**המחלקה להנדסת תוכנה**

**פרויקט גמר –**

אגירה וסיווג של מסלולי גופים פיזיקליים באמצעות למידת מכונה  
Storage and Classification of Trajectories via Machine Learning Methods

**מאת**

**אביתר גולן, 203311733**

**מנחה אקדמי: דר' ראובן יגל אישור: תאריך:**

**אחראי תעשייתי: דר' שי אקו אישור: תאריך:**

**רכז הפרויקטים: דר' אסף שפנייר אישור: תאריך:**

מערכות ניהול הפרויקט:

|  |  |  |
| --- | --- | --- |
| # | מערכת | מיקום |
| 1 | מאגר קוד | <https://github.com/EvjaG/FinalProject> |
| 2 | קישור ליומן | <https://mail313575.monday.com/boards/2070942680/> |
| 3 | קישור לסרטון דוח אלפא | [לחצו כאן לפתיחה](https://postjceac-my.sharepoint.com/:v:/g/personal/evyatargo_post_jce_ac_il/EUwYOYa_ct9Ho3oAkvpgGfgB8gtYa4l7bJQpM-QS8-BXYQ?e=yeRRr9) |

|  |  |
| --- | --- |
| סוג הפרויקט | תעשייתי, חברת hi-tech |
| פרויקט מח"ר | לא |
| פרויקט ממשיך | לא |
| פרויקט זוגי: | לא |

כלי פיתוח

|  |  |
| --- | --- |
| שפות | Python, Java |
| DB | MongoDB |
| יצירת תרשימים לדוחות | StarUML |

**מילון מונחים:**

**למידת מכונה** – תחום במדעי המחשב העוסק בפיתוח אלגוריתמים המאפשרים למחשב לעבד נתונים מתוך דוגמאות ומכך להסיק מסקנות לגבי מידע חדש שלא  
נתקל בו בעבר.

**רשת נוירונים מלאכותית** – משפחת מודלים בתחום למידת מכונה שפותחה בהשראת תהליך הלמידה המתרחש במוח אמיתי. מתחלקת לשלב שבו הרשת מתאמנת על ידי קבלת דוגמאות, ושלב שבו הרשת כבר 'מאומנת' ויכולה לתת ערכים לדוגמאות.

**משקולות** – בחלק מהאלגוריתמים של למידת מכונה, אנו רוצים 'לתקן' תופעות מסויימות על מנת למנוע או לקדם את השפעתן. עושים זאת ע"י הכפלת הערך במשתנה שנקרא *משקולת*, על מנת לשנות את השפעתו של הערך על החישוב בשלבי הלמידה וההערכה. המשקולות יכולות להתעדכן לפי הדוגמאות.

**מסלול** – בפרוייקט זה, מסלול הוא רשימה של נקודות במרחב המתארות את תנועתו של גוף, כאשר הנקודה הראשונה מייצגת את תחילת ההקלטה , וכל נקודה  
 נוספת מייצגת נקודת זמן . פורמט ייצוג המסלול בפרוייקט הוא בקבצי CSV, כאשר כל שורה מייצגת נקודות t, x, y, ו-z. בעתיד אולי יוספו פרטים נוספים לקבצי המסלול (כגון גודל הגוף,מבנה וכו').

תוכן עניינים

[0. מבוא 4](#_Toc92494875)

[1. תיאור הבעיה 4](#_Toc92494876)

[2. סקירת עבודות דומות והשוואה 4](#_Toc92494877)

[3. תיאור הפתרון 4](#_Toc92494878)

[דרישות ואפיון הפתרון 4](#_Toc92494879)

[האתגרים של הפתרון מבחינת הנדסת תוכנה 4](#_Toc92494880)

[המשך הפרק תיאור הפתרון 5](#_Toc92494881)

[רכיבים 5](#_Toc92494882)

[תהליכים ומצבי המערכת 5](#_Toc92494883)

[תיאור הכלים המשמשים לפתרון 5](#_Toc92494884)

[4. מה עשינו עד כה ? 6](#_Toc92494885)

[5. תכנית בדיקות 6](#_Toc92494886)

[בדיקות יחידה 6](#_Toc92494887)

[בדיקות אינטגרציה 6](#_Toc92494888)

[בדיקות מערכת 7](#_Toc92494889)

[6. נספחים 8](#_Toc92494890)

[א. טבלת סיכונים 8](#_Toc92494891)

[ב. טבלת דרישות 9](#_Toc92494892)

[ג. תכנון הפרויקט 9](#_Toc92494893)

[ד. סקר שוק 10](#_Toc92494894)

[ה. תרשימים 11](#_Toc92494895)

[ו. רשימת ספרות \ ביבליוגרפיה 13](#_Toc92494896)

# מבוא

כיום מוערך שיש בעולם כ-47,840 כלי טיס צבאיים (WMMDA) בשימוש, יותר מ-4,500 לוויינים פעילים המרחפים במסלול מעל האטמוספירה ואין-ספור כלי נשק המסוגלים לנוע באוויר או בחלל.

במדינת ישראל ישנו צורך תמידי לשמור על בטחונה, לרבות בטחונה האווירי. במשך שנים רבות נאספו ותועדו מסלולי גופים רבים שנעו שתוך ומחוץ לשטחה האווירי של מדינת ישראל.

# תיאור הבעיה

הבעיה – המצב הקיים כיום: המעקב כיום הוא אחרי גופים ספציפיים בעזרת חיישנים שמוקצים בלעדית לכל גוף. הצפי בעתיד הקרוב הוא היכולת לעקוב אחרי גופים רבים שמסלולם אינו ידוע מראש, לרבות בניית מערכת אוטומטית למעקב אחר גופים, בעזרת חיישנים שלא יוקצו מראש לגופים מסוימים.

# סקירת עבודות דומות והשוואה

נעשית בדיקה לגבי המערכות שבהם משתמשים בחברות לזיהוי וקטלוג מסלולים בצורה אוטומטית (לדוג' ברכבים אוטונומיים), הדורשת תשובה מצד הגורמים מהצד השני. לבינתיים אין תשובה, ולכן ביצעתי בהנחיית ראובן סקר שוק לגבי סוגי מסדי נתונים שיכולים להתאים לכמות גדולה של נתונים. השוואת מסדי הנתונים השונים נמצאת בנספחים. בנוסף , בהמשך המחקר נשווה בין סוגי למידה שונים, כולל למידה עמוקה ולמידת מכונה.

# תיאור הפתרון

## דרישות ואפיון הפתרון

הקמת מערכת הכוללת בסיס נתונים ומערכת לאגירת נתונים בזמן-אמת מחיישנים הקולטים גופים פיסיקליים הנעים במרחב, הכולל תיעוד מסלולם ותכונותיהם. המאגר יאחסן ויקטלג את הנתונים בעזרת הכלים של למידת מכונה ויאפשר זיהוי הגוף בהשוואה לגופים אחרים. נטמיע מס' אלגוריתמים של למידת מכונה ונעזר בהם על מנת למדוד את ביצועי המסלול של גוף חדש וסיווגו.

## האתגרים של הפתרון מבחינת הנדסת תוכנה

* בניית מערכת שתמדוד את ביצועי האלגוריתמים ותשווה ביניהם.
* אין מסלול אמיתיים - מידע מסווג ולכן אין אפשרות לקבלו
* התאמה של נתונים מחיישנים וממאגרי מידע קיימים לצורה הניתנת ללמידה בעזרת למידת מכונה.
* בניית מערכת הלמידה באופן שבו מסלולים באורכים שונים יוכלו להתאים למערכת לומדת, היות ומס' הכניסות מוגבל לגודל מסויים.
* הקמת מערכת מבוזרת שתוכל לאגור ולקטלג את המידע, ולמשוך אותו בשלב יותר מתקדם.

הפתרון אליו הגענו מכיל מס' רכיבים, ובהם:

* מחולל מסלולים, לצורך יצירת מסלולים בסיסיים עליהם נלמד בהמשך (היות והמסלולים הקיימים במערכות של החיל הם חסויים) ומערכת הכותבת מסלולים אלו לקבצי תיעוד וקבצים ללמידה.
* מערכת למידה –ככל הנראה רשת נוירונים, לא בטוח עדיין אם CNN או ANN
* מסד נתונים – נוכל לאחסן בו את כל המסלולים והמידע שקשור ללמידת המכונה לצורך שליפה מהירה במידת הצורך (לדוג' לזיהוי אובייקט הנע במהירות במרחב)

## המשך הפרק תיאור הפתרון

נא לפנות לנספחים לטובת תרשימי זרימה המתארים את פעולות המערכת המרכזיות.

### רכיבים

המערכת בנויה משני חלקים עיקריים;  
מחשב המשתמש  
יכיל את תוכנת הלמידה (כקובץ הרצה) עבור מערכת הסיווג, יוכל לפנות לשרת כדי להוריד את המשקולות המעודכנים במידת הצורך, ולאחסן מסלולים חדשים עם או בלי הסיווג

שרת  
השרת יכיל את מסד הנתונים, המשקולות המעודכנות ללמידה. יספק ממשק משתמש (REST)

ראו את [התרשים בנספחים](#_רכיבי_מערכת) המתאר את מבנה המערכת באופן כללי

### תהליכים ומצבי המערכת

המערכת תכלול ארבעה מצבים עיקריים:

* סיווג:
  + [סיווג של נתונים חדשים](#_סיווג_מחיישנים) - נתונים שייכנסו מתוך חיישנים חיצוניים יעברו סיווג בזמן אמת על ידי המכונה הלומדת
  + [סיווג של מקבצי נתונים](#_סיווג_ממסד) - נניח שיוכנסו נתונים למכונה ממקור חיצוני (לדוג' מנהל הוסיף מסלולים שנרכשו מחברה חיצונית) שלא נלמדו עדיין, אנו נבצע עליהם תהליך סיווג במידה ואין ידע על איזה סוג מסלולים אלו.
* [למידה](#_תהליך_הלמידה): המערכת תיקח את כל הנתונים שיש להם סיווג חיצוני (כלומר ידוע איזה סוג מסלול זה ופרטים נוספים קיימים עליו במידת האפשר), ותבצע תהליך למידה מאפס, עם שאיפה לרמת דיוק שתיקבע ע"י מנהל המערכת. הלמידה תתבצע אוטומטית אחת לתקופה במידה ונוספו נתונים מתאימים למסד מאז הלמידה האחרונה, או ידנית ע"י מנהל מערכת.

### תיאור הכלים המשמשים לפתרון

בפרוייקט השתמשתי עד כה ב-Python לכתיבת הקוד, ובספריות החיצוניות הבאות:

|  |  |
| --- | --- |
| הספרייה | הסיבה שהשתמשתי |
| NumPy | חישוב ויצירת המסלולים, פונקציות מתמטיות נוחות |
| Pandas | בדיקה של נכונות החישובים של המסלולים, חישוב מהירויות, הצגה נוחה יותר להדפסה על המסך |
| OpenCV | המרת מסלולים ממערכים לתמונות, לטובת בדיקה האם רשתות קונבולוציה יתאימו ללמידה של המסלולים |
| Matplotlib | המחשת המסלולים בעזרת אנימציה |

בנוסף, נעשו בדיקות על ביזור השרת למס' מחשבים בעזרת מערכת הקבצים Hadoop ומנוע העיבוד המקבילי Spark, של ארגון Apache, אך כעת הנושא הוקפא למטרת התקדמות עם נושא הסיווג.

# מה עשינו עד כה ?

* יצירת מחולל מסלולים, כולל כתיבה לקבצי PNG וCSV
* יצירת מודול לקריאת הCSV ופרסור (לדו’' מציאת המהירות בכיוון X בזמן T)
* בחירת מסד נתונים לעבודה ראשונית (MongoDB)
* חקר על סוגי אלגוריתמי הלמידה השונים לטובת הפרוייקט
* יצירת מכונה לומדת לסיווג המסלולים
* ביצוע ניסויים ראשוניים ללמידה, ללא בדיקה
* בדיקת המסלולים במחלקת test
* כתיבת מודול לאחסון ושליפת המסלולים ממסד הנתונים

# תכנית בדיקות

על המערכת יבוצעו שלוש סוגי בדיקות, אותן אתאר בפירוט בפרק הקרוב:

## בדיקות יחידה

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| מס' | תיאור | תנאי קדם | צעדים | דאטה | צפי | מציאות | עבר/נכשל |
| 1 | כתיבת מסלולים | התיקייה data צריכה להיות ריקה מתוכן | 1. בחירת מס' המסלולים (X) 2. בחירת Y נקודות למסלול 3. הרצה 4. בדיקה שיש X מסלולים, והחלוקה בין תיקיות train ו-test בערך 80-20 | אין | קבלת 0.8\*X קבצים בתיקייה train, 0.2\*X קבצים בתיקייה test,  כל קובץ עם Y שורות |  |  |
| 2 | אחסון נתונים במסד | המסד צריך להיות ריק טרם תחילת הבדיקה | 1. בחירה של איזשהו סט של דאטה אקראי 2. הכנסת הנתונים למסד 3. משיכת הנתונים בחזרה מהמסד 4. השוואה של הנתונים מול המקור לשגיאות | יווצר על-ידי המערכת, גודל וכמות יקבעו מראש ע"י הבודק | לא יהיו שגיאות |  |  |

בהמשך הדרך ייתכן כי נחליט שנכון כי מערכת קבצים מבוזרת (כדוג' Hadoop) תתאים למטרת הפרוייקט. אם כן,

נצטרך לבנות מערכת להזנת המסלולים לתוך המסד שישב על גבי מערכת הקבצים המבוזרת  
בצורה נוחה ומהירה, ולבדוק גם אותו.

## בדיקות אינטגרציה

* שליפה של מסלול חדש ממסד הנתונים ולמידה עליו, ללא מידע חדש מהחיישנים.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| מס' | תיאור | תנאי קדם | צעדים | דאטה | צפי | מציאות | עבר/נכשל |
| 1 | שליפת מסלולים מהמסד ולמידה עליהם | המערכת צריכה להיות מאומנת מראש | 1. לכתוב X מסלולים עם תוויות ידועות מראש, ולהפריד את התוויות מהמסלולים 2. להעלות את המסלולים למסד 3. בחירת כל המסלולים (X) מהמסד 4. הרצת למידה 5. השוואת התוויות שהתקבלו מהלמידה מול התוויות | אין | קבלת דיוק ברמה A לפחות ( ייקבע על ידי מנהל המערכת) |  |  |
| 2 | אחסון מסלול שעבר למידה במסד עם הסיווג שקיבל כתוצאה מהלמידה | המערכת צריכה להיות מאומנת מראש | 1. המכונה הלומדת תקבל מסלול כלשהו ותסווג אותו 2. נבדוק שהסיווג שהתקבל תואם לתווית הידועה מראש - אחרת נדחה 3. לנתק את הקשר למסד ולפתוח מחדש (אם השרת יהיה REST לא יהיה צורך לנתק את הקשר) 4. נמשוך את המסלול מחדש מהמסד ונוודא שהתווית שבמסד נכונה ושאין שגיאות באחסון המסלול | מסלול כלשהו שידוע לנו הסיווג שלו מראש | התאמה מלאה בין הסיווג הידוע לבין הסיווג במסד, בין המסלול המקורי לבין המסלול במסד |  |  |

## בדיקות מערכת

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| מס' | תיאור | תנאי קדם | צעדים | דאטה | צפי | מציאות | עבר/נכשל |
| 1 | דימוי זיהוי מסלול אמיתי | המערכת צריכה להיות במצב מנוחה, כמוגדר [בתהליכים ומצבי ומערכת](#_תהליכים_ונתוני_המערכת).  צריכות להיות משקולות בשרת שמעודכנות לאותו פורמט של הדאטה. | 1. הרצת מודול חיצוני שידמה חיישנים שכותבים בזמן אמת מסלול (המסלול והסיווג ידועים לנו מראש) 2. קבלת חלק מהנתונים במחשב משתמש וקריאה לשרת לבקשת המשקולות המעודכנות 3. הרצת הסיווג 4. בדיקת התוצאות שאכן יצאו דומות או זהות. | מסלול מופרד מהסיווג שלו. | קבלת 0.8\*X קבצים בתיקייה train, 0.2\*X קבצים בתיקייה test,  כל קובץ עם Y שורות |  |  |

# נספחים

## טבלת סיכונים

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Description / Impact** | **Prob**  **(%)** | **Impact (L, M,**  **H, VH)** | **Risk**  **Code** | **Mitigation Plan / Contingency Plan** |
| Tight Schedules | Due to the college board deciding to begin the second semester earlier, there will be less time to work on the project | 90% | H | 1 | Crunch. |
| No solution to classification problem | We won’t find a suitable model for machine learning to classify plots | 10% | VH | 1 | The project might be cancelled, or frozen over to next year. |
| COVID-19 | A surge in COVID-19 cases might cause me or one of the supervisors  to catch the virus | 60% | M | 2 | I can still work from home if my case won’t be too severe. If only one supervisor becomes ill, I can use the other, if both get sick, I’ll have no feedback, and might have to rely on Assaf. |
| New requirements | Any new requirements to come from the client side | 30% | H | 3 | Deviations from schedule. |

## טבלת דרישות

#TODO

* **דרישות פונקציונליות**

|  |  |
| --- | --- |
| מס' דרישה | תיאור |
| 1 | המערכת שמקבלת את המסלולים צריכה להתאים למסלולים עם X נקודות (ייקבע על ידי הלקוח), ולדעת להתאים מסלולים עם יותר או פחות נקודות |
| 2 | מסד נתונים לאגירת מסלול הגוף |

* **דרישות נוספות ולא פונקציונאליות**
  + **מהירות (זמן): המערכת צריכה לעלות תוך מס' שניות (לא יותר מ-5) מרגע זיהוי גוף ע"י חיישנים חיצוניים**
  + **עומס:**
  + **זמינות: רק בעת הצורך, המערכת לא תפעל תמיד**
  + **קיבול (צריכת זיכרון / CPU):**
  + **ניידות:**
  + **שימושיות:**
  + **תחזוקה:**
  + **אילוצי פלטפורמת מימוש:**

## תכנון הפרויקט

|  |  |
| --- | --- |
| 04.10.21 | <<פגישת הכרות עם הלקוח >> |
| 18.10.21 | תחילת למידה- מערכות מבוזרות לאחסון נתונים |
| 01.11.21 | המשך למידה |
| 15.11.21 | כתיבת דוח הצעת פרוייקט, תחילת תיכון אחסון המערכת |
| 29.11.21 | סיום תיכון ומעבר לעבודה על תיכון מערכת הסיווג |
| 13.12.21 | כתיבת מחולל מסלולים, בדיקת פורמטים לאחסון הנתונים |
| 27.12.21 | החלטה על הפורמט, בניית נתונים ללמידה עתידית |
| 10.01.22 | חקירת סוגי הלמידה למציאת למידה מתאימה |
| 24.01.22 | המשך חקירה |
| 07.02.22 | הגעה למסקנות ותחילת כתיבה של תוכנית עבודה למערכת |

## סקר שוק

1. מסדי נתונים

בדיקת מערכות לאגירת מסלולים במערכות אחרות, לדוג' מערכות רכבים אוטונומיות, מטוסים וכו'.

בדיקת סוגי מסדי הנתונים השונים

לאחר התייעצות עם המנחה הוסכם כי שנינו מעדיף לעבוד עם מסד נתונים לא-יחסי, כלומר מסוג NoSQL. הבסיסים אותם אני בודק למען הפרוייקט הם MongoDB, Cassandra ו-HBase.

**Cassandra**בסיס נתונים מבוזר הכתוב בשפת ג'אווה, מיועד לטיפול בכמות נתונים גדולה העולה על מאות טרה-בייטים.  
הבסיס אינו משתמש במערכת Master-Slave אלא כל מחשב בצומת יוכל הפנות לכל צומת אחר. בנוסף, כל צומת מכיל שכפולים של מס' צמתים אחרים. המסד מבוסס שיטת "אחסון עמודה רחבה", כלומר משתמש בטבלאות עם שורות ועמודות, אך בניגוד מסדי SQL מאפשר ששמות ופורמטי העמודות ישתנו משורה לשורה.   
*חסרונות* הבסיס כוללים חוסר עקביות בעת שחזור מידע לצומת שנפלה, זמן קריאה ארוך מהשרת במידה והמפתח אינו ידוע למשתמש (לדוג' חיפוש של דוגמא ספציפית), וחוסר של תיעוד רשמי מאת הארגון המפתח, קרן Apache. שימושים ראשיים למסד כוללים תוכנות מסרים מיידים, מסחר אלקטרוני ועיבוד מידע מחיישנים בזמן אמת.

**HBase**בסיס נתונים הכתוב גם הוא בשפת ג'אווה, מבוסס על מערכת Bigtable של חברת גוגל, ומשתמש במספר תכונות שלו לרבות ביצוע פעולות על-גבי הזכרון בזמן אמת (לעומת גישה למידע הנמצא על גבי מחשבים רבים, פעולה הלוקחת הרבה יותר זמן), כיווץ מידע במוסד ושימוש במבני נתונים מבוססי Bloom Filter, אשר יכולים לקצר את זמן ההמתנה למשאבים מהדיסקים באופן משמעותי. בדומה ל-Cassandra, גם HBase מבוסס שיטת "אחסון עמודה רחבה". אחת מחוזקות המסד היא השימוש המובנה שלו ב-HDFS, מערכת הקבצים המבוזרת של Hadoop. המסד משתמש במערכת הקבצים HDFS של Hadoop, כך שאין צורך להשתמש במערכת MapReduce. חסרונות עיקריים של המסד כוללים שימוש במודל Master/Slave, דבר התורם לסיכוי לנקודת כשלון יחידה, אין ל-HBase שפת שאילתות מובנית (צריך להשתמש בממשק חיצוני, לדוג' Apache Hive), ותלות-יתר גבוהה בטכנולוגיות המובנות של Hadoop. שימושים שכיחים ל-HBase כוללים אנליטיקת לוגים מקוונת ואתרים מרובי תוכן (כגון נטפליקס, פינטרסט, פייסבוק).

**MongoDB**בסיס נתונים הכתוב בשפות JS, C++ ופייתון, מבוסס שיטת אחסון מסמכים, המציע גרסת קהילה וגרסת עסקים הכוללת גישה מהירה וקלה לנתונים על גבי הרשת, ביקורת נתונים מובנית, אבטחת מידע והצפנה.  
MongoDB מאחסן את הנתונים על גבי המערכת כמסמכי JSON בינאריים, שיטה המאפשרת גישה מהירה למידע. גרסת העסקים גם כוללת זמינות גבוהה ע"י שכפול המידע ליצירת עודפות וריבוי צמתים לאחסון. חסרונות של המסד כוללות חוסר תמיכה רשמי לגרסת הקהילה (לדוג' תיקוני טלאי צריכים להתבצע באופן ידני), וגרסאות סקלביליות גוזלות יותר משאבי זכרון.

## תרשימים

#### תהליך הלמידה

Diagram

Description automatically generatedתיאור תהליך הלמידה של המערכת; יתבצע בהפעלה הראשונית ואחת לתקופה התיקבע מראש ע"י מנהל המערכת

#### סיווג מחיישנים

Diagram

Description automatically generatedתהליך הסיווג של מסלול חדש שנקלט ע"י חיישנים חיצוניים

#### סיווג ממסד

Diagram

Description automatically generatedתהליך הסיווג של מסלולים ממסד הנתונים שעוד לא סווגו

#### רכיבי מערכת

Diagram

Description automatically generatedתרשים המתאר את מבנה המערכת ממבט עליון

## רשימת ספרות \ ביבליוגרפיה

עבור סקר השוק:

[**https://logz.io/blog/nosql-database-comparison/**](https://logz.io/blog/nosql-database-comparison/)[**https://db-engines.com/en/system/Cassandra%3BHBase%3BMongoDB**](https://db-engines.com/en/system/Cassandra%3BHBase%3BMongoDB)[**https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0045-4**](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0045-4)