

*Dedicated to*

*My family*

# Declaration

I certify that

- The work contained in the dissertation is original and has been done by myself under the general supervision of my supervisors.
- The work has not been submitted to any other Institute for any degree or diploma.
- I have followed the guidelines provided by Tezpur University in writing the thesis.
- I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the university.
- Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the dissertation and giving their details in the references.

*parthajit borah*

**Parthajit Borah**



Department of Computer Science & Engineering  
Tezpur University

Napaam, Tezpur- 784028, Assam, India.

Dr. Dhruva Kr Bhattacharyya  
Professor

Phone:03712-275353  
E-Mail : dkb@tezu.ernet.in

---

## Certificate

This is to certify that the thesis entitled “**Detection of Malware and Malware-based Attacks using AI Approaches**” submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Parthajit Borah** under my supervision and guidance.

All help he received from various sources has been duly acknowledged. No part of this thesis has been submitted elsewhere for award of any other degree.

Prof. D.K. Bhattacharyya  
Dept. Of Computer Sc. & Engg.  
Tezpur University.

Signature of Supervisor  
(Dhruva Kr Bhattacharyya)  
Professor  
Department of Computer Science and Engineering  
Tezpur University  
Assam, India-784028

## Certificate

This is to certify that the thesis entitled “**Detection of Malware and Malware-based Attacks using AI Approaches**” submitted to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the award of the degree of Doctor of Philosophy in Computer Science and Engineering is a record of research work carried out by **Parthajit Borah** under my supervision and guidance.

All help he received from various sources has been duly acknowledged. No part of this thesis has been submitted elsewhere for award of any other degree.




Signature of Co-Supervisor  
(Jugal Kalita)  
Professor  
Department of Computer Science  
University of Colorado  
Colorado Springs, CO 80918, USA



## Certificate

This is to certify that the thesis entitled “**Detection of Malware and Malware-based Attacks using AI Approaches**” submitted by **Mr. Parthajit Borah** to Tezpur University in the Department of Computer Science and Engineering under the School of Engineering in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Computer Science and Engineering has been examined by us on .....**08/10/2024**..... and found to be satisfactory.

The Committee recommends for award of the degree of Doctor of Philosophy.

  
Prof. D.K. Bhattacharyya  
Dept. Of Computer Sc. & Engg.  
Tezpur University.

Signature of Principal Supervisor

# Acknowledgment

It is a great pleasure to express my gratitude to all those who provided guidance and support in the successful completion of my doctoral program at Tezpur University. Achieving this milestone would not have been possible on my own. Many people extended their supportive hands, contributing to the success of my work. I would like to sincerely thank everyone who supported and assisted me throughout this journey at Tezpur University.

I am deeply indebted to my supervisor, Prof. Dhruba Kr. Bhattacharyya, for his invaluable guidance, patience, and encouragement throughout my PhD journey. His expertise and insights have been instrumental in shaping my research and helping me overcome numerous challenges. His unwavering support and belief in my abilities have been a constant source of motivation, and I am truly grateful for his mentorship and dedication.

I would also like to extend my heartfelt gratitude to my co-supervisor, Prof. Jugal Kalita, for his continuous support and constructive feedback. His detailed and insightful suggestions have greatly improved the quality of my work. His availability and willingness to assist at every stage of my research have been incredibly helpful. I deeply appreciate his commitment and contributions to my academic and professional growth.

Additionally, I am profoundly thankful to all the lab members and friends who have been a part of this journey. Their camaraderie, encouragement, and support have made this challenging path more enjoyable and manageable. I am grateful for the insightful discussions, collaborative spirit, and moments of shared laughter and perseverance.

I would like to express my sincere appreciation to the faculty of the Department of Computer Science and Engineering at Tezpur University. Their assistance and support in various administrative and academic matters have been invaluable, and I am truly thankful for their contributions.

---

I sincerely thank the members of my thesis review committee and the anonymous reviewers for their precious comments and feedback. Their constructive criticism and valuable suggestions have significantly improved the quality of my work.

A very special mention goes to my parents, who have constantly encouraged and supported me in every walk of my life. Their unwavering faith and love have been my foundation, and I will always be grateful for their sacrifices and guidance. I extend my heartfelt thanks to all my family members for their endless support, understanding, and encouragement throughout this journey.

I would like to thank the Almighty for providing me with the strength, patience, and perseverance needed to complete this journey. Without His blessings, none of this would have been possible.

Finally, I would like to thank those who have directly or indirectly helped me complete my research work in different capacities.

**Parthajit Borah**

# List of Figures

1-1	Stages of a malware attack . . . . .	5
1-2	Detection Approaches . . . . .	6
2-1	A conceptual framework of client-server architecture . . . . .	15
2-2	A conceptual framework of peer-peer architecture . . . . .	15
2-3	Taxonomy of malware . . . . .	20
2-4	Types of malware analysis . . . . .	27
2-5	Flow diagram of dynamic analysis . . . . .	28
2-6	A generic architecture of an automated analysis system . . . . .	29
2-7	Actors involved in MaaS ecosystem . . . . .	31
2-8	Artificial Neural Network . . . . .	37
2-9	The local receptive field for a neuron . . . . .	38
2-10	Illustration of max pooling . . . . .	39
2-11	Graph Neural Network Architecture . . . . .	40
3-1	Dataset Generation and Evaluation Pipeline . . . . .	52
3-2	Testbed Architecture for malware collection . . . . .	52
3-3	Steps in data analysis phase . . . . .	54
3-4	Example of API call information . . . . .	54



## List of Figures

---

3-5	TUANDROMD: Dataset creation framework . . . . .	58
4-1	Feature selection methods . . . . .	66
4-2	Feature selection methods . . . . .	69
4-3	Workflow of the proposed method . . . . .	78
4-4	Performance of FSR in terms of accuracy . . . . .	81
4-5	Performance of FSR in terms of accuracy . . . . .	81
4-6	Performance of FSR in terms of accuracy . . . . .	82
4-7	Performance of FSR in terms of accuracy . . . . .	82
4-8	Performance of FSR in terms of accuracy . . . . .	83
4-9	Performance of FSR in terms of F1 score . . . . .	83
4-10	Performance of FSR in terms of F1 score . . . . .	83
4-11	Performance of FSR in terms of F1 score . . . . .	84
4-12	Performance of FSR in terms of F1 score . . . . .	84
4-13	Performance of FSR in terms of F1 score for Parkinson dataset . . . . .	84
5-1	Taxonomy of Ransomware . . . . .	92
5-2	A generic architecture of malware analysis framework . . . . .	94
5-3	ERAND Detection framework for Ransomware and its variants . . . . .	96
5-4	Number of features vs Classification Accuracy . . . . .	106
6-1	Overview of the proposed method . . . . .	116
6-2	Ensemble feature selection framework . . . . .	117
6-3	Performance of FRAMC in terms of accuracy . . . . .	122
6-4	Optimal range of the size of feature subsets . . . . .	122

## List of Figures

---

7-1	Illustration of the k-NN algorithm in 2D . . . . .	130
7-2	Illustration of the CUDA concept and workflow . . . . .	132
7-3	Framework of the proposed work . . . . .	134
7-4	K vs accuracy graph for Euclidean Distance . . . . .	141
7-5	K vs accuracy graph for Manhattan Distance . . . . .	142
7-6	K vs accuracy graph for Kulczynski Distance . . . . .	143
7-7	K vs accuracy graph for Chebyshev Distance . . . . .	144
7-8	K vs accuracy graph for Cosine similarity measure . . . . .	145
7-9	K vs accuracy graph for Sorgel distance measure . . . . .	146
7-10	K vs accuracy graph for soreson distance measure . . . . .	147
7-11	K vs accuracy graph for tanimoto distance measure . . . . .	148
7-12	K vs accuracy graph for Euclidean distance measure . . . . .	149
7-13	K vs accuracy graph for Manhattan distance measure . . . . .	150
7-14	K vs accuracy graph for Kulczynski distance measure . . . . .	151
7-15	K vs accuracy graph for Chebyshev distance measure . . . . .	152
7-16	K vs accuracy graph for Cosine similarity measure . . . . .	153
7-17	K vs accuracy graph for Sorgel distance similarity measure . . . . .	154
7-18	K vs accuracy graph for Sorenson distance similarity measure . . . . .	155
7-19	K vs accuracy graph for Tanimoto distance similarity measure . . . . .	156
7-20	K vs accuracy graph for Euclidean distance similarity measure . . . . .	157
7-21	K vs accuracy graph for Manhattan distance similarity measure . . . . .	158
7-22	K vs accuracy graph for Kulczynski distance similarity measure . . . . .	159
7-23	K vs accuracy graph for Chebyshev distance similarity measure . . . . .	160

## List of Figures

---

7-24	K vs accuracy graph for Cosine distance similarity measure . . . . .	161
7-25	K vs accuracy graph for Sorgel distance similarity measure . . . . .	162
7-26	K vs accuracy graph for Sorenson distance similarity measure . . . . .	163
7-27	K vs accuracy graph for Tanimoto distance similarity measure . . . . .	164
7-28	Accuracy of Binary and Multi-Class Classification on Ransomware Dataset . . . . .	165
7-29	Accuracy SWaT Dataset . . . . .	165
7-30	CPU vs GPU Time Comparison for Binary Ransomware Dataset . . . . .	165
7-31	CPU vs GPU Time Comparison for Multi-Class Ransomware Dataset . . . . .	166
7-32	CPU vs GPU Time Comparison for Swat Dataset . . . . .	166
8-1	Neural Network Architecture . . . . .	171
8-2	Convolutional Neural Network (CNN) Architecture . . . . .	174
8-3	Overview of the proposed method . . . . .	175
8-4	Feature Data Generation . . . . .	176
8-5	Web-based tool for binary to image conversion . . . . .	177
8-6	Example of Airpush malware visualized as image . . . . .	177
8-7	Illustration of transfer learning . . . . .	180
8-8	Class Distribution of TUANDROMD-X . . . . .	182
8-9	Top 10 categories of TUANDROMD-X . . . . .	183
8-10	Model Performance Comparison on Dataset 1 . . . . .	185
8-11	Model Performance Comparison on Dataset 2 . . . . .	186
9-1	A simple graph with directed and undirected edges . . . . .	191
9-2	Function Call Graph . . . . .	191

## List of Figures

---

9-3	Graph Neural Network Architecture . . . . .	192
9-4	Overview of the proposed method . . . . .	198
9-5	Example of Function Calls . . . . .	199
9-6	Process of binary to FCG generation . . . . .	199
9-7	Application analysis results . . . . .	199
9-8	Web-based tool for binary to FCG conversion . . . . .	200
9-9	Example of malware binary data visualized as a FCG . . . . .	200
9-10	Pipeline of a GNN Model for Malware Detection . . . . .	201
9-11	Class Distribution of TUANDROMD-FCG . . . . .	205
9-12	Top 10 categories of TUANDROMD-FCG . . . . .	206
9-13	Performance of GNN Models on Dataset 1 . . . . .	207
9-14	Performance of GNN Models on Dataset 2 . . . . .	208

# List of Tables

2.1	Malware Propagation Methods . . . . .	22
2.2	Common Malware Evasion Techniques . . . . .	22
2.3	Malware Infectors . . . . .	24
3.1	List of top ranked feature categories for TUMALWD . . . . .	55
3.2	Network level features . . . . .	57
3.3	List of top-ranked features for TUANDROMD . . . . .	60
3.4	Benchmark on TUMALWD and TUANDROMD . . . . .	61
3.5	Datasets and Their Availability Sources . . . . .	63
4.1	A data table . . . . .	74
4.2	Relevance score of each feature . . . . .	75
4.3	Dataset details . . . . .	80
4.4	Top ten features of the datasets . . . . .	85
4.5	Top five features of the datasets . . . . .	85
4.6	Top five features of the datasets . . . . .	86
4.7	Top ten features of the datasets . . . . .	86
4.8	Top ten features for the Parkinson dataset . . . . .	86

## List of Tables

---

5.1	Ransomware dataset with two classes . . . . .	96
5.2	Ransomware dataset with Normal and 11 ransomware subclasses . . . . .	97
5.3	Number of optimal features for each ransomware dataset . . . . .	100
5.4	List of selected features for dataset 1 . . . . .	101
5.5	List of top ranked feature categories for dataset 2 . . . . .	101
5.6	Weightage of the classifiers given by NSGA-II . . . . .	104
5.7	Classification accuracies of dataset 1 . . . . .	105
5.8	Classification accuracies of dataset 2 for each variant . . . . .	105
5.9	Classification accuracies of whole ransomware family . . . . .	106
5.10	Precision, Recall, and F1 score of all the classifiers . . . . .	108
5.11	Comparison of the proposed method with existing methods . . . . .	109
6.1	Top 10 Features Selected by FRAMC for Windows Dataset . . . . .	120
6.2	Top 10 Features Selected by FRAMC for Android Dataset . . . . .	120
6.3	Dataset details . . . . .	121
6.4	The classification accuracy of the Windows dataset using the complete feature space . . . . .	123
6.5	The Precision, Recall, and F1 score of the Windows dataset using the complete feature space . . . . .	123
6.6	The performance comparison on the Windows dataset . . . . .	123
6.7	Precision, Recall and F1 -Score on the Windows Dataset . . . . .	123
6.8	The classification accuracy of the Android dataset using the complete feature space . . . . .	124
6.9	The Precision, Recall and F1 score of the Android dataset using the complete feature space . . . . .	124
6.10	The performance comparison on the Android dataset . . . . .	124

6.11	Precision, Recall and F1 -Score on the Android Dataset . . . . .	124
7.1	Distance measures and their mathematical expressions [1] [2] . . . . .	135
7.2	Ransomware dataset characteristics . . . . .	138
7.3	Characteristics of 20 datasets obtained from UCI repository . . . . .	139
7.4	Value of K for which the maximum accuracy is achieved for the 20 UCI ML Repository Dataset . . . . .	140
7.5	Results of Euclidean Distance . . . . .	141
7.6	Time Comparison between CPU and GPU . . . . .	141
7.7	Result for Manhattan Distance . . . . .	142
7.8	Time Comparison between CPU and GPU . . . . .	142
7.9	Result for Kulczynski distance . . . . .	143
7.10	Time Comparison between CPU and GPU . . . . .	143
7.11	Results for Chebyshev Distance . . . . .	144
7.12	Time Comparison between CPU and GPU . . . . .	144
7.13	Result for Cosine Similarity . . . . .	145
7.14	Time Comparison between CPU and GPU . . . . .	145
7.15	Result for Soergel Distance . . . . .	146
7.16	Time Comparison between CPU and GPU . . . . .	146
7.17	Results for Sorenson Distance . . . . .	147
7.18	Time Comparison between CPU and GPU . . . . .	147
7.19	Results for Tanimoto Distance . . . . .	148
7.20	Time Comparison between CPU and GPU . . . . .	148
7.21	Result for Euclidean Distance . . . . .	149
7.22	Time Comparison between CPU and GPU . . . . .	149

## List of Tables

---

7.23	Result for Manhattan Distance . . . . .	150
7.24	Time Comparison between CPU and GPU . . . . .	150
7.25	Result for Kulczynski distance . . . . .	151
7.26	Time Comparison between CPU and GPU . . . . .	151
7.27	Result for Chebyshev Distance . . . . .	152
7.28	Time Comparison between CPU and GPU . . . . .	152
7.29	Result for Cosine Similarity . . . . .	153
7.30	Time Comparison between CPU and GPU . . . . .	153
7.31	Result for Soergel Distance . . . . .	154
7.32	Time Comparison between CPU and GPU . . . . .	154
7.33	Results for Sorenson Distance . . . . .	155
7.34	Time Comparison between CPU and GPU . . . . .	155
7.35	Results for Tanimoto Distance . . . . .	156
7.36	Time Comparison between CPU and GPU . . . . .	156
7.37	Result for Euclidean Distance . . . . .	157
7.38	Time Comparison between CPU and GPU . . . . .	157
7.39	Result for Manhattan Distance . . . . .	158
7.40	Time Comparison between CPU and GPU . . . . .	158
7.41	Result for Kulczynski Distance . . . . .	159
7.42	Time Comparison between CPU and GPU . . . . .	159
7.43	Result for Chebyshev Distance . . . . .	160
7.44	Time Comparison between CPU and GPU . . . . .	160
7.45	Result for Cosine Similarity . . . . .	161



## List of Tables

---

7.46	Time Comparison between CPU and GPU . . . . .	161
7.47	Results for Soergel Distance . . . . .	162
7.48	Time Comparison between CPU and GPU . . . . .	162
7.49	Results for Sorenson Distance . . . . .	163
7.50	Time Comparison between CPU and GPU . . . . .	163
7.51	Results for Tanimoto Distance . . . . .	164
7.52	Time Comparison between CPU and GPU . . . . .	164
7.53	A ratio of CPU time and GPU time . . . . .	166
8.1	Comparison of Model Computational Efficiencies . . . . .	179
8.2	Dataset Statistics . . . . .	181
8.3	Performance Metrics of Various Models on Dataset 1 . . . . .	184
8.4	Performance Metrics of Various Models on Dataset 2 . . . . .	185
9.1	Snapshot of Function Call Graph (FCG) for Airpush Malware . . . . .	199
9.2	Dataset Statistics . . . . .	204
9.3	Performance of GNN Models on Dataset 1 . . . . .	206
9.4	Performance of GNN Models on Dataset 2 . . . . .	207
9.5	Method Accuracy Comparison . . . . .	207

# Glossary of Terms

API	Application Programming Interface
MaaS	Malware-as-a-Service
CUDA	Compute Unified Device Architecture
CPU	Central Processing Unit
GPU	Graphics Processing Unit
FCG	Function Call Graph
CNN	Convolutional Neural Network
GNN	Graph Neural Network
GCN	Graph Convolutional Network
GAT	Graph Attention Network
GIN	Graph Isomorphism Network
GraphSAGE	Graph Sample and Aggregate
VGG	Visual Geometry Group
ResNet	Residual Network
SWaT	Secure Water Treatment
MAC	Message Authentication Code
LAN	Local Area Network
AP	Access Point
TUANDROMD	Tezpur University Android Malware Dataset
TUMALWD	Tezpur University Windows Malware Dataset