

**SSH ATTACKS CLASSIFICATION & CLUSTERING USING
HONEYPOT COWRIE LOGS**

By

Alif Minda Septiansyah
200001089

MASTER'S DEGREE
in

MASTER OF INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

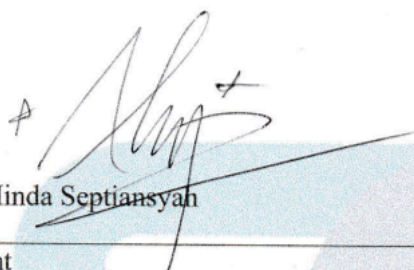


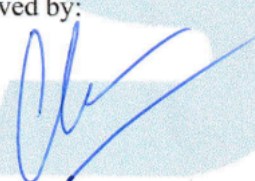
SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

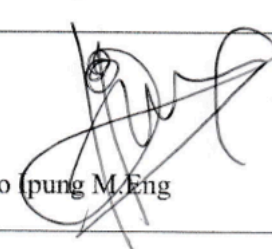
August 2026
Revision after the Thesis Defense on 04 August 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.


Alif Minda Septiansyah
Student
15 August 2026

Approved by:

Dr. Ir. Charles Lim BSc., M.Sc
Thesis Advisor
15 August 2026


Dr. Ir. Heru Purnomo (pung M.Eng
Thesis Co-Advisor
15 August 2026

Dr. Maulahikmah Galinium S.Kom M.Sc
Dean
15 August 2026

ABSTRACT

SSH ATTACKS CLASSIFICATION & CLUSTERING USING HONEYPOT CORWIE LOGS

By

Alif Minda Septiansyah

Dr. Ir. Charles Lim BSc., M.Sc, Advisor

Dr. Ir. Heru Purnomo Ipung M.Eng, Co-Advisor

SWISS GERMAN UNIVERSITY

This research aims to develop an automated framework for cyber threat intelligence by transforming unstructured SSH and Telnet honeypot logs into structured, actionable intelligence. Traditional signature-based defenses and regex rules struggle to dynamically capture complex, obfuscated adversary behaviors. To overcome this, this research proposes a semi-supervised methodology integrating Natural Language Processing (NLP), and machine learning applied to Cowrie honeypot telemetry. The framework utilizes SecureBERT 2.0 to generate semantic embeddings of attacker command sequences and applies Incremental Principal Component Analysis (IPCA) for dimensionality reduction. Subsequently, for behavioral campaign clustering HDBSCAN applied, and command sequences are automatically mapped to the MITRE ATT&CK framework using cosine similarity. The findings demonstrate that the proposed framework successfully processed over 100,000 real-world sessions, accurately identifying distinct attack campaigns—such as brute force, active scanning, and persistence techniques—without human intervention. This Machine Learning-driven approach significantly outperformed traditional methods in extracting fine-grained behavioral patterns. In conclusion, this research provides a scalable, intelligent pipeline that reduces reliance on manual detection rules, enhancing operational cyber threat intelligence and enabling robust, data-driven behavioral attribution of evolving cyber threats.

Keywords: Honeypot Telemetry, Natural Language Processing, Unsupervised Clustering, MITRE ATT&CK Framework, Cyber Threat Intelligence.



DEDICATION

This Thesis dedicated to the future of cybersecurity and artificial intelligence in Indonesia, with the hope that this research contributes to a safer, more resilient digital landscape for the nation.



ACKNOWLEDGEMENTS

Alhamdulillah, all praise and gratitude belong to Allah SWT for His endless blessings, grace, and strength, which enabled me to complete this thesis. This research would not have been possible without the guidance of my advisor, Dr. Ir. Charles Lim BSc., M.Sc, and my co-advisor, Dr. Ir. Heru Purnomo Ipung M.Eng, for their exceptional guidance, patience, and continuous support throughout this research. Their expertise and invaluable feedback have been instrumental in shaping this thesis and guiding me through the complexities of cyber threat attribution and machine learning.

I extend my sincere appreciation to all my lecturers at MIT. The rigorous academic environment and the profound knowledge they have shared with me laid the critical foundation necessary to undertake and complete this research.

I am also highly grateful to my colleagues and peers at the SGU Cybersecurity Lab. The collaborative environment, insightful discussions, and shared passion for security research have been a great source of motivation and technical inspiration during my studies.

To my friends in class (MIT Batch 36) and my hangout friends: thank you for the friendship, cooperation, the usual late-night assignments, and the much-needed breaks. Your encouragement and friendship provided the perfect balance to the demands of academic research, keeping me grounded and motivated.

Finally, words cannot express my profound gratitude to my family. A very special and heartfelt thank you goes to my mother, whose unconditional love, endless prayers, and countless sacrifices have been my greatest source of strength. This milestone would not have been possible without her unwavering belief in me.

Thank you all for being a part of this journey.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR	2
ABSTRACT	3
DEDICATION	5
ACKNOWLEDGEMENTS	6
TABLE OF CONTENTS	7
LIST OF FIGURES.....	12
LIST OF TABLES	13
CHAPTER 1 - INTRODUCTION	15
1.1. Background	15
1.2. Research Problem.....	17
1.3. Research Objectives	19
1.4. Research Questions	19
1.5. Hypothesis.....	20
1.6. Research Scope and Limitation.....	20
1.7. Significance of Study	22
1.8. Thesis Structure.....	22
CHAPTER 2 - LITERATURE REVIEW	24
2.1. SSH Services for Digital Business	24
2.1.1. Digital Business Ecosystem	24
2.1.2. Cyber Threat Landscape of SSH Service	24
2.2. SSH Services and Attack Vectors	26
2.2.1. SSH Architecture & Security Evolution	26

2.2.2. Attacker Motivation & Methods	26
2.2.3. Characteristics of Traffic Log: Interactive vs. Non-Interactive	27
2.2.4. Mechanism Defense Evolution	28
2.3. Deception Technology and Honeypot.....	30
2.3.1. Deceptive Defense Mechanism	30
2.3.2. Medium-Interaction Honeypots and Fake Shells	30
2.3.3. Honeypot Cowrie and Telemetry Architecture	31
2.3.4. Scope Boundaries and Visibility Limitations.....	32
2.4. Data Mining.....	33
2.4.1. Historical Regex Ruled-Based to Semantic AI	33
2.4.2. Alternative Regex baseline (STIX 2.0 MITRE ATT&CK)	33
2.4.3. Command Deobfuscation and Normalization Techniques.....	35
2.4.4. Feature Extraction & Contextual Enrichment.....	36
2.5. Fundamentals of Machine Learning in Cybersecurity	36
2.5.1. Core Operational Paradigm: Supervised and Unsupervised Learning.....	36
2.5.1.1. Supervised Learning.....	37
2.5.1.2. Unsupervised Learning	39
2.5.1.3. Semi-Supervised.....	40
2.5.1.4. Neural Language Processing	41
2.5.2. The Role of Machine Learning in Cybersecurity	42
2.6. Applied Machine Learning Models and Analytical Learning Algorithms	44
2.6.1. Cybersecurity NLP and Semantic Embedding (SecureBERT)	44
2.6.2. Feature Engineering and Multi-Modal Feature Fusion Mechanics.....	48
2.6.3. Dimensionality Reduction (Incremental PCA)	50
2.6.4. Unsupervised Clustering (HDBSCAN)	51
2.6.5. Cluster Validation Metrics	54

2.6.6. Supervised Classification (XGBoost)	60
2.6.7. Evaluation and Metrics for Supervised threat models	61
2.7. Cyber Threat Intelligence	66
2.7.1. Tactical vs Operational Intelligence	66
2.7.2. MITRE ATT&CK TTP Mapping	68
2.8. Related Works	69
2.8.1. Related Works and Research Gaps	69
CHAPTER 3 - RESEARCH METHODS	74
3.1. Research Framework Overview	74
3.2. Phase 1: Telemetry Data Ingestion and Preprocessing	78
3.2.1. Raw Cowrie JSON Log Ingestion and Sessionization	78
3.2.2. Multi-Stage Command Deobfuscation Engine	79
3.2.3. Session Deduplication and Normalization	80
3.3. Phase 2: Feature Engineering & Multi-Modal Transformation	81
3.3.1. Tabular & Statistical Feature Extraction (27-D Feature Space)	81
3.3.2. Semantic Representation via SecureBERT 2.0 (768-D Embeddings)	83
3.3.3. Dimensionality Reduction via Mini-Batch Incremental PCA (768-D to 50-D)	84
3.3.4. Multi-Modal Feature Fusion Matrix Construction (77-D Space)	85
3.3.5. Deterministic STIX 2.0 Baseline (Pseudo-Ground Truth)	86
3.4. Phase 3: Machine Learning-Driven Analytics Engine	88
3.4.1. Supervised Threat Classification via XGBoost	88
3.4.2. Unsupervised Campaign Clustering via HDBSCAN	89
3.5. Phase 4: Cyber Threat Intelligence Mapping & TTP Attribution	91
3.5.1. Automated STIX 2.0 Behavioral Mapping & Attribution Framework	91
3.5.2. Cosine Semantic Similarity Mapping Mechanics	92
3.6. Evaluation & Validation Methodology	94

3.6.1. Supervised Classification Evaluation Framework	94
3.6.1.1. Mathematical Formulation of Discrete Multi-Class Metrics	95
3.6.1.2. Class-Balanced Aggregation & Probability Calibration	96
3.6.1.3. Dataset Partitioning	96
3.6.2. Unsupervised Clustering Validation Framework	97
3.6.2.1. Mathematical Formulation of DBCV	97
3.6.2.2. External Chance-Adjust Benchmarking (ARI & AMI Formulation).....	98
CHAPTER 4 - RESULTS AND DISCUSSIONS.....	101
4.1. System Architecture Overview	101
4.1.1. Flowchart of Research Framework	102
4.2. Experimental Setup & Environmental Baseline.....	105
4.3. Preliminary Parameter Selection & Theoretical Grounding	107
4.3.1. Semantic Representation & IPCA Parameter Bounds	107
4.3.2. Deterministic Taxonomy Similarity Threshold Bounds	108
4.3.3. Density-Based Hyperparameter Configuration (HDBSCAN)	108
4.3.4. Multi-Modal Supervise Classification Parameter (XGBoost)	109
4.3.5. Operational Parameter Reference Matrix.....	110
4.4. Supervised Threat Classification Results (XGBoost)	111
4.4.1. Resolving Severe Class Imbalance and Macro-Weighted	112
4.4.2. Confusion Matrix Analysis	113
4.4.3. Feature Importance and Decision Criteria Analysis	115
4.5. Unsupervised Campaign Discovery and Clustering Results (HDBSCAN).....	117
4.5.1. External Taxonomical Alignment (ARI & AMI).....	118
4.5.2. Discovered Attack Campaign & Behavioral Profiling.....	119
4.5.3. Outlier Isolation & Scanner Noise (Cluster -1).....	122
4.6. Multi-Signal TTP Synthesis and Operational Threat Intelligence	123

4.6.1. Empirical Shift Analysis: Synthesis TTP Distributions	124
4.6.2. Serialization to Interoperable STIX 2.0 CTI Objects.....	127
4.7. Key Contribution and Discussion	129
4.7.1. Scientific and Technical Contribution.....	129
4.7.2. Empirical Insights and Operational Limitation	131
4.7.2.1. Empirical Realities of Threat Telemetry	131
4.7.2.2. The Imperative of Integrating Human-in-the-Loop (HITL) Workflows	132
4.7.2.3. Empirical Parameter Assessment to Advanced Optimization.....	132
4.8. Alignment with Research Question and Hypotheses Verification.....	133
4.9. Chapter Summary.....	136
CHAPTER 5 - CONCLUSIONS AND RECCOMENDATIONS	141
5.1. Conclusion.....	141
5.2. Research Limitation	142
5.2.1. Telemetry Scope.....	142
5.2.2. Extreme Long-Tail Class Imbalance and Macro-Weighted Divergence	143
5.2.3. Unannotated Zero-Day Clusters and Taxonomy Rigidity.....	143
5.2.4. Static Batch Processing and Temporal Concept Drift.....	144
5.3. Recommendations and Future Work.....	145
GLOSSARY	148
REFERENCES.....	151
CURRICULUM VITAE	157

LIST OF FIGURES

Figures	Page
Figure 1.1. Fishbone Diagram.....	19
Figure 2.1. Attack vectors of all Vulnerabilities (ENISA, 2025).....	25
Figure 2.2. Most identified initial Vector (ENISA, 2025)	25
Figure 2.3. The Pyramid of Pain	67
Figure 3.1. Research Framework	74
Figure 4.1. System Architecture.....	101
Figure 4.2. Flowchart of Research Framework.....	102
Figure 4.3. Multi-Class Confusion Matrix	114
Figure 4. 4. Sample of Serialization output 1	128
Figure 4. 5. Sample of Serialization output 2.....	128

LIST OF TABLES

Table	Page
Table 2.1. Taxonomy of SSH Defense Mechanisms	28
Table 2.2. Cowrie JSON Telemetry & Behavioral Mapping	32
Table 2. 3. Comparative verified TTP vs Regex	34
Table 2.4. Comparative of NLP Generations	45
Table 2.5. Token counts for code and text across categorized pretraining dataset	46
Table 2.6. Document embedding Performance (Bi-Encoder).....	47
Table 2.7. Methodological Comparison of Unsupervised Clustering Algorithms.....	53
Table 2.8. Classification models report.....	61
Table 2.9. Comparison of Frameworks	70
Table 3.1. Comprehensive Research Alignment and Hypothesis testing Matrix.....	76
Table 3.2. Sample of Feature Extraction.....	82
Table 3.3. Sample of Semantic Embedding	83
Table 3.4. Sample of dimension reduction.....	84
Table 3.5. Sample of Feature Fusion matrix construction	86
Table 3.6. Sample of STIX 2.0 Baseline.....	87
Table 4.1. JSON Logs after Parsing	105
Table 4.2. Experimental & Environmental Specification	106
Table 4.3. Operational Parameter and Theoretical Ground	110
Table 4.4. Per-Class Supervised Classification Report.....	112
Table 4.5. Top XGBoost Feature Importances	116
Table 4.6. Discovered HDBSCAN Campaign Clusters	119

Table 4.7. Synthesis Context TTP Distribution (Phase 3-Baseline)	124
Table 4.8. Final Synthesis Context TTP Distribution (Phase 4c)	125
Table 4. 9. Thesis Hypothesis and Research Alignment.....	134
Table 5. 1. Strategic Framework enhancements & Future Work.....	146

