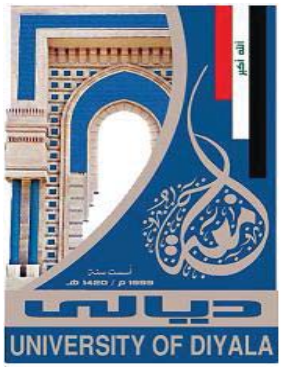
Ministry of Higher Education and Scientific Research
University of Diyala
College of Science
Department of Computer Science



Classification Method of Advanced Persistent Threat (APT) Malware Using Advanced Machine Learning

A Thesis

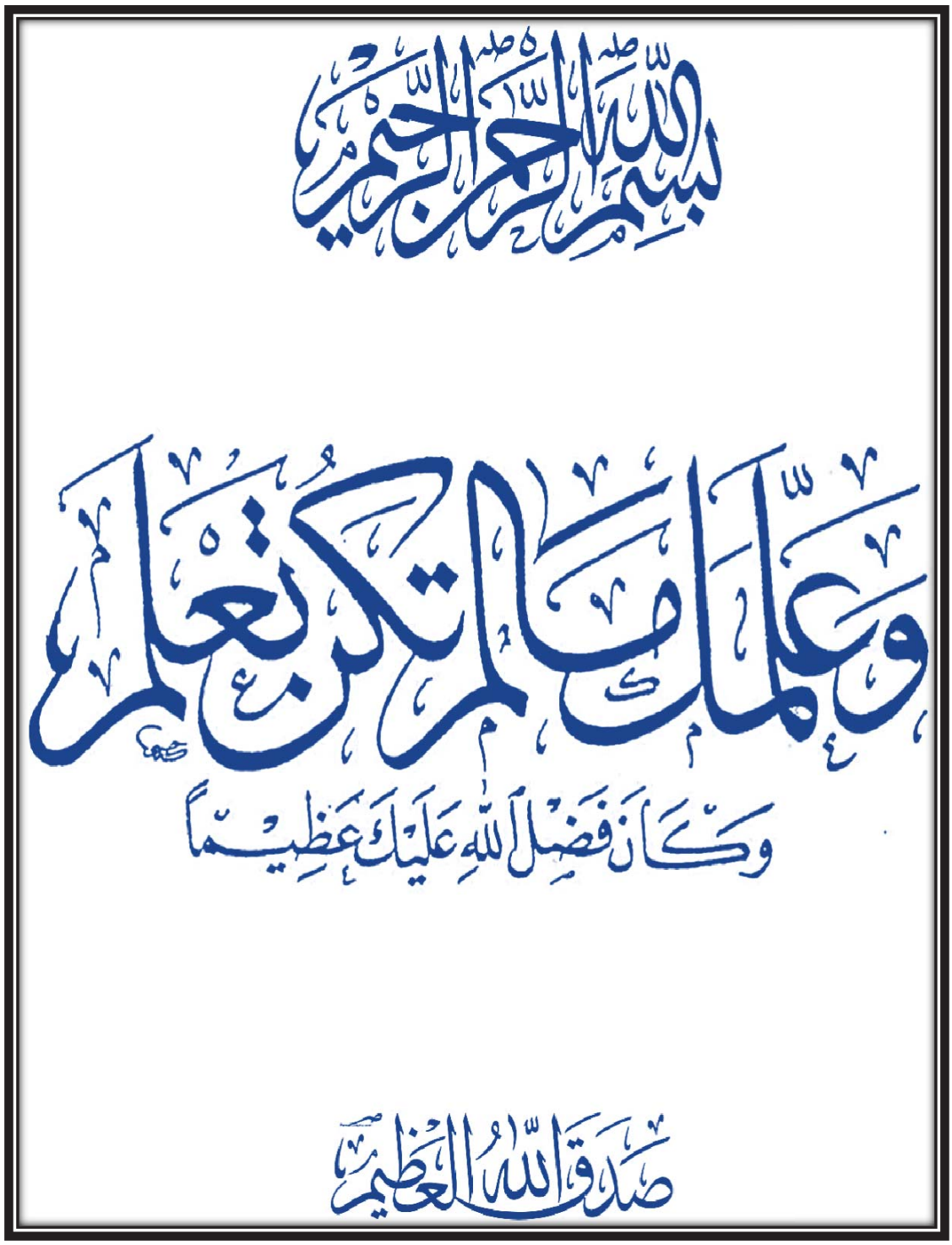
Submitted to the Department of Computer Science\ College of Science\ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Emaan Jalal Khalifa
B.Sc. Computer Science Dept./ Diyala University 2006

Supervised By
Prof. Dr. Dhahir Abdulhade Abdulah

2022 A.D. *IRAQ\ Diyala* *1444 A.H*



سورة النساء

الآية (113) جزء (5)

Dedication

I would like to dedicate this work to

To the one who honored me by bearing his name

To the one who gave the most precious things to reach this degree (my beloved father), may God prolong his life.

To her prayers and words were companions of brilliance and excellence (my beloved mother); may God prolong his life.

To my brother, the martyr (Mustafa), may God have mercy on him; thanks to him, I was accepted into the Master's degree.

To my brother and sisters, the ones who gave me unstopped support, help, and encouragement when I was disappointed,

To my soul, my life, and my happiness, my daughter, Rawnaq.

Emaan Jalal Khalifa

Acknowledgments

First and foremost, thanks to Allah SWT for his blessing and mercy who has guided me in completing this thesis.

I wish to express my thanks and deep gratitude to my supervisor ***Prof. Dr. Dhahir Abdulhade Abdulah*** *for his supervision of this research, and for his invaluable advice, criticism, encouragement, help, and trust throughout the development of this work to be in the best manner.*

Also, I extend grateful thanks to my college (College of Science), all the professors and staff of the Department of Computers Science at the University of Diyala, who contributed and helped me in any way to the success and completion of this study, ***especially*** ***Prof.Dr. Ziyad Tariq Mustafa Al-Ta'i.***

Finally, I am sincere gratitude thanks to my family for their unlimited love, encouragement, patience, and support during all stages of my study which helped me in achieving my goals.

Thank you all!

Emaan Jalal Khalifa

Linguistic Certification

I certify that this thesis entitled **"**Classification Method of Advanced Persistent Threat (APT) Malware Using Advanced Machine Learning" *was prepared by* "Emaan Jalal Khalifa" *and was reviewed linguistically. Its language was amended to meet the style of the English language.*

Signature:

Name: Dr.Ghazwan Mohammed Jaffar

Date: / / 2022

Scientific Certification

*I certify that this thesis entitled "*Classification Method of Advanced Persistent Threat Malware Using Advanced Machine Learning*" was prepared by "*Emaan Jalal Khalifa*" and has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.*

Signature:

Name: Assist. Prof. Mohammed Najm Abdulla

Date: / / 2022

Scientific Certification

*I certify that this thesis entitled "*Classification Method of Advanced Persistent Threat Malware Using Advanced Machine Learning*" was prepared by "*Emaan Jalal Khalifa*" and has been evaluated scientifically; therefore, it is suitable for debate by the examining committee.*

Signature:

Name: Assist. Prof. Dr. Ibrahim Zeghaiton Chaloob

Date: / / 2022

Supervisors Certification

*We certify that this research entitled "*Classification Method of Advanced Persistent Threat (APT) Malware Using Advanced Machine Learning*" was prepared by "*Emaan Jalal Khalifa*" under my supervision at the University of Diyala collage of Science Department of Computer Science as partial fulfillment of the requirements needed to award the degree of Master of Science in Computer Science.*

Signature:

Name: Prof.Dr. Dhahir Abdulhade Abdulah

Date: / / 2022

Approved by the University of Diyala Faculty of Science Department of Computer Science.

Signature:

Name: Assist. Prof. Dr. Bashar Talib Al-Nuaimi

Date: / / 2022

Title: Head of Computer Science Department

Examination Committee Certification

*We certify that we have read the thesis entitled "*Classification Method of Advanced Persistent Threat (APT) Malware Using Advanced Machine Learning*" and, as an examination committee, examined the student "*Emaan Jalal Khalifa*" in the thesis content and that, in our opinion, it is adequate as fulfill the requirement for the Degree of Master in Computer Science at the Computer Science Department, University of Diyala.*

(Chairman)

Signature:

Name: Prof Dr. Abdul Monem S. Rahma
Date: / / 2022

Signature:

Name: ***Assist. Prof. Dr. Thekra Hyder A. Abbas*** ***(Member)***
Date: / / 2022

Signature:

Name: ***Assist. Prof. Dr. Abdulbasit Kadhim Shukur*** ***(Member)***
Date: / / 2022

Signature:

Name: ***Prof Dr. Dhahir Abdulhade Abdullah***
(Supervisor)
Date: / / 2022

Approved by the Dean of the College of Science, University of Diyala

(The Dean)

Signature:

Name: Prof. Dr. Tahssen Hussein Mubarak
Date: / / 2022

ABSTRACT

Advanced Persistent Threat (APT) is a complex type of attack that steals personal data by staying in the infected system for a long time. APT represents sophisticated attacks that are executed in multiple steps. Recently, detecting and classifying APT attacks using Machine Learning (ML) or Deep Learning (DL) algorithms has become a common approach for analyzing network traffic for signals and anomalous behaviors. APT attack detection approach that uses behavior analysis and evaluation approaches encounter many issues. Network traffic analysis to detect a common APT attack is one of the solutions for dealing with this situation.

In this thesis, we propose two systems for classifying APT malware using advanced machine learning. The first, binary-class classification identifies two-class APT malware and normal malware; the second, multi-class classification identifies 15 APT malware and normal malware. Moreover, each system has two classification subsystems: ML based on Random Forest Classifier (RFC), Light Gradient Boosting Machine (LightGBM), and DL using a hybrid Convolution Neural Network (CNN) with Long Short-Term Memory networks (LSTM).

The main methods used are Exploratory Data Analysis (EDA) for detecting and removing outlier data, Extra Tree Classifier (ETC) for selecting essential features, and Synthetic Minority Oversampling Technique (SMOTE) for solving the unbalance data problem.

A reliable APT Malware dataset with 11,107 samples spread over 16 unique malware classes. Each proposed system is studied separately, and the performance results of the machine and deep learning algorithms are compared based on the accuracy value. At the same time, four case studies were conducted

to evaluate the performance of ML algorithms and the impact of using Feature Selection (FS) and SMOTE technology on their results.

The machine learning results demonstrated the significant impact of feature selection and SMOTE technology on the performance of both proposed systems. The binary class classification system results show that machine learning has better performance than deep learning, with the random forest accuracy being around 0.999723 and LightGBM accuracy being 0.999480, while the CNN-LSTM hybrid has an accuracy of around 0.914798. The results of the multi-class classification system illustrated that machine learning has the best performance than deep learning; the LightGBM accuracy is about 0.999727, the random forest has an accuracy of about 0.999632, while hybrid CNN-LSTM achieved an accuracy of about 0.798206.

Furthermore, the comparison of the results of the proposed systems with the results corresponding to the previous works proved that the two proposed systems have the highest classification accuracy, and this indicates the effectiveness of the proposed system for detecting and classifying the attack with a high determination and avoiding the significant risks that it causes.

List of Content

Contents		*Page No.*
Abstract		I
Lists of Contents		III - V
List of Abbreviations		V - VI
List of Algorithms		VI
List of Figures		VII-IX
List of Figures in Appendix (A, B, C)		X - XI
List of Tables		XII
Chapter One: Introduction		**1 - 7**
1.1	Overview	1 - 2
1.2	Related Works	2 - **4**
1.3	Problem Statement	5
1.4	Aim of Thesis	**5**
1.6	Layout Of Thesis	5-6
Chapter Two: Theoretical Background		**8 - 37**
2.1	Introduction	8
2.2	Malware	8 - 9
2.2.1	Malware Analysis	9
2.2.2	Malware Detection	10
2.3	Advanced Persistent Threat (Apt)	11
2.3.1	A Formal Definition of Advanced Persistent Threat	11
2.3.**2**	Basic Characteristics of Advanced Persistent Threats	11- 13
2.3.**3**	Detection Of APT Attacks	13
2.3.**4**	APT Groups and Operation	14
2.3.**5**	Apt Life Cycle	14 - 15
2.4	Data Pre-Processing.	16
2.4.1	Data Cleaning	17
2.4.2	Dropping Duplicate Raw	17
2.5	Exploratory Data Analysis (EDA)	17
2.5.1	Outlier	17 - 19
2.6	Data Scaling	19
2.7	One-Hot Encoder (OHE)	20
2.7	Synthetic Minority Oversampling Technique (Smote)	20
2.8	Feature Selection (FS)	21
2.8.1	Extra Tree Classifier (ETC)	21
2.9	Cross Validation	21 - 22
2.10	Classification Algorithm	23

2.11	Machine Learning	22 - 27
2.11.1	Light Gradient Boosting Machine	23 - 25
2.11.1.1	Advantages of Light Gradient Boosting Machine	24
2.11.1.2	Parameter of Light Gradient Boosting Machine	25
2.11.2	Random Forest Classifier (RFC)	25 - 27
2.11.2.1	Parameters of Random Forest Algorithm	27
2.12	Deep Learning (DL)	28 - 35
2.12.1	Convolutional Neural Network (CNN)	28 - 33
2.12.1.1	Basic Structure Of CNN	29 - 31
2.12.1.2	Convolution Neural Network Elements	31 - 33
2.12.2	Long Short-Term Memory (LSTM)	33 - 35
2.12.3	CNN–LSTM Model	35
2.13	Overfitting	36
2.14	Confusion Matrix (Performance)	36-37
Chapter Three: The Proposed System		***38-57***
3.1	Introduction	38
3.2	Proposed System	38 - 39
3.2.1	Information Dataset	39 - 40
3.2.2	Cleaning / Preprocessing Dataset	40 - 41
3.2.3	Constructs Classes Based on Behavioral of The Malware Dataset	42 - 43
3.2.3.1	Binary–Class Dataset	43
3.2.3.2	Multi –Classes Dataset	44
3.2.4	Feature Selection Using Extra Trees Classifier	43 - 44
3.2.5	Cross-Validation	45
3.2.6	Balance Training Dataset Using Smote Technique	45 - 47
3.2.7	Classification Models	47 - 57
3.2.7.1	Random Forest and LightGBM Classification	47 - 49
3.2.7.2	CNN –LSTM Deep Learning Classification	50 - 57
Chapter Four: Experimental Results and Evaluation		***58-111***
4.1	Introduction	58
4.2	Implementation Environment	58
4.3	Dataset	58 - 59
4.4	Results of the Proposed System	60
4.4.1	The Results of Preprocessing Phase	60 - 65
4.4.2	The Results of Constructs Classes Phase	66 - 67
4.4.3	The Results of Feature Selection Using Extra Trees Classifier Phase	68
4.4.4	Cross Validation Phase	68
4.5	Results of Balance Dataset Phase	69 – 70

4.6	Performance Results for The Classification Phase	70
4.6.1	Results of The Binary–Class Classification System	71 - 96
4.6.2	Results of Multi-Class Classification System	97 - 110
4.4	Comparison With Previous Related Studies	111
Chapter Five: Conclusions and Suggestions		***112 - 114***
5.1	Conclusion	112-114
5.2	Suggestions For Future Works	114
	References	115- 122

List of Abbreviations

Abbreviations	***Full Form***
1D CNN	1-Dimensional Convolutional Neural Network
AI	Artificial Intelligent
APT	Advanced Persistent Threat
BN	Batch Normalization
C&C	Command And Control
CL	Convolutional Layer
CNN	Convolutional Neural Network
CV	Cross Validation
DL	Deep Learning
DNS	Domain Name System
DT	Decision Tree
EDA	Exploratory Data Analysis
EFB	Exclusive Feature Bundling
ET	Extra Tree
ETC	Extra Tree Classifier
FN	False Negatives
FP	False Positive
FS	Feature Selection
F_t	Forget Gate
GBDT	Gradient Boosting Decision Tree
GOSS	Gradient-Based One-Side Sampling
GRU	Gated Recurrent Unit
IDS	Intrusion Detection System

I_t	Input Gate
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
NIST	National Institute of Standards and Technology
OHE	One-Hot Encoder
Organization	Computer Network Used
O_t	Output Gate
OTP	One-Time Password
ReLU	The Rectified Linear Unit
RF	Random Forest
RFC	Random Forest Classifier
SMOTE	Synthetic Minatory Oversampling Tree Technique
TN	Tue Negatives
TP	True Positives
XGBoost	eXtreme Gradient Boosting
σ	Standard Deviation
μ	Mean

List of Algorithm

Algorithm No.	*Algorithm Title*	*Page No.*
Algorithm (2.1)	Random Forests Classifier	26
Algorithm (3.1)	Remove Outlier	41
Algorithm (3.2)	Extra Tree Classifier for Feature Selection	44
Algorithm (3.3)	Synthetic Minority Oversampling Technique SMOTE	46
Algorithm (3.4)	Random Forest Algorithm for Classification APT Malware Attack	48
Algorithm (3.5)	Light GBM Classification Algorithm	49
Algorithm (3.6)	Proposed Hybrid CNN-LSTM Classification Algorithm	57

List of Figures

Figure No.	*Descriptions*	*Page No.*
Figure (2.1)	Different Types of Malware	9
Figure (2.2)	Organization Of Malware Detection Techniques	10
Figure (2.3)	Attack Procedure	13
Figure (2.4)	Typical Stages of An Apt Attack	15
Figure (2.5)	Typical Data Preprocessing Tasks for Building Operational Data Analysis	16
Figure (2.6)	Box-And-Whiskers Plot	18
Figure (2.7)	An Illustration of the K-Fold Technique Using K = 5	23
Figure (2.8)	Special process of LGB algorithm. (a) Histogram-based algorithm; (b) Obtain difference value by histogram value; (c) Level-wise and leaf-wise strategies	24
Figure (2.19)	Random Forest Classifier	26
Figure (2.10)	Block Diagram of Deep Learning	28
Figure (2.11)	A Typical Architecture of the CNN	30
Figure (2.12)	2 X 2 Size of Max Pooling	32
Figure (2.13)	Example of a Fully-Connected Neural Network	33
Figure (2.14)	Commonly used activation functions: (a) Sigmoid, (b) Tanh, (c) ReLU, and (d) LReLU	33
Figure (2.15)	Basic Structure of The LSTM Model	34
Figure (2.16)	The Architecture of The Hybrid CNN-LSTM Model.	36
Figure (3.1)	Block Diagram of The Proposed Systems	40
Figure (3.2)	Constructs Classes in The Binary-Class System	43
Figure (3.3)	Constructs Classes in The Multi-Class System.	44
Figure (3.4)	Block Diagram of The Coding Classes in The Training Apt Malware Attack.	51
Figure (3.5)	Label Encode for Binary-Class and Multi-Class Classification Systems	52
Figure (3.6)	One-Hot Encoding Technique	53
Figure (3.7)	Block Diagram of the Proposed Hybrid CNN-LSTM Network Algorithm	55
Figure (4.1)	Dataset After Check Missing Data.	61
Figure (4.2)	2D Histogram of the Relation between Attributes and Classes in the Dataset	62
Figure (4.3)	3D Histogram Of The Relation Between Attributes And Classes In The Dataset.	63
Figure (4.8)	A Histogram of Outliers Identification and Processing for Six Features in the Input Data Set	65

Figure (4.9)	The Data Description for The Binary Class Dataset Includes the APT Malware Attack Class &The Normal Malware Attack, Class	66
Figure (4.10)	Description of The Multi-Class Dataset.	67
Figure (4.11)	Selection of Relevant Features from Dataset using Extra Tree Classifier Algorithm	68
Figure (4.12)	Applying SMOTE Technique to a binary–class classifier system, the horizontal line indicates the class name and the vertical line indicates the number of the sample in the Training Dataset	69
Figure (4.13)	Applying SMOTE Technique on Multi-Class Classifier System, the Horizontal Line Indicates the Class Name &the Vertical Line Indicates the Number of the Sample in the Training Dataset	70
Figure (4.14)	Confusion Matrix for the Binary–Class Classifier system using Random Forest Algorithm without Feature Selection and SMOTE Technique	72
Figure (4.15)	Confusion Matrix for the 8- Fold Cross Validation Using Random Forest Algorithm Without Feature Selection and Without SMOTE Technique	135
Figure (4.16)	Confusion Matrix for the Binary–Class Classifier system using LightGBM Algorithm without Feature Selection and SMOTE Technique	73
Figure (4.18)	Histogram of Mean Accuracy using Random Forest Algorithm	74
Figure (4.19)	Histogram Of Mean Accuracy Using LightGBM Algorithm	74
Figure (4.20)	Confusion Matrix for The Binary–Class Classifier System Using Random Forest Algorithm Without Feature Selection and With Smote Technique	76
Figure (4.22)	Confusion Matrix for The Binary–Class Classifier System Using LightGBM Algorithm Without Feature Selection and With Smote Technique	77
Figure (4.23)	Confusion Matrix for the 8- Fold Cross Validation using Light GBM Algorithm without Feature Selection and with SMOTE Technique	141
Figure (4.24)	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case (red) in Case of without Feature Selection and with SMOTE using Random Forest.	78
Figure (4.25)	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case	78

	(red) in Case without Feature Selection and with SMOTE using Light GBM	
Figure (4.26)	Check Overfitting of the Random Forest with Training Score (Black) vs. the Cross–Validation Score (Red) in Case of without Feature Selection and with SMOTE Technique	79
Figure (4.28)	Check Overfitting of the LightGBM with Training Score (Black) vs. the Cross–Validation Score (Red) in Case of without Feature Selection and with SMOTE Technique	80
Figure (4.30)	Confusion Matrix for the Binary–Class Classifier system using Random Forest Algorithm with Feature Selection and without SMOTE Technique	82
Figure (4.32)	Confusion Matrix for the Binary–Class Classifier system using Light GBM Algorithm with Feature Selection and Without SMOTE Technique	83
Figure (4.34)	Histogram with Feature Selection and Without SMOTE using Random Forest	84
Figure (4.35)	Histogram with Feature Selection and Without SMOTE using LightGBM.	84
Figure (4.36)	Confusion Matrix for the Binary–Class Classifier system using Random Forest Algorithm with Feature Selection and SMOTE Technique	86
Figure (4.38)	Confusion Matrix for the Binary–Class Classifier system using LightGBM Algorithm with Feature Selection and SMOTE technique	87
Figure (4.40)	Histogram with Feature Selection and SMOTE Using RF	88
Figure (4.41)	Histogram with FS and SMOTE Using LightGBM.	88
Figure (4.42)	Check Overfitting of the Random Forest with FS and SMOTE.	89
Figure (4.44)	Check Overfitting of the LightGBM with FS and Smote	90
Figure (4.46)	Optimal Accuracy Value for the Random Forest and LightGBM of the Binary Class Classification System Over Four Cases Study	92
Figure (4.47)	Constructed Hybrid CNN-LSTM Model with Binary Class Classification System	93
Figure (4.48)	The Loss of the Proposed Hybrid CNN+LSTM.	94
Figure (4.49)	Show the Accuracy Of CNN+LSTM.	95
Figure (4.50)	Performance comparison Machine Learning and Deep Learning for Binary-Class Classifier System.	96

Figure (4.51)	Fold -0 Confusion Matrix for the Multi–Class Classifier System using Random Forest Algorithm Without Feature Selection and SMOTE Technique	98
Figure (4.53)	Confusion Matrix for the Multi–Class Classifier System using LightGBM Algorithm without Feature Selection and SMOTE Technique	99
Figure (4.55)	Histogram of Mean Accuracy using Random Forest	100
Figure (4.56)	Histogram of Mean Accuracy using LightGBM	100
Figure (4.61)	Histogram without Feature Selection and with SMOTE Using Multi-Class System and Random Forest	102
Figure (4.62)	Histogram without Feature Selection and with SMOTE Multi-Class System and LightGBM	102
Figure (4.73)	Optimal Accuracy Value for the Random Forest and LightGBM of the Binary Class Classification System Over Four Cases Study.	106
Figure (4.74)	Constructed Hybrid CNN-LSTM Model with Multi-Class Classification System	107
Figure (4.75)	The loss function of the multi-class using CNN with LSTM	109
Figure (4.76)	The Accuracy of the Multi classification System	109
Figure (4.77)	Performance comparison Machine Learning and Deep Learning for Multi-Class Classifier System	110

List of Figures in Appendix (A,B,C)

Figure No.	***Descriptions***	***Page No.***
Figure (A.4)	2D Histogram of the Relation between Attributes	125
Figure (A.5)	3D Histogram of the Relation Between Attributes	125
Figure (A.6)	2D Histogram of the Relation between Attributes	126
Figure (A.7)	3D Histogram of the Relation Between Attributes	126
Figure (B.17)	Confusion Matrix for the 8- Fold Cross Validation using Light GBM Algorithm without Feature Selection and Without SMOTE Technique	135
Figure (4.18)	Histogram of Mean Accuracy using Random Forest Algorithm	74
Figure (4.19)	Histogram Of Mean Accuracy Using LightGBM Algorithm	74
Figure (4.20)	Confusion Matrix for The Binary–Class Classifier System Using Random Forest Algorithm Without Feature Selection and With Smote Technique	76

Figure (B.21)	Confusion Matrix for the 8- Fold Cross Validation using Random Forest Algorithm without Feature Selection and with SMOTE technique	139
Figure (B.27)	Overfitting Data of the Random Forest in Case of without Feature Selection and with SMOTE	143
Figure (B.29)	Overfitting Data of the LightGBM in Case of without Feature Selection and with SMOTE	145
Figure (B.31)	Confusion Matrix for the 8- Fold Cross Validation using Random Forest Algorithm with Feature Selection and without SMOTE Technique	147
Figure (B.33)	Confusion Matrix for the 8- Fold Cross Validation using Light GBM Algorithm with Feature Selection and without SMOTE Technique	150
Figure (B.37)	Confusion Matrix for the 8- Fold Cross Validation using Random Forest Algorithm with Feature Selection and with SMOTE Technique	153
Figure (B.39)	Confusion Matrix for the 8- Fold Cross Validation using Light GBM Algorithm with Feature Selection and with SMOTE Technique	155
Figure (B.43)	Overfitting Data of the Random Forest in Case of Feature Selection and with SMOT	157
Figure (B.45)	Overfitting Data of the LightGBM in Case of Feature Selection and with SMOTE.	160
Figure (C.52)	Confusion Matrix for the 9- Fold Cross Validation Using Multi-Class System and Random Forest Algorithm Without Feature Selection and SMOTE Technique	162
Figure (C.54)	Confusion Matrix for the 9- Fold Cross Validation using Light GBM Algorithm without Feature Selection and SMOTE Technique.	165
Figure C.57	Confusion Matrix for the 10- Fold Cross Validation Using Multi-Class System and Random Forest Algorithm Without Feature Selection and With SMOTE Technique	167
Figure (C.58)	Confusion Matrix for The10- Fold Cross Validation Using Multi-Class System and Light GBM Algorithm Without Feature Selection and With SMOTE Technique	170
Figure (C.59)	Check Overfitting of The Multi-Class System and Random Forest with Training Score (Black) Vs. The Cross-Validation Score (Red) In Case of Without Feature Selection and with SMOTE	173

Figure C.60	Check Overfitting of the Multi-Class System and LightGBM with Training Score (Black) vs. the Cross–Validation Score (Red) in Case of without Feature Selection and with SMOTE.	176
Figure C.63	Confusion Matrix for the 10- Fold Cross Validation using Multi-Class System and Random Forest Algorithm with Feature Selection and without SMOTE Technique	179
Figure C.64	Confusion Matrix for the 10- Fold Cross Validation using Multi-Class System and Light GBM Algorithm with Feature Selection and without SMOTE Technique.	181
Figure C.65	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case (red) in Case of with Feature Selection and without SMOTE using Multi-Class System and Random Forest.	182
Figure C.66	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case (red) in Case of with Feature Selection and without SMOTE using Multi-Class System and LightGBM.	182
Figure C.67	Confusion Matrix for the Multi–Class Classifier System using Random Forest Algorithm with Feature Selection and SMOTE technique	185
Figure C.68	Confusion Matrix for the Multi–Class Classifier System using LightGBM Algorithm with Feature Selection and SMOTE Technique	188
Figure C.69	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case (red) in Case of with Feature Selection and SMOTE using Multi-Class System and Random Forest.	188
Figure C.70	Histogram of Mean Accuracy Cross-Validation V-Fold Values with Error Bars (Blue) vs. the Ideal Case (red) in Case of with Feature Selection and SMOTE using Multi-Class System and LightGBM.	189
Figure C.71	Check Overfitting of the Multi-Class System and Random Forest with Training Score (Black) vs. the Cross–Validation Score (Red) in Case of with Feature Selection and SMOTE.	191
Figure C.72	Check Overfitting of the Multi-Class System and LightGBM with Training Score (Black) vs. the Cross–Validation Score (Red) in Case of with Feature Selection and SMOTE	194

List of Tables

Table No.	*Descriptions*	*Page No.*
Table (2.1)	Differences Between APTs and Traditional Malware Attacks	12
Table (2.2)	Summary of Classic Apt Attack Cases	13
Table (2.3)	The main parameters of LightGBM	25
Table (2.4)	Parameters of Random Forest Algorithm	27
Table (2.5)	Two-class confusion matrix	36
Table (4.1)	APT and Normal Malware Samples Dataset	59
Table (4.2)	Details Dataset After and Before Dropping Duplication Rows	60
Table (A.3)	The EDA Statistical Summary of the Dataset Features	126
Table (A.4)	The Results of the Selecting Relevant Features Using Extra Tree Algorithm	131
Table (4.5)	The Accuracy Value of the Binary–Class Classifier System without Feature Selection and SMOTE Technique	74
Table (4.6)	Accuracy Value of the Binary–Class Classifier System without Feature Selection and with SMOTE Technique	75
Table (4.7)	Accuracy Value of the Binary–Class Classifier System with Feature Selection and Without SMOTE Technique	81
Table (4.8)	Accuracy Value of the Binary–Class Classifier System with Feature Selection and SMOTE Technique	85
Table (4.9)	compares random forest and LightGBM with four cases study of the binary class classification system.	91
Table (4.10)	Model Summary of the Proposed Hybrid CNN+LSTM Networks in Binary–Class Classification System.	94
Table (4.11)	Accuracy Value of the Multi–Class Classifier System Without Feature Selection and SMOTE Technique	97
Table (4.12)	Accuracy Value of the Multi–Class Classifier System without Feature Selection and with SMOTE Technique	102
Table (4.13)	Accuracy Value of the Multi–Class Classifier System with Feature Selection and without SMOTE Technique	103
Table (4.14)	Accuracy Value of the Multi–Class Classifier System with Feature Selection and SMOTE Technique	105
Table (4.15)	comparison between Random Forest and LightGBM with Four Cases Study of the Multi Class classification system	105
Table (4.16)	Model Summary of the Proposed Hybrid CNN+LSTM Networks in Multi-Class Classification System	108
Table (4.17)	Comparison Between the Proposed Systems and Related Methods	111

Chapter One
General Introduction

Chapter one
General Introduction

1.1 Overview

In recent years, Advanced Persistent Threats (APTs) have become a new security danger for companies and governments, which is a new type of network attack that can freely use multiple attack techniques. APT attackers use small companies as stepping-stones to gain access to large organizations by avoiding all detection [1]. The most dangerous malware is that developed by APTs, since it applies a high level of sophistication, and targets important victims [2].

APT attacks are a form of attack that uses advanced attack methods to carry out long-term persistent cyber-attacks on specific targets. APT attacks are becoming more frequent. Advanced attack methods, long duration, and a high degree of threat are three main characteristics of APT attacks [3]. APT attacks have two main objectives: one is to steal critical data and the second is to destroy system infrastructure [4].

Advanced Persistent Threat (APT) attack is a persistent, targeted attack on a aparticular organization and is performed through several steps. The primary goals of APT are data exfiltration and espionage. APT is therefore viewed as a newer and more sophisticated type of multi-step attack [5]. APTs are long-term network attacks on specific targets with attackers using advanced attack methods. APTs are more advanced than other forms of attacks. APT is advanced because it uses advanced attack tools and methods [6].

Advanced persistent threat attacks threaten both public and private institutions worldwide and will continue to do so. These attacks offer a severe threat that is hard to see in their early stages because the attackers use various ategies to remain unnoticed as long as possible and to dodge effectively [7].

APT attacks have become the most massive threat to companies and governments as they are increasingly becoming the target of these attacks. APT attacks target the victim's network to gain useful information or to compromise the network to destroy the victim's systems or to steal the target's data without getting caught [8].

In this thesis, a classification method of APT Malware is proposed which uses Machine and Deep learning algorithms. The proposed system divided into two branches and follow same procedure for training and testing the APT Malware dataset with 11,107 samples spread over 16 different unique malware classes: binary, multi class classification systems and each of them include two branches for classification ***Machine learning branch*** based on Random Forest (RF) and Light Gradient Boosting Machine algorithms (LightGBM); ***Deep learning branch*** based on proposed combing hybrid Convolutional Neural Networks (CNN) algorithm and Long Term-Short Memory algorithm (LSTM) are used as a classifier to classify the attack APT malware or normal malware.

The proposed binary–class system obtains an optimal accuracy value of **0.999723** with random forest,**0.99948** with LightGBM, and **0.914798** with hybrid CNN-LSTM. The proposed multi-class system obtains optimal accuracy values of **0.999727** with LightGBM, **0.999632** with random forest, and **0.798208** with hybrid CNN-LSTM.

1.2 Related Work

Several efforts have tried identifying and categorizing the APT problem with increased cyberattacks. This section reviews related work in APT detection and compares with them as shown below:

- ***Ghaffir et al.(2018) [9]:*** The system is composed of three-layers detection, i.e., threat detection, alert correlation, and attack prediction which provided an accuracy of 84.4%. The system needs to be tested for real-time APT signature covering all seven phases of the APT life cycle.

- ***Yan, et al. (2020) [10]:*** This study introduces a new feature that shows the relationship between a DNS request and the response message using deep learning to evaluate DNS request records. Based on the suspicious value, the system assesses DNS activity for threats. This study uses 4, 907, 147, 146 DNS request records (376, 605, 606 after DNS Data Pre-processing) from a large university network to test the system's authenticity and correctness. Experiments reveal that our technique detects suspicious DNS behavior with 97.6% accuracy, 2.3% false positives, and 96.8% recall. The suggested approach detects unusual DNS activities in APT.
- ***Chen, W. et al. ,in (2020) [11]***: In this paper, a new gene model combining malware behavior knowledge graph is proposed. Researchers use malware information to create the APT gene pool. The gene pool should contain APT genetics. Theoretically, genetic traits can help us identify IoT malware and APT. Genetic similarity algorithms can't be employed directly. Instead, a genetic similarity algorithm will identify APT malware.. The experiment is 85% accurate. The model can recognize APT gene organization properties. The program compares sample genes to the gene pool and outputs gene similarity to identify APT.

- ***Laurenza ,G. et al., in (2020) [2]:*** In the paper, the authors offer malware triage for early APT detection. The proposed solution necessitates a significant training time and must be retrained if new APT samples or classes are found. In this paper, they go from multi-class to one-class classification, which reduces run time and increases modularity while maintaining over 90% precision and accuracy.

- ***Zimba ,A., et al. ,in (2020) [7]:*** This article proposes a new APT detection technique based on semi-supervised learning and complex network properties to overcome attack network dynamics. The goal of APT-AN is small-world networking. APT uses finite state machines. The method is supported by information from Los Alamos' 17,684-host enterprise network. The method analyzes huge datasets for C&C and victim APT attack features. APT hosts are ranked. In our framework, three-stage APT detection is 90.5% accurate. Results show model can identify APT hosts.

- ***Zhang, et al. (2021)[12]***. proposed an attribution classification method of APT malware in IoT using the ML approach. The method analyzes samples, pre-processes the acquired behavioral data, constructs a behavioral data set of malware samples, then uses the TF-IDF method to perform the feature representation forming a vector matrix and calculates the chi-square value of the high latitude feature vector to perform feature selection. SMOTE-RF model is used in the multi-class model to train and test sets for predicted output with accuracy of 80%.

- ***C. Do Xuan and M. H. Dao(2021) [13]***.The proposal of a combined deep learning model to detect APT attacks based on network traffic is a new approach, and there is no research proposed and applied yet. In the

experimental section, combined deep learning models proved their superior abilities to ensure accuracy on all measurements from 93 to 98%.

- ***F. J. Abdullayeva (2021)[14]:*** In this study, a deep neural network model built by adding layers was evaluated on a public database and compared to existing methods; the new method showed superior results in detecting APT attacks. This approach uses a machine learning dataset that includes APT1, Crypto, and other attacks. The architect's accuracy was 98.32%.

1.3 Problem Statement

APT attacks are challenging to detect and allow hackers to hide within the network for months. While the hackers remain in the system, the company experiences data loss and outages regularly without knowing the cause of the problems; Traditional detection technologies cannot identify them efficiently.

For the reasons mentioned above, therefore, the ***main problem*** that depends on this work is that detecting and classifying APT malware attack systems, which depend on deep and machine learning, still need an analytical study of the data and improved accuracy in identifying APT malware organizations.

1.4 The aim of the Thesis

The main aim of this work is to design and implement a system for **detection and classification** of the Advanced persistent threat (APT) malware accurately and rapidly based on machine learning and deep learning algorithms, which that help security analysts in government institutions, organizations, and large companies to discover and identify APTs early and accurately to avoid losses caused by APTs attackers as well as develop strategies to counter this type of attack.

1.5 Layout of Thesis

Chapter One (General Introduction)

The first chapter (General Introduction), the other chapters in this thesis are follows as:

Chapter Two (Theoretical Background)

This chapter provides a background and overview about malware, advanced persistent threat (APT), basic characteristics, and APT life cycle, theoretical background and techniques that are used in this thesis.

Chapter Three (Proposed System Design)

This chapter describes the proposed APT malware classification system with its design and implementation.

Chapter Four (Experimental Test Results)

This chapter explains the results and evaluation that have been getting from the proposed system.

Chapter Five (Conclusions and Suggestions for Future Work)

This chapter presents the conclusions of this work. Furthermore, it provides suggestions for future work.