

UniversityofSetif1-FerhatAbbas

People'sDemocraticRepublicofAlgeria

MinistryofHigherEducationandScientificResearch

UniversityofSetif1-FerhatAbbas



FacultyofSciences

THESIS

PresentedattheFacultyofSciences
DepartmentofComputerScience
FortheAwardoftheDegreeof

MASTER

Domain:ComputerScience
Specialization:Cybersecurity

TITLE

ExplainableAIforZero-DayAttackDetectionin IoTNetworks

Preparedby:

YoucefFLIFLA&AliTahaZAIM

Presentedon **Juin20,2026** ,beforeajurycomposedof:

President:	NASRIKhaled	MCBUniversityofSetif1-FerhatAbbas
Supervisor:	BOULKHIOUTYoussef	MCBUniversityofSetif1-FerhatAbbas
Examines:	GOUDJILLakhdar	MCAUniversityofSetif1-FerhatAbbas

2025/2026

Acknowledgment

First and foremost, praise be to God for granting us the mental and physical abilities that enabled us to complete this modest thesis, and we beseech Him to crown our work with success.

Youssef BOULKHIOUT
June 24, 2026

I thank Allah for granting me the strength, patience, and perseverance to complete this work. I am also proud of the dedication and effort invested throughout this journey. Finally, I express my sincere gratitude to my family and friends for their continuous support, encouragement, and belief in me. Alhamdulillah for everything.

Ali Taha ZAIM
June 24, 2026

I would like to express my sincere gratitude to everyone who supported me throughout the completion of this work.

I would also like to thank my colleagues and friends for their encouragement, valuable discussions, and continuous motivation during the development of this project.

Finally, I am deeply grateful to my family for their unwavering support, patience, understanding, and encouragement throughout my studies. Their confidence in me has been a constant source of strength and inspiration.

To all those who contributed, directly or indirectly, to the completion of this work, I extend my heartfelt thanks.

Youcef FLIFLA
June 24, 2026

Contents

Acknowledgment . . . . . **i**

Contents . . . . . **ii**

List of Figures . . . . . **viii**

List of Tables . . . . . **x**

Introduction **1**

I Intrusion Detection Frameworktem **3**

I.1 Introduction 3

I.2 Intrusion Detection Frameworktems (IDS) 3

I.2.1 Definition of IDS 3

I.2.2 Types of IDS 4

I.2.3 Architecture of IDS 6

I.2.4 AI-Based IDS 7

I.2.4.1 Machine Learning Methods 7

I.2.4.2 Deep Learning Methods 8

I.2.5 IDS Challenges 10

I.3 Zero-Day Attacks 10

I.3.1 Definition of Zero-Day Attack 10

I.3.2 Characteristics of Zero-Day Attacks 11

I.3.3 Lifecycle of Zero-Day Attacks 12

I.3.4 Types of Zero-Day Attacks 12

I.3.5 Detection Methods 13

I.4 Explainable Artificial Intelligence (XAI) 14

I.4.1 Definition of XAI 14

I.4.2 Importance of XAI in IDS 14

I.4.3 Types of Explainability 15

I.4.4 Common XAI Methods 15

I.4.5 XAI in Cybersecurity and IDS 16

I.4.6 Challenges of XAI 17

I.5 Conclusion 17

II State of the Art **19**

II.1 Introduction 19

II.2 Zero-Day Attack Detection 19
II.2.1 Attention-Based Fusion Models 20
II.2.2 Multilayer Perceptron Combined with Explainable AI 21
II.2.3 Autoencoder-Based Zero-Day Detection 21
II.2.4 Adaptive Deep Learning with Open Set Recognition 22
II.2.5 Zero-Shot Learning for Zero-Day Detection 23
II.2.6 Unsupervised Anomaly Detection for Zero-Day Attacks 24
II.2.7 Out-of-Distribution Detection Baselines 25
II.2.8 Interoperability and Comparative Frameworks 26
II.3 Intrusion Detection Frameworktems and IoT Security 27
II.3.1 Machine Learning and Deep Learning for IoT IDS 27
II.3.2 Fog Computing-Based Detection Architectures 29
II.3.3 Foundational Surveys on Anomaly-Based IDS 29
II.4 Explainable Artificial Intelligence in Cybersecurity 31
II.4.1 Explainable AI: Foundational Concepts and Taxonomy 32
II.4.2 SHAP and LIME Applied to Network Intrusion Detection 32
II.4.3 Adversarial Approaches to XAI in IDS 34
II.4.4 XAI-Based DDoS Detection in IoT Networks 34
II.4.5 Transformer-Based Models with Integrated Explainability 35
II.5 Loss Functions and Training Strategies for Imbalanced Classification 36
II.5.1 Focal Loss for Class Imbalance 36
II.5.2 Supervised Contrastive Learning 37
II.5.3 Optimization and Learning Rate Scheduling 38
II.6 Comparative Synthesis Across All Works 39
II.7 Research Gap 41
II.7.1 Limitation of Current Zero-Day Detection Approaches 41
II.7.2 Limitations of Existing IDS in IoT Environments 42
II.7.3 Limitations of Explainable AI in Cybersecurity 42
II.7.4 Identified Research Gap 43
II.8 Conclusion 43

III Methodology **45**
III.1 Introduction 45
III.2 Edge-IIoTset Dataset 46
III.3 Overview of the Proposed Framework A: 47
III.4 Dataset and Experimental Configuration 48
III.4.1 Global Hyperparameters 48

III.5 Data Preprocessing 49
III.5.1 Feature Selection and Column Removal 49
III.5.2 Missing Value Imputation 49
III.5.3 Categorical Encoding 50
III.5.4 Feature Standardisation 50
III.5.5 Label Encoding 50
III.6 Zero-Day Holdout Strategy 50
III.7 Stage 1: Autoencoder Anomaly Gate 51
III.7.1 Architecture 51
III.7.2 Training Procedure 52
III.7.3 Threshold Calibration. 53
III.8 Stage 2: 1D-CNN Classifier with Feature Attention 53
III.8.1 Feature Attention Module 53
III.8.2 1D-CNN Architecture 54
III.8.3 Global Average Pooling and Embedding Projection 54
III.9 Composite Loss Function and Training Strategy 56
III.9.1 Cross-Entropy Loss 56
III.9.2 Online Hard Triplet Loss 56
III.9.3 Training Schedule 57
III.10Stage 3: Metric-Learning Zero-Day Detection 57
III.10.1 Centroid Construction 57
III.10.2 Embedding Space Validation 58
III.10.3 Embedding Visualisation 59

III.10.4 The Fused Decision Function 59
III.11Zero-Day Explainability 60
III.11.1 Distance Profile Plot 60
III.11.2 Radar Chart 61
III.11.3 Intra-Class Distance Distributions. 61
III.11.4 Feature Deviation Analysis 61
III.12SHAP-Based Explainability for Known Attack Classes 61
III.12.1 KernelSHAP Configuration 62
III.12.2 SHAP Visualisations 62
III.13Evaluation Protocol 63
III.13.1 Stage 1 — Binary Anomaly Detection 63
III.13.2 Stage 2 — Known Attack Multiclass Classification 63
III.13.3 Stage 3 — Zero-Day Detection 63
III.13.4 ROC Curve Analysis 64
III.14Reproducibility and Artefact Management 64
III.15Proposed framework: Cascaded IDS 64
III.16Preprocessing 65
III.16.1 Feature Dropping and High-Cardinality Handling 65
III.16.2 Train/Test Split and Normalisation 66
III.17Neural Network Architecture: DeepResidualIDS 66
III.17.1 Residual Block 66
III.17.2 Full Network Architecture 67
III.18Stage 1 — Binary Classification 68
III.18.1 Objective and Training 68
III.18.2 Training Procedure 69
III.19Stage 2 — Multiclass Attack Classification 69
III.19.1 Focal Loss 69
III.19.2 Supervised Contrastive Loss 70
III.19.3 Combined Training Objective 70
III.19.4 Balanced Mini-Batch Sampling 71

III.20Zero-Day Detection: CentroidDetector 71
III.20.1 Motivation and Design 71
III.20.2 Mahalanobis Distance 72
III.20.3 Adaptive Per-Class Thresholds 72
III.20.4 Fused Anomaly Score 73
III.21Cascade Inference Pipeline 74
III.21.1 End-to-End Flow 74
III.21.2 New-to-Old Index Mapping 75
III.22Explainable Artificial Intelligence (XAI) 75
III.22.1 Overview of the XAI Module 75
III.22.2 SHAP DeepExplainer 75
III.22.2.1 Stage 1 SHAP Analysis 76
III.22.2.2 Stage 2 SHAP Analysis 76
III.22.3 Integrated Gradients on Mahalanobis Distance 77
III.22.3.1 Interpretation of IG Attributions 77
III.23Training Infrastructure and Hyperparameter Configuration 77
III.23.1 Hyperparameter Summary 77
III.23.2 Reproducibility and Device Management 78
III.24Conclusion 78
IV Results & Comparative Analysis **81**
IV.1 Introduction 81
IV.2 framework A— Experimental Results 82
IV.2.1 Stage 1 — Autoencoder Anomaly Gate 82
IV.2.2 Stage 2 — 1D-CNN Classifier 83

IV.2.3 Stage 3 — Zero-Day Detection 84
IV.2.4 Embedding Space Analysis 85
IV.2.5 SHAP Feature Importance 87
IV.2.6 Zero-Day Explainability 89
IV.3 framework B— Experimental Results 91

IV.3.1 Stage 1 — Binary Classification 92
IV.3.2 Stage 2 — Multiclass Attack Classification 92
IV.3.3 Stage 3 — Zero-Day Detection 94
IV.3.4 SHAP DeepExplainer — Stage 1 95
IV.3.5 SHAP DeepExplainer — Stage 2 97
IV.3.6 Integrated Gradients — Stage 3 97
IV.4 Architectural Overview 99
IV.4.1 frameworkA— 1D-CNN with FeatureAttention andAutoencoder Gate 99
IV.4.2 framework B— Deep Residual MLP with SupCon and Mahalanobis Fusion 100
IV.4.3 Key Architectural Differences 101
IV.5 Dataset and Experimental Setup 101
IV.6 Stage 1 — Binary Classification (Normal vs. Attack) 102
IV.6.1 Gate Mechanisms 102
IV.6.2 Stage 1 Evaluation Results 103
IV.7 Multiclass Classification of Known Attack Types 104
IV.7.1 Honest Multiclass Results (Report B) 104
IV.8 Zero-Day Detection 106
IV.8.1 Overall Zero-Day Detection Metrics 106
IV.8.2 Per-Class Zero-Day Detection Recall 107
IV.9 Known-Attack False Positive Rate Under the Zero-Day Filter 109
IV.10Embedding Quality 109
IV.11 Explainability Analyses 110
IV.12Soft-Label Assignments for Zero-Day Classes 111
IV.13Limitations of Both Frameworktems 112
IV.13.1 Shared Limitations. 112
IV.13.2 framework ASpecific Limitations 113
IV.13.3 framework BSpecific Limitations 113
IV.14Identified Gaps and Future Research Directions 114
IV.14.1 Online Adaptation and Concept Drift 114

IV.14.2 Temporal and Flow-Level Modelling 114
IV.14.3 Improving Fingerprinting and XSS Detection 114
IV.14.4 Unified XAI for Zero-Day Detection 115
IV.14.5 Adversarial Robustness 115
IV.14.6 Multi-Dataset Generalisation 115
IV.15 Summary Comparison 116
IV.16 Conclusion 117
Conclusion **118**
Bibliography **125**
Abstract **127**

ملخص:

تُقدّم هذه الأطروحة بنيتَيْن لأنظمة كشف التسلل (IDS) قائمتَيْن على التعلم العميق، وذلك بهدف اكتشاف هجمات اليوم الصفري على مجموعة البياناتegdE -IIoTset التي تحتوي على 2.2 مليون عينة، وتشمل 51 فئة هجومية، منها 6 فئات محجوزة لتقييم هجمات اليوم الصفري. النظام الأول هو نظام كشف تسللهرمي متسلسل (Cascaded Hierarchical IDS)، يعتمد على شبكة MLP ذات وصلات متبقية مع دالتَيْ خسارة تباينية و بؤر ية، ومسافة ماهالانو بيس فيفضاء تضمين ذي 46 بعُدًا، إضافة ءإلى دمج الدرجات. يحقق هذا النظام نسبة استدعاء لليوم الصفري بلغت 68.46%، ومعدل إنذار كاذب 31.00%، وقيمة

F1=0.8654، ودقة 86.86% على الهجمات المعروفة. أما النظام الثاني فهو نظام هجين مبني على CNN مع وحدة انتباه وخسارة ثلاثية (،)TelpirTssoL يستخدم مسافات إقليدية إلى مراكز الكتل. يصل هذا النظام إلى نسبة استدعاء 18.20%، ومعدل إنذار كاذب 11.90%، وقيمة F1=0.8418، ودقة 94.90%على الهجمات المعروفة. يدمج كلا النظامَيْن قابلية ءالتفسير متعددة المراحل باستخدام SHAP والتدرجات المتكاملة (Integrated Gradients). يكشف تحليلSHAP المشترك عن نفس العشر خصائص الأكثر أهمية على المستوى العالمي، في حين تُظْهر التدرجات المتكاملة بصمات شاذة فر يدة لكل فئة من فئات اليومالصفري.

مفاتيح: كشف هجمات اليوم الصفري، نظام كشف التسلل، إنترنت الأشياء، التعلم العميق، التعرف على المجموعات المفتوحة، الذكاء الاصطناعي القابلللتفسير.

Résumé :

Cette thèse présente deux architectures IDS à apprentissage profond pour la détection d'attaques zero-day sur le jeu de données Edge-IIoTset (2,2 millions d'échantillons, 15 classes d'attaque, dont 6 réservées pour l'évaluation zero-day). La première, un IDS hiérarchique en cascade, utilise un MLP résiduel avec pertes contrastive et focal, une distance de Mahalanobis dans un espace de plongement 64-D, et une fusion de scores. Elle atteint 86,64% de rappel zero-day à 13,00% de FPR, F1=0,8654 et 86,86% de précision sur les attaques connues. La seconde, un IDS hybride à CNN avec module d'attention et perte triplet, utilise des distances euclidiennes aux centroïdes. Elle obtient 81,02% de rappel zero-day à 11,09% de FPR, F1=0,8418 et 90,94% de précision sur les attaques connues. Les deux Frameworktèmes intègrent une explicabilité multi-étape (SHAP, Integrated Gradients). L'analyse SHAP croisée révèle les dix mêmes caractéristiques globalement importantes ; Integrated Gradients montre des signatures d'anomalie uniques par classe zero-day..

Mots-clés : *Détection d'attaques zero-day, Frameworktème de détection d'intrusion, Internet des objets, Apprentissage profond, Reconnaissance en ensemble ouvert, Intelligence artificielle explicable.*

Abstract:

This thesis presents two deep learning IDS architectures for zero-day attack detection on the Edge-IIoTset dataset (2.2M samples, 15 attack classes, 6 held out for zero-day evaluation). The first, a Cascaded Hierarchical IDS, uses a residual MLP with contrastive and focal loss, Mahalanobis distance in a 64-D embedding, and score fusion. It achieves 86.64% zero-day recall at 13.00% FPR, F1=0.8654, and 86.86% known-attack accuracy. The second, a Hybrid CNN-Based IDS with feature attention and triplet loss, uses Euclidean centroid distances. It achieves 81.02% zero-day recall at 11.09% FPR, F1=0.8418, and 90.94% known-attack accuracy. Both Frameworktems include multi-stage explainability (SHAP, Integrated Gradients). Cross-Frameworktem SHAP reveals the same ten globally important features; Integrated Gradients shows unique anomaly signatures per zero-day class.

Keywords: *Zero-Day Attack Detection, Intrusion Detection Frameworktem, Internet of Things, Deep Learning, Open-Set Recognition, Explainable Artificial Intelligence.*