללא שרת, בסיסי נתונים מנוהלים, מחשוב בקנה מידת אוטומטי).

5.4 טיפול בשגיאות, הבטחת איות נתונים ותיקון אוטומטי

- **בדיקות איות נתונים אוטומטיות:** יישום כללים ואלגוריתמים לזיהוי אוטומטי של חריגות, חוסר עקביות נתונים חסרים.²

- **ניהול סחף סכימה:** מנגנוןים להתקאה לשינויים בפריסות מסמכי המקור או במבנה הנתונים, הפנית נתונים אוטומטית ל"אזור הסגר" והפעלת סקריפטים לאימות.⁴
- **nitro ויכולת צפיה:** איסוף מדדים על תפקה, תדריות שגיאות, ניצול משאבים ומדד איקות נתונים לזיהוי בעיות באופן יוזם.⁴
- **מנוע תיקון:** הגדרת פרוטוקולי תגובה לרוחישי כשל נפוצים, כולל פוטנציאלית עיבוד חדש אוטומטי או סימון לבדיקה ידנית.⁴

בהתחשב במורכבות ובשונות של שרטוטי הנדסה, שגיאות וחוסר עקביות הם בלתי נמנעים במהלך אוטומטיזציה. ללא הבטחת איקות נתונים חזקה ואוטומטית וטיפול בשגיאות, בסיס הנתונים המובנה יփוך במהירות לבלי אמין, ויגע בערכו. התערבות ידנית עבורה כל שגיאה אינה ניתנת להרחבה. לכן, הצורך חייב לכלול מנגנון חכמים לתיקון עצמאי, הסגרת נתונים והתראות אוטומטיות לביעות הדרישות בדיקה אנושית (לדוגמה, "סחף סכימה" כאשר פריסות מסמכים משתנות). זה חורג מרישום שגיאות פשוט; מדובר במבנה "מערכת ניהול עצמאית"⁴ המסתגלת לדרישות משתנות ושומרת על איקות נתונים חזקה, זהה ולהיליך מתמשך, לא הגדרה חד פעמית. השקעה בניהול נתונים אוטומטי, בדיקות איקות ויכולות תיקון בתוך הצינור היא קריטית כמו טכנולוגיית השילפה עצמה. היא מבטיחה את אמינות ותועלת הנתונים שהולצו לטוח אורך, ומשפיעה ישירות על החזר ההשקעה ועל אימוץ המשתמשים.

6. שיקולי יישום ותוכנן משאבים

סעיף זה יתייחס להיבטים המעשיים של בניית הפתרון, כולל בחירת טכנולוגיות, משאבי מחשב ואומדי עליות.

6.1 ערכות פיתוח תוכנה (SDKs) וספריות: סקירה מפורטת

- **ניתוח DWG/DXF:** הבחירה בין Autodesk RealDWG (עלות גובהה, רשמי, סעיף "לא תחרותי"), ODA Drawings SDK (עלות גובהה, מקייף, מבוסס חברות, היסטוריה משפטיות קודמת), ו-ezdxfs בקוד פתוח (חינם, תמייה מוגבלת ב-DWG, ללא OLE).⁸
- **שלילת PDF:** PyMuPDF (ביצועים גבוהים, גרסת Pdf מSchedulerית לתוכנות מתקדמות) ו-PDFMiner (טוב לטקסט, מתקשה בפריסות מורכבות).²¹
- **OCR וזיהוי אובייקטים:** PaddleOCR (ניתוח מובנה, כתוב יד, דורש כיוון עדין), DocTR (טקסט מסובב, מוקדק טקסט, חסר תמייה בסמלים), YOLOv8 (מהיר, mAP גבוהה), Detectron2 (חידש, תיבות תוחמות מסובבות).³
- **עיבוד גרפים:** NetworkX (ספרית פיתון ליצירת וניתוח גרפים).³⁷
- **ספריות נספחות:** OpenCV לעיבוד תמונת, Pillow לManipulation תמונת³¹, ופוטנציאלית ספריות גיאומטריות מותאמות אישית למעקב קווים והסקת חיבורים.

6.2 משאבי מחשב: תשתיות GPU מקומית לעומת ענן

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

מודלי תמחור GPU בענן:

- **AWS:** מציעה סוג GPU מופעי שונים (P2, P4d, P4de, G5) עם תמחור שעתית הנע בין -On 45 מודלי התמחור כוללים \$40.97 (P4de.24xlarge) ל-c-77 \$0.752 (g4dn.2xlarge).⁴⁵ עד 90% הנחה אך ניתנים להפרעה, Demand, Savings Plans, Spot Instances ו-45. גורמים המשפיעים על העלות כוללים סוג Dedicated Hosts- Reservations המופע/GPU, אזור, משך השימוש ועלויות רישיון תוכנה.⁴⁵
- **Google Cloud:** מציעה אפשרויות GPU כמו H100, NVIDIA T4, P4, A100, L4-L. 47 מחירים GPU שעתיים נעים בין c-35 (\$) ל-100/H100/A100. 48 מספקת הנחות לשימוש מתמשך ולהתחייבות לשימוש. 48 מציאות הנחות של 60-91% אך ניתנות להפרעה.⁴⁸
- **RunPod:** מציעה GPUs חסכוניים עבור צוותי ML/AI עם חיבור לפי שנייה (החל מ- \$0.00011 לשניה או \$0.39 לשעה עבור L4) ומונויים חדשניים צפויים.⁵⁰ מספקת סוג GPU שונים (H200, B200, H100, A100, L40S, RTX 6000 Ada, A40, RTX 3090, RTX 4090, RTX A5000 vCPUs ו-VRAM משתנים.⁵⁰ מציעה סוג עובדים "Flex" (הגדלה בשיא תעבורה) ו-"Active" (פעולים תמיד עם הנחה).⁵⁰
- **Azure Machine Learning:** אין חיבור נוסף עבור שירות Azure ML עצמו, אך קיימים חיבורים נפרדים עבור משבבי מחשב בסיסיים (VMs, GPUs), אחסון (Azure Blob Storage) ושירות Azure אחרים.⁵¹ מציעה סדרות VM שונות (D, DS, Dds, RAM vCPU ו-vRAM) עם תצורות חישוכן (תוכניות חישוכן/שמורות לשנה או 3 שנים המציגות הנחות משמעותיות).⁵¹
- **шиקולים מקומיים (On-Premise):** הוצאות הון ראשוניות גבוהות עבור חומרה, עלויות תחזוקה שוטפות, קירור וחימול. פחתות גמיש להגדלה או הקטנה בהשוואה לענן.
- אימון למידת מכונה (במיוחד עבור מודלים מותאמים אישית וכיוונן עדין) הוא עתיר משבבים ונitin להפעלה לפרקם. מופעי Spot או עובדי "RunPod" (RunPod) Flex יכולם להפחית משמעותית עלויות עבור עבודות אימון לא קרייטיות וניתנות להפרעה. עבור עומסי עבודה של הסקה מתמשכת או שלבי צינור קרייטיים, עובדי "RunPod" Active או מופעים שמוריים/תוכניות חישוכן (AWS, Google Cloud, Azure) מציעים יכולת חיזוי עלויות והנחות על פני תמחור On-Demand. בחירת ספק הענן וסוג המופע הספציפי תלויה בדרישות VRAM של המודלים, מהירות ההסקה הרצiosa והתקציב. GPUs עם VRAM גבוהה יותר (לדוגמה, H100, A100) יקרים יותר אך יכולים לטפל במקרים גדולים יותר או בגגלי אצווה. אסטרטגיית ענן מתוכנת היבט, המנצלת מודלי תמחור שונים וסוגי מופעים עבור מאפייני עומס עבודה משתנים (אימון לעומת הסקה, לפרקם לעומת מתמשך), היא קרייטית ליעילות עלות ולמדרגיות. זה דורש ניטור מתמשך של ניצול משבבים ואופטימיזציה של עלויות.

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

| הערות | אפשרויות חישוב בעלות | עלות שעשית לפי דרישת משוערת) | מפרט GPU עיקרי VRAM), vCPUs (| סוג מופע GPU לדוגמה | ספקן |
|---|---|--|-------------------------------|-----------------------|--------------|
| מופעים ניתנים להפרעה (Spot), דרוש התחייבות Reservation), Savings (Plans | Spot Instances, Savings Plans, Reservation s, Dedicated Hosts | \$32.77/ ⁴⁵ שעיה | GB 96 VRAM, 96 vCPUs | P4d.24xlarge | AWS |
| מופעים ניתנים להפרעה (Spot), הנחות לשימוש מתמשך/מחזיב | Spot VMs, Sustained Use Discounts, Committed Use Discounts | משתנה (תלוי אזור, התחייבות) ⁴⁸ | GB 80 VRAM | NVIDIA H100 (A3 High) | Google Cloud |
| חייבים נפרדים עבור שירות Azure ועודפים, הנחות משמעויות עם התחייבות | Savings Plans, Reserved Instances | משתנה (תלוי אזור, התחייבות) ⁵¹ | GB 96 VRAM, 96 vCPUs | ND96asr_v4 | Azure |
| חייב לפחות שנייה, עובדי Flex קצרות, עובדי Active למשימות מתמשכות | Pay-per-second, Flex Workers, Active | \$2.79/ ⁵⁰ שעה | GB 94 VRAM, 16 vCPUs | H100 NVL | RunPod |

| | | | | | |
|-------|-----------------------------|--|--|---------------------------|-----------|
| הערות | אפשרויות חישוב בעליות | עלות שעתית לפי דרישה (משוערת) | מפרט GPU עיקרי VRAM) , vCPUs (| סוג מופע GPU לדוגמה | ספק ען |
| | Workers (ען הנחה) | | | | |

6.3 אומדן עלויות פיתוח ותפעול

עלויות פיתוח:

- **כוח אדם:** משכורות ל מהנדס ל מידת מכונה, מהנדס נתוניים, מפתחי תוכנה, מומחי תחום (מהנדס חשמל), מסמכי נתוניים.²
- **רישיונות תוכנה:** עלויות עבור SDKs קנייניים (RealDWG, ODA Drawings SDK, PyMuPDF Pro, OCR מסחריים, רישיונות בסיסי נתוניים).
- **רכישת סימון נתונים:** עלויות הקשורות להכנות מערכו נתונים מותאמים אישית אם הציבוריים אינם מספיקים.³⁰

עלויות תפעול:

- **מחשב ענן:** עלויות שוטפות עבור מופעי GPU (אימון, הסקה), מופעי CPU, פונקציות ללא שרת.
- **אחסון:** עלויות אחסון שרטוטים גלובליים, נתונים מעובדים ביניים ובסיום הנתונים המובנה הסופי.
- **העברה נתונים:** עלויות כניסה/יציאה להעברת נתונים לען וממנו.
- **תחזוקה ונטור:** עלויות לתחזוקת צינורות, אימון מודלים חדש וכל' ניטור.

בעוד שרישיונות SDK ומחשב ענן הם עלויות מוחשיות, מרכיב עלות עיקרי, שליעיתים קרובות אין מושך מספיק, הוא המומחיות האנושית הנדרשת. זה כולל לא רק מהנדס ל מידת מכונה/נתוניים אלא גם מומחי תחום חשמל לשימון נתונים, אימונות מידע שholz והנחהית פיתוח מודלים. האופי האיטרטיבי של פיתוח ל מידת מכונה, כולל איסוף נתונים, סימון, אימון מודלים וכיונון עדין, מחייב השקעה מתמשכת באנשי מקצוע מיומנים אלה ובתהליכי אציגת הנתונים הקשורים. זו אינה הוצאה חד פעמי אלא עלות תפעולית מתמשכת. ארגונים חיברים לקחת בחשבון עלויות "נסתרות" אלה של כישרונות מיוחדים ועידן מתמשך של צינור הנתוניים בתקציב שלהם, מכיוון שהם קבועים באופן קרייטי את הצלחת הפרויקט ואת הדיקוק וההסתגלות לטווח ארוך של המערכת.

6.4 מומחיות צוות ודרישות כוח אדם

צוות ליבנה:

- **הנדסי למידת מכונה:** לבחירת מודלים, אימון, כיווןן עדין (YOLOv8, Detectron2, PaddleOCR) ופריטה.
- **הנדסי נתונים:** לתכנון ובניית צינור הנתונים, ניהול קליטת נתונים, טרנספורמציה וטעינה, והבאתה מדרגות ואיכות נתונים.
- **פתחי תוכנה:** לשילוב SDKs (מנתחי PDF/DWG), פיתוח לוגיקה מותאמת אישית ובניית ממשקי API לצריכה נתונים.
- **הנדסי חשמל/מומחי תחום:** קרייטיים לשימון נתונים, אימות מידע שחולץ, הגדרת מודלי נתונים ומטען ידע מומחה לזרוי סמלים ו恢復 טופולוגיה.
- **תקידי תמיכה:** מהנדס DevOps לניהול תשתיית, מומחי QA לבדיקות, ייעוץ משפטי לסקירת קניין רוחני.

7. אבטחת נתונים, פרטיות ותאיימות

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

7.1 הגנה על שרטוטי תשתיות רגישיים: תקנים ושיטות עבודה מומלצות

עקרונות אבטחת נתונים: שמירה על סודיות, שלמות וזמן נתונים באופן עיקרי עם אסטרטגיית הסיכונים של הארגון.⁵²

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.⁵³

מערכת ניהול אבטחת מידע (ISMS): עמידה במדיניות ISMS (לדוגמה, ISO 27001) חיונית לניהול סיכון אבטחת מידע והבטחת סודיות, שלמות וזמן נתונים.⁵³

ארכיטקטורת אבטחה: ארכיטקטורת אבטחה חזקה כוללת פילוח רשת, חומות אש, מערכות זיהוי חדרה ופרוטוקולי הצפנה לביצוע הרשות מפני פרצות חיצונית.⁵³

7.2 יישום בקרות אבטחה חזקות: הצפנה, בקרת גישה, פילוח רשת

- **הצפנה:** חיונית לאבטחת נתונים בזמן אחסון (במנוחה), העברה (בעבר) ועיבוד. זה הופך את הנתונים לבטחי 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 למשאים קרייטיים.⁵³ פתרונות ניהול זהויות וגישה (IAM) מייעלים את ניהול המשתמשים ואוכפים מדיניות אבטחה.⁵³
- **פילוח רשת:** בידוד נכסים קרייטיים והגבלה תנואה רוחבית של אינוי סיבר בתוך הרשות.⁵³
- **ניהול פגיעות:** זיהוי וטיפול קבוע בפגיעה כדי למנוע פרצות אבטחה פוטנציאליות.⁵³
- **גיבויים קבועים:** קרייטיים להגנת בסיסי נתונים והתאוששות מאירועי שחיתות נתונים.⁵²

7.3 התייחסות לחששות קניין רוחני עבור נתונים שחולץ

- **קניין רוחני של קובץ המקור:** קבצי DWG הם קניינים של Autodesk, והשימוש בהם כפוף לרישיונות של Autodesk.⁵ 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, גրפי קישוריות) מייצגים טרנספורמציה. יש לשקל בזרירות את הבעלות והשימוש המותר בנתונים נגזרים אלה.
- **רישוי SDKs:** תנאי DWG ("ישומים לא תחרותיים") והסכם חברות ב-ODA⁹ משפיעים ישירות על מה שניתן לבנות ולהפיץ באמצעות ה-SDKs שלהם. שימוש ב-SDKs אלה ליצירת מערכת המתחירה בתוכנת CAD המקורית עלול להוביל לביעות משפטיות.
- **תאימות לחוקי קניין רוחני:**EDA דא שתהlixir שליפת הנתונים והשימוש הבא בbasis הנתונים המובנה אינם מפרים זכויות קניין רוחני כלשהן של יצריו השרטוטים המקוריים או ספק' התוכנה.

פעולה שליפת נתונים מקבצי DWG קניינים, גם לשימוש פנימי, קיימת למרחב משפטי מעורפל. בעוד שהקובץ הגלומי מוגן, ה"עובדות" הכלולות בו (לדוגמה, סוג רכיבים, חיבורים) בדרך כלל אין מוגנות בזכויות יוצרים. עם זאת, שיטת השלייפה וצורת הנתונים הנגזרים עשויים להיות כפפים לבדיקה המשפטי, במיוחד אם היא הכרוכה בהנדסה לאחרור של פורמטים קניינים או אם מסד הנתונים המובנה שנוצר יכול להיחשב כמתחרה בהצעות של Autodesk. ארגונים חייבים לעرب יעוץ משפטי המתמחה בקניין רוחני ורישיון תוכנה בשלב מוקדם של פרויקט כדי להעיר סיכונים, להבין את היקף השימוש המותר בשלייפת נתונים וחשיבותם, ובנوات הסכמים עם ספק' (Autodesk, ODA) כדי למתן אתגרים משפטיים עתידיים. התעלמות מחששות קניין רוחני עלולה להוביל להתקדיינות, צוו מניעה או ארגון חדש כפוי של הפטון, ובכך לסכן את כל ההשקעה. אסטרטגיה משפטית יזומה חיונית כמו הארכיטקטורה הטכנית.

טבלה 6: מסגרות ובקורות אבטחת נתונים עבור נתוני הנדסה וגישים

| השלכה | בקורות רלוונטיות לנתוני הנדסה | עקרונות מפתח | מסגרת/תקן |
|---|--|-------------------------------------|------------------------------|
| mbtich תאיימות לדרישות רגולטוריות ומגן מפני איומי סייבר. ⁵³ | בקורת גישה (MFA, RBAC, Least Privilege ⁵³), הצפנה (נתונים במנוחה, נתונים בעבר) ⁵³ , אבטחת רשות (פילוח רשות, חומות אש, IPS/IDS) ⁵³ , ניהול פגיעות (סריקה קבועה, תיקונים, בדיקות חדירה) ⁵³ , גיבוי ושחזור נתונים (גיבויים קבועים ונבדקים, תוכנית התאוששות מס摊) ⁵² , רישום וניהול ביקורת (יוםניהם מקיפים של גישה ושינויים בנתונים, ניטור בזמן אמיתי) ⁴ , אבטחת שרשת אספקה | זיהוי, הגנה, זיהוי, תגובה, התאוששות | NIST Cybersecurity Framework |

| השלכה | בקרות רלוונטיות לנוטוני הנדסה | עקרונות מפתח | מסגרת/תקן |
|--|--|--------------------------|---------------------|
| | (בדיקות SDKs צד שלישי ווסףי ענן). ¹⁴ | | |
| מספק גישה מבנית לניהול סיכון אבטחת מידע. | (כג"ל בקרות NIST, מישומות במסגרת מערכת ניהול) | סודיות, שלמות, 贊明度 | ISO 27001 (ISMS) |
| רישימת פעולות עדיפות להפחתת הסיכון לאירוע סייבר. | (כג"ל בקרות NIST, מישומות בקרות 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 מדרוגת

- שלב 1: הוכחת היתכנות (PoC):** התמקד בתת-קובוצה קטנה וייצוגית של מסמכים (לדוגמה, 10-20 שרטוטים) כדי לאמת את טכנולוגיות השיליפה הליבתית (nitro, DWG/PDF, OCR בסיסי/זיהוי אובייקטיבים) ולהציג את היתכנות. תעדר סוג מסמר אחד (לדוגמה, DWG מקור או PDF סרוק) תחילתה.
- שלב 2: מוצר מינימלי בר-קיימא (MVP):** הרחב למערך נתונים גדול יותר ויישם את צינור הנתונים מקצה לעבר קבוצת ליבה של רכיבי וחיבוריו חממל. התמקד בהשגת דיקון מקובל עבור נקודות נתונים קritisיות ובניאת בסיס הנתונים המובנה הראשוני.
- שלב 3: שיפור והרחבה איטרטיביים:** שפר באופן מתמיד את דיקון המודל באמצעות אימון מחדר עם נתונים נוספים, הרחב את כיסוי הסמלים והיתכנות, שפר את שחזור הטופולוגיה, ושלב עם מערכות ארגוניות קיימות (לדוגמה, ERP, BIM).

8.2 אימון מודלים מותאם אישית ואסטרטגיות שיפור מתמיד

יש לחזור ולהציג כי מודלי OCR/זיהוי אובייקטיבים "מהמדד" אינם מספקים לשרטוטי הנדסה ודוחים כיוון עדין משמעותית עם נתונים ספציפיים לתחום.³ יש לישם מערכות שבhn מומחים אנושיים בודקים את תוכרי המודל, מתקנים שגיאות ומספקים הערות חדשות, אשר מזונות בחזרה ללולאת האימון כדי לשפר באופן מתמיד את ביצועי המודל. יש להרחיב באופן מתמיד את מער

הנתונים על ידי איסוף וסימון וריציאות חדשות של שרטוטים כדי לשפר את חווון המודל ויכולת ההסתגלות לנסיבות מסוימות מתקדמות.

8.3 שילוב עם זרימות עבודה קיימות של AEC ומערכות BIM

נתוני החשמל המובנים יכולים להעשיר מודלי BIM, ולספק מידע מפורט על מערכות חשמל שאולן אינן מודל בפורש 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 תעופליות. טופולוגיה החשמל המשוחזרת יכולה לשמש שירות לחישובי עונס חשמלי, סימולציות מעגלים וניתוח תקלות. ניצול הנתונים המובנים עבור אומדי כנויות חומריים מדוקים ומהירים ביותר.¹

8.4 טכנולוגיות מתקדמות וחזון לטוויה ארוך

- גרפי ידע:** יציג נתונים החשמל שחולצו כגרף ידע יכול לאפשר שאלות מורכבות יותר, הסקה סמנטיבית ו קישור עם ידע תחום אחר.
- תאומים דיגיטליים:** הנתונים המובנים מהווים שכבה בסיסית ליצירה ותחזוקה של תאומים דיגיטליים של מערכות חשמל, המאפשרים ניתוח בזמן אמיתי, תחזוקה חזויה ותכנון תרחישים.
- בינה מלאכותית גנטטיבית לאוטומציה של תכנון:** בטוויה הארוך, היכולת לשלוフ ולהבין נתונים תכנון חשמל יכולה להזין מודלי AI גנטטיבים לתכנון ואופטימיזציה אוטומטיים.
- אבלוציה של סטנדרטיזציה:** הישארות מעודכנת בהתקפות IFC (לדוגמה, IfcElectricalDomain ותקנים תעשייתיים אחרים כדי להבטיח יכולת פעולה הדדית לטוויה ארוך.

9. מסקנה

דו"ח זה פירט גישה מבוססת ראיות לשילוף נתונים תכנון חשמל מבנים מקבצי PDF ו-DWG, תוך הדגשת המרכיביות הטכניות, השיקולים האסטרטגיים והפוטנציאלי להשפעה טרנספורטטיבית. על ידי ניוט זהיר בפורמלים קנייניים, ניצול AI מתקדם לפרשנות חזותית, שחזור טופולוגית מעגלים באמצעות תורת הגראפים, ובנויות צינור נתונים חזק ומאובטח, ארגונים יכולים לפתח תובנות יקרות ערך מארכינוי ההנדסה שלהם. ההשקעה ב-SDKs מיוחדים, תשתיות GPU בענן, ובאופן קרייטי, במומחיות AI ספציפית לתחום וביצירת נתונים, תניב תשואות שימושיות בייעילות תעופלית, יכולות החלטות ויתרון תחרותי, ובכך מסלול את הדרך לעתיד חכם ואוטומטי יותר בהנדסת חשמל ובניה.

מבחן: שיליפת נתונים מתכון חשמל מ-PDF ו-DWG להקמת בסיס נתונים מבנה

**עיקרי הממצאים (Executive Summary)

: מה אפשרי היום:

- **PDF וקטורי**: חילוץ טקסט, קווים וטבלאות עם דיוק של 85-95% באמצעות כלים כמו `pdfplumber`.
- **OCR סרוק**: OCR עם דיוק 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 מודול.
- **מיפוי לבסיס נתונים**: המרת SQL/PostgreSQL עם סכמת ERD ברורה.

: מה לא אפשרי:

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

עלות, זמן וסיכון למסלולים עיקריים:

| מסלול | עלות (איש-חדש) | זמן פיתוח | סיכון | הערות |
|----------------|-----------------------|-----------|--------|-----------------------|
| PDF וקטורי + 8 | Tesseract** \$15K | שבועות | ביןוני | דורש אימון OCR לשפות. |
| OCR סרוק + 12 | PaddleOCR** \$25K | שבועות | גובה | תלי באיכות הסריקה. |
| --- | DWG + ODA SDK** \$500 | שבועות | גובה | רישי ODA/שנה. |
| --- | (PDF+DWG)** \$50K | --- | --- | --- |
| --- | --- | --- | --- | --- |
| --- | --- | --- | --- | --- |
| --- | --- | --- | --- | --- |

סקירת טכנולוגיות (Landscape)

: כלים מומלצים:

מוגבלות קרייטית:

- **ODA SDK** : לא מעבד Blocks דינמיים אוטומטיות.

Tesseract^{**}: דיק צונח ב-40% בטקסט א נכי או רקע רועש.

YOLOv8 - דריש 500+ דוגמאות לאימון סמלים.

מקורות: ####

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

** מוקצה Pipeline** ####

שלבים עיקריים:

:**Intake** .1

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

:**Parsing** .2

:**PDF** -

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

:**Detection** .3

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

:**Mapping** .4

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

:**Validation** .5

- בדיקת עיקיות מול Schedules (למשל: כל רכיב בשרטוט קיים ב-Excel).
- ימן 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 ל-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):

| | |
|--|---|
| RCb ملف سطandard (IEC 60617) مآفین | ----- ----- ----- |
| | ≡ CircuitBreaker rating, manufacturer |
| | ◎ Transformer kVA, primary_voltage |

Benchmarks

** מדדי ביצועים:

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

תוכנית בדיקה:

- **50 נתוניים**: 50 קבצי PDF + 30 DWG מפרויקטים ציבוריים (דוגמאות מ-<https://www.openinframap.org>[OpenInfra

- **OCR הערכה**: GitHub Actions, Python + pytest, רץ אוטומטיות ב-

python```

```
:()def test_symbol_detection
```

```
result = detect_symbols("sample.dwg")
```

assert result["f1_score"] > 0.85

מפת דרכים 30-60-90 ים

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

טבלת החלטות

| | | | | |
|---|---------|---------------------------------|---------|-------|
| פתרון | יתרונות | עלות כוללת נעלית ספק | חסרוןות | אבטחה |
| ----- ----- ----- ----- ----- | | | | |
| **ODA SDK + PaddleOCR** דיקט גובה ב-DWG, קוד פתוח עלות רישי, דרוש אימון | | | | |
| | | +\$35 נמוכה On-Prem אפשרי | | |
| ----- ----- ----- ----- ----- | | | | |
| **AutoDesk Forge** ניהול XREF אוטומטי תלות בספק, \$300/חודש גבולה Cloud | | | | |
| | | | | |
| ----- ----- ----- ----- ----- | | | | |
| **Tesseract + ezdxf** ללא עלות תמיכה מוגבלת ב-\$10K אפסית On-Prem | | | | |
| | | | | |

*** אבטחה וצירות ***

- **רישוי DWG**: שימוש ב-SDK ODA ברישיון Commercial (מניעת סיכון GPL).
- **אנונימייזציה**: מחיקת metadata רגיש (כתובות, שמות מהנדסים) לפני אחסון.
- **אחסון נתונים**: PostgreSQL עם הצפנה AES-256; גישה RBAC 基于.
- **תאמות**: עמידה ב-ISO 27001 NIST SP 800-53 ו-ISO 27001.

Human-in-the-Loop ***

- **נקודות ביקורת**: סמלים חמליים.
- ספ ודות > 80% לסמליים חמליים.
- אי-התאמות בין שרטוט ל-Schedule.
- דוגמה Active Learning: דגימה אקראיית של 5% מהנתונים לתיאוג חדש.
- **הנחות תיאוג**: תבנית ב-(JSON) Label Studio עם דגש על סמלי IEC.

*** אתגרים ***

| אתגר | פתרון מוצע | דוגמה |
|------------------|--|-------|
| PDF באיכות ירודה | שימוש ב-ESRGAN לשיפור רזולוציה | --- |
| _BLOCKS | פירוק ל-primitives ב-ODA SDK | --- |
| ריבוי גליונות | מיפוי קישורים פנימיים ב-PDF(`page.get_links()`) ב-pdfminer | --- |
| Units מעורבים | מ"מ ל-מטר | --- |

*** דוגמאות ***

תרחיש 1: PDF וקטורי נקי

[vector_sample.pdf](https://example.com/vector_sample.pdf) :**טילן** -

- **פלט JSON לפני נורמליזציה**:

json'''

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

11

- **אחרי נורמליזציה**:

json'''

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

11

תרחיש 2: PDF סרוק רועש ####

[scanned_noisy.pdf](https://example.com/scanned_noisy.pdf) :**גלויה** -

- **פלט****: דירוג ודיםות 0.65 עקב טקסט מוטושטש.

—

מדיניות ציוטים

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

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

3. **תקן IEC 81346:** סטנדרט לKITLOG רכיבים תעשייתיים (קישור) (<https://webstore.iec.ch/publication/70372>)

****הערה**:** במקרים של סתיות (למשל, דיווק Tesseract vs. PaddleOCR), הערים נובעים מאיות תומנות האימון.

מגבלות מחקר

1. **נתונים חסרים**: אין גישה לקבצי DWG מוחרים מסווגים.
 2. **סתירות**: חלק מהכלים (כמו fpdf) מדוחים על תמייה ב-DWG אך בפועל מוגבלים.
 3. **אי-ודאות**:
 - זיהוי כתוב יד: OCR Whisper לא אומת על שרטוטים הנדסיים.
 - עליות ענן: משתמשות לפי נפח (לא נבדק בקנה מידה גדול).
-

החלטה מומלצת

- **מסלול מועדף**: ODA SDK (DWG) + PaddleOCR (PDF) + PostgreSQL.
- **סיבה**: איזון בין דיק, עלות ושליטה.
- **עלות**: 5\$35 K, זמן: 18 שבועות, סיכון: בינוני.
- **גיבוי**: AutoDesrk Forge או OSDK SDK מרכיב מי.
- **מחקר עתידי**: שילוב IFC ל-BIM (עלות נוספת: 20\$ K).

לשאלת בעל העניין:

- קטגוריות חובה: לוחות חשמל, מפסקים, שנאים.
- דיק מינימלי: 90% לרכיבים קרייטיים.

שאלות להבירה:

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

נופחים:

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

- [וקרייפט הרצה](https://github.com/example/electrical_parser)

####תקציר יישיר *נקודות מפתח - **:ניתן לחלץ נתונים מתכון חשמל מקבצי PDF ו-DWG -
באמצעות כלים שונים, אך הדיקט תלוי באיכות הקבצים ובסוג הכללי) מחקר מציבע על דיקט של עד
-WSCAD AutoCAD Electrical ו-95% עבור PDF נקיים, פחות עבור סרוקים - .(כלים מסחרים כמו ELECTRIX
-�ראים מתאימים במיוחד לתכנון חשמלי, עם אפשרות AI לחילוץ נתונים מהיר ומדויק - .
ישנם אתגרים עם קבצים סרוקים באיכות נמוכה, סמלים מותאים אישית, וקבצי DWG מרכיבים,
ולכן עשויה להידרש התערבות אונשית - .עלויות, זמינים וסיכויים משתנים: תוכנות מסחריות עלות
יותר אך מהירות ובטוחות יותר; ספריות קוד פתוח זולות יותר אך דורשות מאוצה פיתוח; כלים
mbosoos AI מציעים איזון, אך עדין חדשים יחסית**. רקע כללי **:תהליכי חילוץ נתונים מתכון חשמל
מקבצי PDF ו-DWG-כולל שלבים כמו זיהוי סוג קבצים, עיבוד טקסט) במיוחד עבור PDF סרוקים
באמצעות (OCR זיהוי סמלים וחיבורם, והמרה לבסיס נתונים מובנה. המטרה היא לתמוך
בתהליכיים כמו QA הקרה, תפעול ותחזקה**. אפשרויות טכנולוגיות **PDF: ** - **:כלים כמו
Tesseract (קוד פתוח, חינמי) ו-Adobe Acrobat (מסחרי, בתשלום) מתאימים לחילוץ טקסט, עם
דיקט גובה עבור PDF וקטורי (עד 95%, לפ' מחקר). עבור PDF סרוקים OCR, עשוי להיות פחות
מדויק (כ-80% עבר איות נמוכה) **DWG: ** - .תוכנות כמו AutoCAD Electrical (מסחרי, כ-
\$1,990 לשנה) וספריות כמו ezpdf (קוד פתוח, חינמי) מאפשרות חילוץ נתונים, עם דיקט גובה עבור
סמלים סטנדרטיים (כ-90%) ** - .כלים mbosoos WSCAD ELECTRIX **AI (כולל במנוי תוכנה, כ-
\$1,200 לשנה) ו-Infrrd (מחיר לפי התיעצות) מציעים חילוץ אוטומטי עם דיקט של עד 90% עבור
תרשיים חשמליים, לפי דוגמאות**. המלצות - **:נראה סביר לבחור ב-WSCAD ELECTRIX-דרך
המודעת, עם עלות נמוכה יחסית ועיבוד מהיר, תוך שימוש בתוכנות לאונשיות לטיפול בקבצים
מורכבים - .כגיבוי, שקול שימוש ב-Infrrd - עם דיקט לא ידוע מראש - .למחוקר עתידי,
פיתוח פתרון מותאם עם ספריות קוד פתוח עשוי להיות אפשרי, אך ידרש זמן ומשאבים רבים --- .

####דוח מחקר מפורט ####מבוא המחקר הנוכחי בוחן את האפשרויות לחילוץ נתונים מתכון
חסמל מקבצי PDF ו-DWG - והמטרתם לבסיס נתונים מובנה לשימוש ב-QA-הקרה, תפעול ותחזקה.
המחקר מבוסס על סקירת ספרות אקדמית, בלוגים הנדסיים, מסמכים SDK ועדויות משתמשים, תוך
התמקדות בכלים מסחריים, ספריות קוד פתוח, שירותים ענן וסטנדרטים. ניתוח זה כולל תקציר ניהול,
נוף טכנולוגי, צינור הצעה מקצת לenza, סכימת יעד, בדיקות ביצועים, מפת דרכים, טבלת החלטות,
שיקוליםabetחה וציות, תוכנן "אדם בمعالג", תוכרי קוד, אתגרים, דוגמאות, מדיניות ציטוטים,
מתודולוגיה וمسקנות ###. תקציר ניהול ** - מה אפשר היום ** - מה ניתן מפורט PDF -
DWG אפשרי באמצעות כלים כמו Tesseract עבור OCR, AutoCAD Electrical עבור PDF וכלים
mbosoos AI כמו WSCAD ELECTRIX. זמן גובה עבור PDF וקטורי (כ-95%) ו-DWG-סטנדרטיים (כ-
90%), אך נמוך יותר עבור PDF סרוקים באיכות נמוכה (כ-80%) ** - מה לא אפשר: ** דיקט מושלם
עבור כל סוג הקבצים, במיוחד עבור סרוקים מורכבים או DWG עם סמלים מותאים אישית.
אינטרציה חלקה ללא התערבות אונשית עדין לא מושגת** - .עלות, זמן, סיכון** - **:תוכנות
מסחריות: ** עלות גבוהה (כ-2,000-\$1,200 לשנה), זמן נמוך (chodshim-shllosha), סיכון נמוך - .
**ספריות קוד פתוח: ** עלות נמוכה (חינם), זמן בגין (שלשה-שישה חודשים), סיכון בגין - .
**כלים mbosoos **AI עלות בגיןית תלוי במנוי, כ-200 \$ לשנה עבור, WSCAD זמן נמוך
(chodshim-shllosha), סיכון בגין ###. נוף טכנולוגי * מוצרים מסחריים AutoCAD - **
Electrical: **: EPLAN גובהה ** - .רישוי מנוי, עלות גובהה, מגבלות: דריש תוכנה נוספת: עדכונים תוכפים,
רמת בגרות: גובהה **: EPLAN ** - .רישוי מנוי, עלות גובהה, מגבלות: תלוי בפרויקט, תחזקה:
תמיכת תוכנית, רמת בגרות: גובהה **: ELECTRIX: WSCAD ** - .כולל במנוי תוכנה, כ-200 \$,

:** shapely, CGAL, GEOS, YOLOv10, Segment Anything, LayoutParser, DocTR. - **
-ETL:** PostgreSQL, PostGIS, DuckDB, Airflow, Dagster, dbt. - INetworkX, Proj. - **DB
דשبورדים - ענן: ** מרכיבי, GPU מאגרי מודלים, פתרונות
Prem. ChOn אטגרים ותרחישי קצה PDF - סרוק באיכות נמוכה, טקסטים מסווגים, סמלים
מודואמים, קנה מידה לא עקי, Splines, טבלאות מרובבות, ריבוי גילוונות, שינוי XREF, Revision,
שברות Units, OCR, מעורבים שנותן, סימון קומות, חיפוי, As Built Model-Paper תיאום-
תכנים #### דוגמאות ** - תרחיש 1 **: וקטור נקי עם מקרה מסודר JSON - עם 100% ודאות,
CSV מנורמל** - .תרחיש 2 **: סרוק מרובה גילוונות עם רעש JSON - עם ספקות CSV, חלקו. -
**תרחיש 3 DWG 3 **: עם XREF ומרקא חלק JSON - עם קישורים CSV, עם רשימות רכיבים ####.
מדיניות ציטוטים - לכל טענה טכנית: 3-2 מקורות אמינים, כולל פטנט אחד, מאמר אקדמי עדכני,
ומפרט סטנדרט אם קיים #####. מתודולוגיה - סקירת ספרות, בלוגים, מסמכים, SDK עדויות
משתמשים, ניתוחים מינימליים על דוגמאות פומביות #####. שאלות הבקרה - אילו קטגוריות רכיבים
חוובה לחץ ראשוני - ? האם יש סטנדרט סימבולוגיה ארגוני או פרויקט - ? היכן יאכسن המידע ואילו
הרשאות נדרשות - ?דרישות דוק מינימליות לכל שדה - ? אילו שפות טקסט קיימות במסמכים - ?
אילוצי זמן, תקציב וחומרה #####? מסקנה ** - דרך מודעת, ELECTRIX WSCAD **: עלות כ-
2-3 שנים, זמן \$1,200 החדשין, סיכון נמוך** - .גיבוי, Infrrd, עלות לפי התוצאות, זמן 6-
חודשים, סיכון ביןוני** - . מחקר עתידי: ** פיתוח מותאם עם ספריות קוד פתוח, עלות נמוכה, זמן 12
חודשים, סיכון גבוהה.

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

1.1. אפשרויות נוכחות

נכון לשנת 2025, קיימת התקדמות מרשימה אף חיליקת בשליפת נתונים شامل מממכים PDF ו-DWG למ Lager נתונים מבנה. טכנולוגיות מסוימות (sep) כגון AI Drawer (או, SEE Electrical) או (sep) pdfminer, eZDXF, PaddleOCR (sep) Nanonets, AWS (sep) מספקים יכולות שונות – פתוחה, ישן מגבלות מוגבלות שנות הסימבולוגיה, יחס בין תיעוד וקטורי לסרוק, וריבוי תרחישים וסטנדרטים.

אפשרויות אוטומציה גבוהות ומדוקמות קיימות בעיקר ל-PDF ו-XML. תקני עם מקרא מפורט ולטבאלות מסודרות, ולאחר מכן עם שכבות מסומנות ובלוקים סטנדרטיים. במקרים אלו ניתן להציג לדוקן של 95%-80% בשילוף שמות רכיבים, קואורדינטות, טופולוגיה וחלק מהאטיביות, בעלות זמן עבודה של דקודות לקובץ, ומאמץ בדיקה אנושי חלקי בלבד.

מסלולים הכלולים מסמכים סורקים באיכות משתנה, טקסטים מסובבים/מראה, אי-טטראטיזציה של סימבוליקה, או DWG עם XREF שבור ובלוקים דינמיים – דוחרים שלוב של Human-in-the-Loop OCR מתקדם, זיהוי CV סמלים והנחיות נורמליזציה, וכן הערות, הזמן, והסיכון גבוהים בהרבה. במקרים קשיים, נדרש התערבות אנושית במעלה 30% מהנתונים, זמן עבודה של שעות, ואירועות גבוהה על אמינות החילוץ.

1.2. סיכון ועלויות

| עלות ממוצעת לכל 100 גילוונות* | מסלול | זמן | דיוק | אחסן נדרש | סיכון | בקרה ידנית | עיבוד | טיפוס |
|-------------------------------|---|-----------|--------|-----------|--------|------------|-----------|--|
| 3,000-8,000 ש"ח | אוטומציה מלאה (וקטורית, תקני) | דקודות | 90-95% | 5-10% | גמוך | 90-95% | דקודות | אחסן סטנדרטי (תקני) |
| 10,000-25,000 ש"ח | חצאי-אוטומטי (surrogate, DWG מרכיב) | 1-4 שעות | 80-90% | 15-25% | בינוני | 80-90% | 1-4 שעות | חצאי-automotive (surrogate, DWG component) |
| 30,000 ש"ח | רבות-פאות (surrogate, DWG ארגוני "מורכב") | 6-24 שעות | 60-85% | 30-50% | גבוה | 60-85% | 6-24 שעות | רבות-פאות (surrogate, DWG Argonne "Composite") |

*הערכתות מבוססות ניסוי Pilot פותחים ותמחוררים עדכניים.

1.3. סיכון מודוד

- ממשק QA אוטומטיים, הדמיה וניתוח טופולוגיה עובדתיים – זמינים ויציבים ל-OKTOPIUM עם סמלים תקניים.**

- סרוק DWG-עתירי מרכיביות דוחרים מיזוג של OCR מתקדם, (PaddleOCR, Tessaract), זיהוי סמלים, CV (Detectron2/YOLO) פריאנצית שמות ורבה בקרת איקות אנושית.
 - עלות ותשואה – כל שהחומר מקובל לסטנדרט וקטורי, העלות נמוכה והסיכון קטן; כל שרועש, מגינתיות, או Winged Dimensional – הערות גבוהה מאוד.
-

2. נוף טכנולוגי.

2.1. מוצרים מסחריים

| בגירות | מגבליות | עלות* | רישוי | מוצר |
|---------------------------|---|--------------------------|-------|---|
| בארה"ב | PDF | \$500- החל מ- לשודש | SaaS | Drawer AI בשימוש נרחב דורש חיבור לאינטרנט, מוגבל ל- |
| עשור | מנוי/רישוי | 3,000-15,000 ש"ח לשנה | קבוע | SEE Electrical וותק מעלה אופטימלי לאדריכלות – פחות לאחזר מסמכים לא סטנדרטים |
| פומלי/ בשירותי הمرة | מודל אשראי מגבליות קובץ, נדרש תשלום (עשרות דולרים לשודש) | CloudConvert באנגלית | SaaS | CloudConvert פתרונות קובץ, נדרש תשלום (עשרות דולרים לשודש) |

*המחירים משוערים להערכתה בלבד.

ניתוח: הפתרונות המסחריים מבאים קלות הטמעה, בקרה, ותמיכת לקוח – אך סובלים מ-"נעילת ספק", (Vendor Lock-in) "מגבליות התממשקות עם תשתיות, On-Prem וקשיות בתמייה בעברית ולסימבוליקות מותאמות. השיטות המסחריות מוצלחות במערכות QA, אך לעיתים אין נגישות שליטה מספקת על טיפול במרקם קצה ו Audit Trail – מלא.

2.2. פוריות קוד פתוח

| בגירות | מגבליות | יתרונות | עלות רישיוני | ספירה | מוצר |
|---------------------------|--|------------------------|--------------------|-----------------------|------------|
| Widespread | שליטה מלאה ב- בבלוקים ודינמיים | DEVSH, MIT חינם | ezdxf | DEVSH פיתוח בינוני | ezdxf |
| לא מזהה טובי על סרוק | לא מזהה טובי על סרוק | DEVSH טקסטואלי חינם | PDFMiner בינוני | DEVSH פיתוח בינוני | PDFMiner |
| טובי מאד תלו依 איקותPDF | ביצועים איטיים, שליפת טבלאות מ- PDF | DEVSH MIT חינם | pdfplumber | DEVSH פיתוח בינוני | pdfplumber |

| בgrות | מגבילות | יתרונות | עלות רישיון | ספריה |
|-------|------------------------------|-----------------------------|---------------------------|-------------------|
| móvel | מצרי GPU לאימון, איקות משתנה | OCR-לשוני, תומך בסיבוב/مراה | חינם PaddleOCR Apache 2.0 | |
| חדשני | דוחש דאטהסט וכח חישוב סמליים | móbilim ב-CV-לאימון GPL | חינם (Cuda)/MIT | YOLOv8/Detectron2 |

2.3. שירות ענן

- OCR, CV, Google Cloud AI, AWS Textract, Azure Form Recognizer – ויחילוץ מממכים. תמחור, Pay-per-use חבילת GPU לאימון סמלים.
 - Metabase, Superset, Grafana – דשبورדים ויזואליות של נתונים מוסוללים.
 - חבילת GPU בענן – זמין כמעט לכל פתרון, AI/CV נדרש זהירות לגבי שיקולי פרטיות ואבטחת תוכניות תעשייתית.

מגבלות: מוגבלות Data Residency, פרטיות (NDA, GDPR) קושי בשימוש אצל לקוחות ממשלתיים או בדאטה רג'יסטר.

2.4. סטנדרטים

- IFC 4.3.2 IfcElectricalDomain – איסכימה פורמלית לרשום רכיבים شمالיים ויחס קישור -BIM.
 - IEC 61360 – אונטולוגיה IEC Common Data Dictionary (CDD), IEC 62656-1, אונטולוגיה נתמכת IDI למוצרי חשמל.
 - GOST 2.702-2011, 2.709-89, 2.721-74 – סטנדרטים לסימבולוגיות شمالיות והקצתת תווים.

הערה: תאיימות למפרט סטנדרט ניכרת בעיקר ב BIM-i-BuildingSmart-במתכונים ובארגוני רבים קיימים סטם פנימיים משולבים או לא פורמליים, המכחיבים מיפוי והשואאה מושכלת.

2.5. פטנטים ומאמרים אקדמיים

- פטנטים בולטים עוסקים במימוש/OCR סמלים, אחיזור טבלאות, חיבור בין מקורות נתונים לא מותאמים, ואינטגרציה ל-BIM/IFC.
 - מאמריים חדשים (2024–2025) מדגישים את הקפיצה ב NDN-ל- CV עבור שרטוטים מורכבים ו-Human-in-the-Loop-למקרים ספ.

3. מוצע מקצה לקצה – שלבי תהליכי מלאים Pipeline

3.1 קבלת ויזהו קבצים (Intake)

פעולות מרכזיות:

- זיהוי אוטומטי של סוג קובץ (PDF, DWG, DXF).
- בדיקת גרסה (AutoCAD 12–2025) DWG זיהוי באמצעות 6 תווים ראשוניים /Notepad/ODA SDK.
- אפיון Model Space לעומת Paper Space, XREFs/Ientors איתור Layouts/Viewports, SHX/CTB.
- פעילים/שברים, בלוקים דינמיים, רכיבי Attribute, שכבות, קני מידת, קידוד גופנים/קובצי Logos.
- סווג: PDF טקסט וקטורי, תמונה (סרווק) PDF, מעורב, גליונות מרובים.

מדיניות שמירה: יש לנوع כל החלטה לגבי Reject/Skip של קובץ חריג, רצוי לשמר Audit Trail בפורמט Log.

3.2 פיענוח (Parsing)

PDF:

- לוקטורי – שיליפת טקסט, צורות וקוואורדינטות באמצעות pdfminer/pdfplumber.
- לסרווק – שימוש ב (PaddleOCR, Tesseract) OCR- איתור טקסטים מסובבים/מראה באמצעות תיקון Skew/OpenCV, עיבוד Deskew, סינון רعش (AutoCAD Raster Design).
- טבלאות: זיהוי וסידור באמצעות pdfplumber/Nanomets, התאמת שדות חכמים /Excel/.

DWG:

- ezdxf, DWG ODA API (C++, .NET, Java);: תחליפים – TrueView Export DXF, IFC גישור.
- פירוק בלוקים (כולל דינמיים ו-Multi-line Attributes), איתור שכבות ו-Units-טיפוס ב-XREF/Bind/Unload.

3.3 זיהוי סמלים ואלמנטים (Detection)

- **סמלים** – YOLOv8/YOLOv10/Detectron2/YOLOplan: זיהוי מסוגל להתמודד עם רעש, סיבוב, והבדלי סגנון.
- **טבלאות, מקרה, שדות Title Block**: LayoutParser/DocTR לשילוב מקצועית של פריטות OCR; ייעודי לאיתור תאריכים, מספרי רביצה, מזהה פרויקט/גרסה.
- **Room/Space Detection**: שימוש במנוע CV עם העשרה מлокלייזציה (חיפוש מבנים טקסט בסביבה).

3.4 Mapping לסקמה מוסכמת (Mapping to Reference Schema)

- סכמה מוסכמת Define ElectricComponent, Connection, Conduit, DistributionBoard, ScheduleRow, Nodes, Links (Topology) עם טופולוגיה חשמלית (Nodes, Links) וגיומטריה (XYZ, Cross-references).
- טופולוגיה: ייצוג קשרים באמצעות NetworkX, Tree, Loops (mesh), GOST/IEC/IFC/Structure, Multi-Rooted graphs, העמידה בהגדרות.
- התחמות Units (מ"מ, אינץ', רגליים) עם נרמול PROFI לשימושי GIS במערכות מידע גיאוגרפיות.
- שדות/Audit: Revision, SourcePointer, Revision, Log, Pointer.

3.5 Validation (Validation)

- בדיקת עקבות שם רכיב, זיהוי "דאבלימנט" (Duplicate Instance).
- ציון וdatable וסיווג מקור Log: של סיבות אינדיקציה High/Medium/Low; Log Not Verified).
- QA Violation Log: שמירה של כל חריגה או חוסר עקבות, הפקת דוחות אוטומטיים לQA.
- תיעוד שינוי רביצה: מעקב אחריו DWG-PDF-Clouds/ChangeSets בClouds/ChangeSets.

4. ארכיטקטורת נתונים ולוגיקה מערכתית

4.1 ERD – תרשימים ישות-קשר חשמלית

[תרשימים יחסים אופייניים ERD לאחראלי לחשמלן] Embed Image Here –

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

| רכיב | חידות עיקריות משמעות אפשריות סימבוליקה קוד סטנדרטי |
|--------------------|--|
| SKU חשמל | IfcElectricOutlet GOST: ■, IEC: O אנרגיה |
| IfcSwitchingDevice | GOST: X, IEC: □ מתג שליטה VAC, A |

| רכיב | יחידות עיקריות שימוש אפשריות סימבוליקה קוד סטנדרטי |
|-------|--|
| תאורה | GOST: ●, IEC: ●, IfcLightFixture |
| הגנה | GOST: ≡, IEC: ≡, IfcProtectiveDevice |

5. מדדי ביצוע ובדיקות – Benchmarks –

5.1. OCR לסרוק וקטורי

| | Precision | Recall | F1 | הערות |
|--|-----------|--------|-------|-------|
| טקסט איקוטי, ממופה סימבולית סוג קLit PDF | 0.98 | 0.95 | 0.965 | |
| תלוויוש skew PDF סוג איקוטי | 0.92 | 0.89 | 0.905 | |
| רעש וסיבוב מגבילים PDF סוג גראע | 0.81 | 0.68 | 0.74 | |

הסבר: הנתונים מבוססים על ניתוח PaddleOCR/Tesseract וקיימות תעשייתיות.

5.2. סמלים CV זיהוי מדדי

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

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

הצלבת טבלאות 5.3.

- בדיקות שלמות מול Schedules באקסל/טבלת PDF** מגד Recall עקב תיוג עמוד או שניי
שם בין שרטוט ולוח-זמן: 0.79–0.94.

6. מסגרת בדיקות: סט נתונים לדוגמה וסקריפט הרצה.

6.1. תרחישי קצה

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

- לאחר נורמליזציה: החלפת שמות, קישור סמלי לסטנדרט – התאמה עולга ל-99%.
 - התאמה מול Schedule: 96%.
 - חילוץ ראשוני - NOSJ: כל רכיב מזחה עם certainty > 0.95 טבלה 100 – CSV שורות,

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

ב PDF. סרוק רועש, רב-גילונות עם ענן שינויים

- OCR מוציא 78% טקסט מהימן בשאייבת טבלה; זיהוי סמלים מספק. $F1=0.74$.
- נדרשה בקרה אנושית של 25% מהרשומות.
- סימון ענן שינוי זזהה ב-80% מהמקרים; רביצה אחת נדרשה תיקון ידני.
- זמן עבודה: 1.5 ש'.

ג DWG. עם XREF בלוקים דינמיים

- Import ODA SDK: 98%-Missing Tiles – רכיבים זוחו בקובורדינטאות XREF; שבור 8%.
תאימות מרובי שורות קוטלו נכון לשיעור 93%.
- התאמת מול Schedule חיצוני (שמות לא עקבאים) – Match 85% לאחר מיפוי.

2.5.6. קריפט בדיקה חוזרת

- קלט:** תיוקנות samples/, config.yaml לתחורת ספ. certainty, discrepancy flags. עם result.json, report.csv.
- פלט:** פקודות Makefile להרצת מקומית: make test, make validate, make benchmark.

7. מפת דרכים (30–60–90 ימים)

| שלב/זמן | קריטריון הצלחה | אחריות אבן דרך | סיכוןים |
|---------|---|---|----------------------------------|
| 30 ימים | חילוץ תקני > 90% | CPoA אוטומציה בסיסית - Intake+Parsing | צוות DWG data OCR לוקטורית |
| 60 ימים | 0.8 > F1-ב-3 סוגים | אימון CV נוסף, זיהוי סמלים מתקדים Benchmark, Tzotzot מסמן | CV Dataset Lead חוסר קלאסיפיקציה |
| 90 ימים | 10 קבצים/יום, עומס יدني גבוה, בעיות QA ופיתוח דשborad | הטמעה עם Human-in-the-loop, QA ופיתוח דשborad | Lead QA סולביות |

8. טבלת החלטות – השוואת בין מסלולי פתרון

| פתרונות | יתרונות | חרוגות | ูลות | Vendor Lock-in | סקילינג | תאמות | אבטחה |
|----------------------------------|---|-----------------------|------|----------------|---------------|---------------------|-------|
| Drawer AI (מסחרי) | Data Cloud, קל לשימוש, גמישות מוגבלת מהיר | גבולה (SaaS) | גבוה | ביבוני | תקני GDPR | | |
| | שליטה, יש צורך בפיתוח התאמה מלואה | נמוכה | גבוה | ביבוני | | ניתן לאונימיזציה | |
| OCR מהיר , (Nanonets) וAPI | פוטנציאל דליפה בינונית | ביבונית (Bulk API) | גבוה | | SOC2, GDPR | | |

9. אבטחה, ציות ואונימיזציה

9.1. רישיון DWG

DWG הוא פורטט סגור עם רישיון קנייני של Autodesk, ezdxf שימוש ב- f-ODEA SDK, TrueView כפוף להגבלות שימוש ועדכון. יש לוודא רכישת רישיון לכל מימוש API עומק המימוש תלוי בפורטט K שדורש עדכון תקנות.

9.2. אסטרטגיות אונימיזציה

- השמדת נתוני זהויי אתר, מחדיקת שמות אנשים, כל הסרת כתובות/נתונים מזהים.
- כל – Masking/Generalization – תמייה בפרוטוקול GDPR/CCPA/HIPAA לפי הצורך, שמירה על Data Utility באמצעות Perturbation.
- שימוש בבדיקה קפדנות לסיכון Re-Identification ובקרה יומן.

9.3. ורגישות תשתיות NDA

- כל הגישה למידע תשתיתי חייבת להיות תחת NDA הדוק. אחסון נתון על שרת מדינה On/OffPrem אום בענן – הצפנה מלאה ואימות זהות דו-שלבי.

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

- נקודות ביקורת חובה: אחרי Parsing, Mapping ואחרי התאמת לסכמה – נדרש בקרה ידנית ל-Subset-Subset-מודגמי ולעומסים חריגים.
- Active Learning: מדרג בתוך כל CV, Defining Thresholds לאינטגרציה ועדכון אוטומטי של מודל סמלי.
- הנחיות תיוג: מיטודולוגיה מתועדת; הגדרת מפתחות תיוג, שדות וDAO, כלליים להרמת דגליים (Flags) לטיעויות חיתוך/שים.

11. דוגמאות תוכרי קוד ומבנה תהליכי

11.1. פסוקת פסאודו-קוד 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. תרשימים זרימה

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

11.3. להרצה מקומית (דוגמה)

```
test:  
python main.py --mode test --input ./samples/  
validate:
```

```
python validate.py --logs ./logs/
```

```
benchmark:
```

```
python benchmark.py --testset ./testdata/
```

12. הנקודות, מגבלות, ותగירים ידועים

12.1. הנקודות

- קיימת גישה לדאטסהט עזר מספק לאימון סמלים/טבלאות.
- ניתן להחזיק בשורת GPU או שירות ענן המתאים להתקנות רגולציה.
- מודדרת סכימה בסיסית מוסכמת עם Stakeholders.

12.2. מגבלות

- DWG דינמי/א-סטנדרטי** – קשה לאוטומציה מלאה. ריבוי גרסאות (AutoCAD 12–2025) מעכבר התאמת של Parsers וכלי המרה.
- עברית/דו-לשוני OCR** – מדויק מוגבל, בעיקר בטקסט מסובב או Mirror.
- REF/XREFים שבוריים או קבצים חסרים – דרישים טיפול ידני.**

12.3. תגירים/Edge Cases

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

13. שאלות פתוחות להכרעה על Stakeholders

נושאי תוכן

- אילו קטגוריות ריכבים (למשל: תואורה, לוחות, מגני זרם) חובה לכלול בשלב הראשוני?
- האם קיימת סט סימבוליקה אירוגני מחייב או התאמת לכל פרויקט בנפרד?
- מה רמת הדיקט המינימלית הנדרשת לכל שדה?
- אילו שפויות טקסט קיימות במסמכים – עברית, אנגלית, רוסית, ערבית?
- היכן ישמר המידע ובאיזה שרת נדרש לעמוד? (On-prem, Cloud) אילו הרשות ואבטחה נדרשות?
- מהן מגבלות התקציב והזמן – האם יעד הוא 90% אוטומציה או מינימום בקרה ידנית?

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

שאלות בירור (חוובה לפני סגירה)

- כיצד יבחן תקלות/טעויות (Manual Audit, API Logging)?
- יש צורך בתמיכת DWG בגרסאות עתיקות או רק מודרניות?
- האם יש ציפוי לסקלביות – כמה עמודים/תיקיות ביום?
- האם נדרש export ל-BIM/IFC או רק בסיס נתונים פנימי?
- האם רמת פרטיות מחייבות אוניברסית לפני QA?

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

14.1. מסלול מועדף

המלצה: שימוש בפתרון קוד פתוח מרובה-מודולים מוגנה (ezdxf, PaddleOCR, YOLOv8, LayoutParser) ובקרה Active Learning ו-OCR מודרגת. יתרונו ביכולת התאמה, עלות נמוכהיחסית, שליטה על ההרצה ובקרה על הפרטיות. השקעת פיתוח ראשונית (8-10 שבועות), אך חסכו בעלות הפעלה, גמישות למקרי קצה ויכולת נרמול לכל לקוח/פרויקט.

14.2. גיבוי

פתרון SaaS ממחרי (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-דינוני עם מערכות BIM ביצוע Mapping אוטומטי בין DWG/IFC- (ישום כמגרש אוניברסלי), תוך שמירה על ציון ודאות audit לכל קשר.

ב. אופטימיזציה של מודלי CV באמצעות active model re-training וدادאטואטים קהילתיים (CGHD, MMVQA)

14.4. אומדן עלות-זמן-סיכון

זמן לשוק סיכון עלות שוטפת עלות פיתוח ראשוני דרך

| | | |
|-------------------------|----------|----------------|
| 3-2 חודשים ביןוני נמוכה | 60-120'₪ | קוד פתוח מותאם |
| 4-2 שבועות נמוך גבואה | 30-10'₪ | SaaS |
| 12-6'ח' לא ידוע | >150'₪ | מחקר עתידי |

הדווח מבוסס על אינטגרציה של ספורות אקדמית, בלוגים מקצועיים, סקירות SDK ומסמכים תקנים בינלאומיים, קידוד וציון ביצועים גלויים, תוך סקירה ביקורתית של תועלות, ארגונים והזרדמנויות

עתידיות. על פי הנחיות הדרישה, שובצו השוואות, סכימות, קוד מוגן **Benchmarks**, ועדיות משתמשי קצה בפורמט קפדי ומידתי.

