תקציר מנהלים**:**

**בעקבות ההתפתחות הטכנולוגיה המהירה והחדשנית, עולם הסייבר נהיה מורכב יותר ויותר, קיימות רוגלות העוקבות בחשאי אחר הרגלי הגלישה של המשתמשים השונים ומדליפות מידע פרטי על המשתמשים לגורמים שונים החל מחברות וארגונים וכלה בגורמים זדוניים. למעשה, החיים האישיים של המשתמש הפרטי הופכים לחשופים יותר, וכיום ניתן לגלות מידע רב על המשתמש באופן מהיר ופשוט. התפתחות העולם הטכנולוגי גורמת לאנשים רבים להתקין ללא יודעין אפליקציות ריגול, אשר מאפשרות גניבת מידע אישי ורגיש כגון, כרטיסי אשראי, חשבונות וסיסמאות בנקים וכו'.כיום, מערכות רבות מתמקדות בזיהוי רוגלות מבוססות חתימה אלקטרונית ולעיתים לא ניתן להבדיל בין נתונים זדוניים לבלתי מזיקים ובעקבות אתגר זה, נבע הצורך בפרויקט זה. מטרת הפרויקט היא לאפשר גילוי וניתוח רוגלות זדוניות באמצעות אלגוריתמי למידת מכונה ו** PE (Portable Executable) **על בסיס רשתות נוירונים מלאכותיות. כדי להשיג את המטרה שלנו, בנינו מערכת קוד פתוח המיועדת לשימוש המשתמש הפרטי אשר תכריע ותסווג האם קובץ הוא זדוני או לא.**

**המערכת מוצגת בעזרת ממשק ידידותי למשתמש ומכילה פונקציות לזיהוי הרוגלות, ניתוח נתונים וייצור תחזיות אשר יאפשרו להבדיל בין נתונים זדוניים ובלתי מזיקים.סט הנתונים המקורי כלל 487 עמודות אשר מייצגות חלק מסוים מתוך נתוני הקובץ ו114,844 רשומות אשר מייצגות קבצים שונים וסיווגם (זדוני או לא זדוני).הפרויקט כולל מספר אבני בניין: צמצום סט הנתונים בעזרת** Feature Selection **והורדת מימד** Dimension Reduction, **מציאת הפרמטרים אשר משמעותיים לקביעת סיווג הקובץ, טעינת המידע המצומצם לתוך מחסן נתונים, כתיבת מודל חיזוי מסוג** ANN **והתאמתו באופן אוטימלי עד להשגת מדדי הפרויקט. כמו-כן, יצרנו ממשק** WEB **אשר ישמש את המשתמש הפרטי להעלאת קבצים וקבלת הסיווג וכן הצגת גרפיים ומידע סטטיסטי אשר ישקפו את הנתונים השונים אשר נובעים מכל שלב בפרויקט.**

**בפרויקט זה נעשה שימוש בשפת פייתון לכתיבת קוד האלגוריתמים, מסד נתונים מסוג** MySQL **כדי לאחסן ולאחזר את הנתונים, תוכנת** Flask **כדי לממשק את הקוד לממשק ה**WEB, **ושימוש ב**PowerBI **על מנת להציג את נתוני החיזוי השונים.במהלך כתיבת מסמך זה ניתחנו מספר רב של שאלות מחקר אשר סייעו במציאת מודל חיזוי אופטימלי אשר עונה על כלל היעדים אשר הוגדרו בתחילת הפרויקט. כחלק מתוצאות פרויקט זה, ניתן לראות כי המודל סיווג נכונה** 95.19% **אחוזים מהקבצים אשר נבדקו.**

**הצעדים להמשך עבור פרויקט זה כוללים הנגשה של האתר לכלל הציבור שכן כיום האתר נגיש באופן פנימי בלבד כחלק מסביבת השרתים של מכללת אפקה. כמו כן, המלצתנו היא ליצור מערכת ידידותית למשתמש אשר יודעת לעבד קבצים ממשיים מפורמטים שונים, לחלץ מהם את ה**PE Headers **וכן להעשיר את סט הנתונים, אלגוריתמי למידת מכונה והנתונים הגרפיים באופן שוטף על סמך מידע חדש שנכנס למערכת על ידי משתמשי הקצה.**

**במסמך זה ניתן למצוא סקירה רחבה של האלגוריתמים והאנליזות אשר ביצענו עד לקבלת התוצאה הסופית של פרויקט זה וכן המלצות לשלבים הבאים כדי להרחיב את הפתרון המוצע.**

**Executive summary:**

Due to the constant and rapid progress of technology, the cyber world is becoming more and more complex. We are witnessing a rise up in the number of malwares that secretly spies on different users and leaks private information on those users to different organizations and groups which can have, in many cases, harmful and malicious intentions.

The personal lives of private users are much more exposed to the public, and harmful groups and individuals can identify such information in many simple ways. The technology evolvement can cause many people to install without knowing malicious spyware that allows access to personal and sensitive information such as credit card details, bank account details, and even passwords.

These days, many systems are focusing on identifying malware based on electronic signatures and it is not always possible to classify whether the data is malicious or not. The goal of this project is to identify malicious malware using machine learning algorithms and Portable Executable (PE) Headers based on Artificial Neurons Network (ANN). In order to achieve this goal, we built an open code system that designs for end-user use and classifies if a file is malicious or not. The system is represented by a user-friendly web interface that enables malware identification by inserting CSV files, analyzes the PE data, and generates predictions that will determine whether the file is malicious or not based on a predefined data set.

The original data set consisted of 487 columns (features) and 114,844 rows, each column represents a part of the PE headers, and each row stands for a file that was already classified (malicious or not malicious).

This project contains numerous milestones: narrowing down the original data set using Feature Selection and Dimension reduction, identifying the key parameters that are significant for the file classification, loading the reduced data into a Data Warehouse, writing an ANN prediction model, and optimizing it in order to achieve the project’s Key Performance Indicators (KPIs). In addition, we created a web interface that will allow the end-users to upload files and receive the file final classification and represent graphic and statistical data that will reflect the project information and analysis of each phase.

In this project, we used Python code for writing the different algorithms, MySQL Database for restoring and retrieving the data, Flask software for integrating the code and the web interface, and PowerBI for presenting the different predictions data.

We have investigated and analyzed numerous questions that assisted with identifying the required prediction model and optimizing it until achieving the project’s goals and KPIs. As part of the project results, we were able to successfully classify 95.19% of the files that were tested.

The future next steps for this project include opening the website to the public as part of the internet network (today, website access is possible only for internal Afeka Academic College users). In addition, our recommendation is to create a user-friendly system that can process different format files, extract the PE headers and enrich the data set, machine learning algorithms, and graphic data dynamically based on new information that was added by the end-users.

This document contains a thorough review of the algorithms and analysis that we performed as part of this project and recommendations for expanding the suggested system and solutions.