**המרכז האקדמי רופין**

**בית-הספר להנדסה**

**המחלקה להנדסת חשמל ומחשבים**

**ספר פרויקט לשנת תש"פ**

**נושא הפרויקט:**

**Deep Learning for Malware Classification**

מבצעים:

עדן אביטן ת.ז 203286992

רועי אספורטס ת.ז

תומר עפרוני ת.ז 205711682

מנחה:

מנחה פנימי: ד"ר בני סלומון

השנה: תשפ״ב

תוכן עניינים

[1 תקציר 2](#_Toc85895066)

[2 מבוא 3](#_Toc85895067)

[2.1 כיצד קשור או משתלב הפרויקט לתחום כללי כלשהו 3](#_Toc85895068)

[2.2 הגדרת הבעיה 3](#_Toc85895069)

[2.3 האתגר הטכנולוגי 4](#_Toc85895070)

[3 דרכי פתרון הבעיה 4](#_Toc85895071)

[4 תוצר מצופה מהפרויקט 4](#_Toc85895072)

[5 תיאור רעיון דומה שיכול להוות השראה 5](#_Toc85895073)

[6 סיכונים, אי וודאות ואילוצי הפרויקט 5](#_Toc85895074)

[7 מקורות קריאה 6](#_Toc85895075)

[רשימת נספחים 7](#_Toc85895076)

[נספח א - סכמת בלוקים 8](#_Toc85895077)

[נספח ב - טבלת אבני דרך ותוצרים 9](#_Toc85895078)

# תקציר

בעולמנו כיום עולה השימוש בטכנולוגיה בצורה כבירה ועם עלייה זו, צצים אין ספור איומים על המשתמשים השונים. מתקפות הסייבר בעולם מהוות סיכון מרכזי לפגיעה במרחב הקיברנטי של היעד במטרה לגנוב ממנו מידע ואף להסב לו נזק.

בנוסף לכך, תחום טכנולוגי אחר אשר מתפתח במהירות אדירה הינו הבינה המלאכותית – מצב שבו מנסים לדמות את יכולות החשיבה האנושיות ומבנה מוח האדם בעזרת אמצעים טכנולוגים. נושא זה מסתעף לתחומים רבים, אם כי הנושא העיקרי שמעניין אותנו הוא "הלמידה העמוקה" – תחום הנגזר מתוך למידת המכונה, אשר מתבסס על רשתות נוירונים (מודל מתמטי חישובי שפותח בהשראת תהליכים מוחיים/קוגניטיביים) בעלות מספר רב של שכבות.

פרויקט זה בא לשלב בין שני העולמות הללו. מטרתנו היא לדמות, לשחזר ואף להשוות את תוצאותינו לניסוי [1] שנערך באוניברסיטת סאן חוזה כחלק מתזה שבוצעה בתחום. הרעיון הכללי הוא בניית מודל למידה עמוקה אשר בהינתן תמונה של נוזקה (Malware) יידע לסווגה למשפחת נוזקות ספציפית.

תמונה של נוזקה – זוהי בעצם תוצאת המרה של נוזקה לכדי תמונה. כך נוכל לבצע סיווג על בסיס למידה עמוקה באמצעות עיבוד וניתוח התמונה.

בנוסף לכך, נרצה להשוות את תוצאות הסיווג כתוצאה משימוש ב-CNN לתוצאות סיווג שונות, המבוססות על מודלים שונים מעולמות למידת המכונה, אותם נבנה. זאת בכדי לנסות ולבדוק עלות אל מול תועלת (כגון: זמני ריצה, ביצועים וכו') בין מודלים פרימיטיביים אל מול מודלים חדישים כדוגמת CNN.

מכיוון שמדובר במודל סיווג נתונים, התוצאות אותן נרצה לספק יהיו הסתברויות שינועו בין 0 ל-1, כך שככל שנתקרב ל-1 אזי מדובר בזיהוי נכון.

בפרויקט זה בנינו ארבעה מודלים. שלושה מהם, הם מודלים פרימיטיביים שמטרתם להוות נקודת ייחוס למודל ה-CNN אותו רצינו לחקור, והם: LDA, QDA, GNB. כל אחד מאלה רץ (אימון ומבחן) על מאגר הנתונים כאשר כל מודל נחקר בשלושה אופנים שונים, כך שבכל אחד מהם התמונות במאגר הנתונים הוגדרו להיות:

32x32, 64x64, 128x128 פיקסלים.

התוצאות שהשגנו אכן מוכיחות את עליונות השימוש ב-CNN אל מול המודלים השונים שנבדקו בפרויקט זה.

בתוצאותינו ניתן לראות כי בתהליך המבחן על מאגר הנתונים 128x128 להוסיף השוואת נתונים ביחס לאחרים וביחס למחקר + מסקנות.

# מבוא

תשתית החברה כיום מתבססת ברובה על מחשבים ואינטרנט. אבטחת הסייבר היא אומנות ההגנה על רשתות, מכשירים ונתונים מפני גישה בלתי מורשת או שימוש פלילי ובנוסף מהווה אבן יסוד בכל הנוגע להבטחת הסודיות, תקינות וזמינות המידע. לפי המחקר [1] עליו אנחנו מתבססים, בשנת 2018 הובחנו כ-669 מיליון סוגי נוזקות חדשות.

תחום העולה גם הוא בעולם הטכנולוגי בעקבות תוצאות וממצאים מחקריים טובים הוא תחום הבינה המלאכותית.

מתחום זה נגזרת הלמידה העמוקה אשר בה נעסוק בפרויקט זה. למידה עמוקה היא תחום העוסק באלגוריתמים ומודלים בהשראת מבנה ותפקוד המוח, מודלים אלו נקראים רשתות עצביות מלאכותיות.

ניתן לראות שילובים בין העולמות הללו בתוכנות אנטי וירוס שונות אשר מבססות את יכולת ההבחנה שלהן על בסיס בינה מלאכותית על מנת לאתר ולזהות פעולות ותהליכים זדוניים.

עקב העובדה כי תהליכים וקבצים זדוניים משתנים בקצב מהיר, קיים אתגר עבור עולם אבטחת הסייבר המתבסס על בינה מלאכותית לזהות ולנתר את אותם האיומים. בנוסף, מודלים ואלגוריתמים מבוססי למידה עמוקה מצריכים יכולות מחשוב גדולות על מנת לספק תוצאות בזמן סביר.

* 1. הגדרת הבעיה

הבעיה אותה הפרויקט בא לפתור היא בעיית סיווג נוזקות למשפחות באמצעות למידת מכונה. כלומר, סיווג שיטתי של איומים אלו, המאיימים על מערכות תקשורת ואינטרנט המהוות תשתית עבור ארגונים רבים. ביצוע והצלחה בפתירת בעיה זו, כרוך בקשיים רבים והתאמות משמעותיות.

* 1. שיטות עבודה וטכנולוגיות
     1. סביבת העבודה שבה השתמשנו היא Jupyter Notebook אשר תרוץ על Google Colab.
     2. הפרויקט כולו נכתב ב-Python, כאשר השתמשנו במגוון ספריות שונות, כגון:
* Keras אשר שימשה ליצירת מודל ה-CNN.
* Scikit-Learn אשר שימשה ליצירת המודלים LDA, QDA, GNB.
* Pandas & Numpy אשר סייעו בעיבוד המידע.
  + 1. מאגר הנתונים שעליו התבססנו הוא Malimg Dataset המכיל כ-9764 תמונות בסקלה של גווני הצבע האפור. את המאגר נמיר לשלושה גדלים שונים (פיקסלים) והם: 32x32, 64x64, 128x128. זאת על ידי שימוש בספריית Pillow.
    2. הפרויקט וקוד המקור מנוהלים באמצעות GitHub.

1. למידת מכונה, למידה עמוקה ומודלי למידה
   1. למידת מכונה

תת-תחום במדעי המחשב ובבינה מלאכותית. התחום עוסק בפיתוח אלגוריתמים המיועדים לאפשר למחשב ללמוד מתוך דוגמאות, ופועל במגוון משימות חישוביות בהן התכנות הקלאסי אינו אפשרי.

תחום זה מכיל שלושה סוגי למידה עיקריים והם:

* למידה מונחית (Supervised Learning): מאופיינת על ידי תיוג מאגר המידע לקטגוריות, כך שבתהליך הלמידה של המערכת היא תוכל להשוואת את תוצאותיה לתיוג שהוגדר מראש, ובעצם ללמוד לסווג את המידע על סמך תשובות שהוגדרו, תהליך זה יקרא שלב האימון (Train). לבסוף, המערכת תידרש לסווג קלטים חדשים שאיננה מכירה לקבוצה המתאימה, תהליך זה יקרא שלב הבדיקה (Test).
* למידה בלתי מונחית (וUnsupervised Learning): מאופיינת על ידי היכולת לקבל מידע שאינו מתויג לקטגוריות, ובעצם לדעת לנתח ולאגד אותו ללא התערבות אדם.
* למידת חיזוק (Reinforcement Learning): מאופיינת על ידי היכולת להעניק תגמול על התנהגות נכונה ו-״להעניש״ כאשר ההתנהגות אינה נכונה. המערכת תקבל משוב חלקי בסיום העבודה ותדע להסיק מסקנות לגבי השלבים אשר אותם יש לשפר/לשמר.

בעת אימון מודלי למידת מכונה, יש להיזהר ממצב של התאמת יתר (Overfitting) אשר עלול לגרום לכך שהמערכת תדע לסווג בצורה טובה את הדאטה שניתן לאימון, אך לא תצליח לעשות זאת עבור מידע חדש.



איור 1: הצגת קו חלוקה בצורות התאמה שונות

פרויקט זה מתמקד בעיקרו בלמידת מונחית, כאשר תייגנו את התמונות ל-25 משפחות שונות של נוזקות.

* 1. למידה עמוקה

היא חלק ממשפחה רחבה יותר של שיטות למידת מכונה המבוססות על רשתות עצביות מלאכותיות בשילוב עם למידת מאפיינים. הרעיון מאחורי שיטה זו היא היכולת ללמד מחשבים לבצע עבודה אשר נראית טבעית למוח האדם אך מסובכת לביצוע המחשב.

בלמידה עמוקה המודל יקבע אילו מהפיצ׳רים רלוונטיים להסקת המסקנות ואילו לא, לעומת למידת מכונה, אשר בה יש לבצע עבודת הכנה מוקדמת של סידור הדאטה ובחירת הפיצ׳רים המתאימים.

במקרה שלנו, הפיצ׳רים באמצעותם נרצה לפתור את בעיית הסיווג הם הפיקסלים של התמונות אותם נכניס כקלט למודלים.

* + 1. רשת עצבית מלאכותית

רשת עצבית מלאכותית (ANN – Artificial Neural Network) היא מודל מתמטי חישובי, שפותח בהשראת תהליכים מוחיים או קוגניטיביים המתרחשים ברשת עצבית טבעית ומשמש במסגרת למידת מכונה. הרשת מכילה לרוב מספר גדול של יחידות מידע המקושרות זו לזו, כאשר צורת הקישור בין היחידות מזכירה את אופן חיבור הנוירונים במוח ומכאן נובע שמה.

ברשת מספר רב של נוירונים כאשר כל אחד מהם מבצע חישוב פשוט יחסית. אל הנוירון מתקבל קלט אשר עליו מבוצע עיבוד באמצעות נוסחה מוגדרת שתוצאתה מהווה הפלט של אותו הנוירון. פלט זה מועבר הלאה אל נוירונים אחרים ברשת אשר מבצעים תהליך דומה.

למחוק

הרשת מאופיינת על ידי מספר תכונות, בהן צורת החיבור של הנוירונים (כיצד הנוירונים מחוברים בין השכבות השונות), משקלי ההשפעה של הנוירונים (כמה משפיע כל נוירון על ערך הנוירון הבא, עד כמה הוא משמעותי לעומת שאר הנוירונים באותה שכבה) ופונקציית ההפעלה שהיא הפונקציה שמוציאה את ערך הפלט של הנוירון, זוהי פונקציה לא לינארית המשקללת, בהתאם לפונקציה, את ערכי הנוירונים בכניסה יחד עם משקליהם.

לשכתב!

הרשת מאופיינת על ידי מספר תכונות כגון, צורת חיבור הנוירונים המהווה האופן בו הנוירונים מחוברים האחד לשני בין השכבות השונות, פונקציית ההפעלה (Activation Function) הקובעת האם קלט הנוירון ברשת הוא חשוב או לא וזאת באמצעות פעולות מתמטיות.

* 1. מודלי למידה
     1. LDA
     2. QDA
     3. GNB
     4. CNN

להוסיף גם סוגי שכבות

## 