Chapter 4

Dataset Description

4.1 Introduction

In the proposed framework, when identifying zero-day attack paths two information sources are utilized, namely (i) Path Dataset-D1, which comprises system object dependencies and network-wide provenance to extract potential attack paths and (ii) Attack Dataset-D2, which includes labeled traffic examples of benign and attack instances, including zero-day exploit traces. In applying the both, it ensures that the framework is considered on both the path identification and the attack behavior classification aspects of the research problem.

4.2 Dataset 1: Path Dataset

4.2.1 Description

Path Dataset is formed by simulating the dependencies of system objects on a network within different hosts. Every record has a record of an object (process, file, socket, registry entry) and the dependency relations that the instance has towards other objects. With the help of these relations, a System Object Dependency Graph (SODG) is created, in which the edges represent read/write/invoke dependencies.

- Total Features Generated: 18
- Features to be used: 12 (some of the were selected because of the relevance to propagation analysis)
- Rationale behind choosing: The 12 features are connected to the identification of suspicious initiation of intrusion propagation that is capable of involving the 0-day exploits.

Table 4.1 Feature Table (Path Dataset)

Feature No.	Feature Name	Description	Contribution to Zero-Day Attack Identification
1	Object ID	Unique identifier of system object	Provides traceability across system
2	Object Type	Process/File/Socket/Registry	Differentiates critical object categories
3	Operation Type	Read/Write/Execute/Invoke	Detects abnormal execution chains

Feature No.	Feature Name	Description	Contribution to Zero-Day Attack Identification
4	Source Object	Parent object generating event	Identifies origin of suspicious dependencies
5	Destination Object	Target object affected	Tracks attack propagation targets
6	Timestamps	Time of operation	Helps in temporal correlation of events
7	Dependency Edge Weight	Probability of influence	Captures infection likelihood across objects
8	Event Frequency	No. of times interaction occurs	Detects anomalous repetition (attack loops)
9	Data Volume	Bytes transferred	Identifies unusual large transfers
10	Execution Path Depth	Depth of dependency graph	Helps detect multi-hop hidden propagation
11	Anomaly Flag	Derived via thresholding	Indicates suspicious propagation
12	System Call Type	Read, fork, exec, open, etc.	Critical in recognizing unseen exploit triggers

4.2.2 Path Dataset Simulation Details

Simulation Tool: C++/ Python system call tracer to build the graph and Yed Editor. Another dataset D1 called Path Dataset is discussed in the following chapter 5.

Table 4.2 Simulation Parameter Table

Parameter	Value
No. of Hosts Simulated	10
OS Environment	Linux (Ubuntu 20.04)
Monitoring Duration	48 hours
No. of Object Instances Captured	~50,000
Avg. Graph Size	8,000 nodes / 25,000 edges
Dependency Types	Read, Write, Execute, Invoke
Noise Filtering	Events < 5 occurrences discarded
Graph Representation	Directed labeled graph

4.3 Dataset 2: Attack Dataset

4.3.1 Description

Dataset 2 (D2) is an openly available benchmark dataset and is called the Celosia Zero-Day Attack Dataset which is composed of labelled network traffic flows of benign and malicious network traffic. This information has been observed to be applicable in the supervised learning and anomaly detection research since it contains the zero-day attacks of abnormal traffic behaviours and lacks the exploit patterns signatures.

D2 is the standard data set that is utilized to ensure that the results can be reproducible and can be appropriately compared. Usage of this dataset in details, preprocessing, feature engineering, model training, and performance evaluation is demonstrated in Chapter 6, Chapter 7 and Chapter 8 further.

a. Total Features Generated: 25
b. Count of features Selected to work with: 15 (selected feature after feature importance analysis based on Boruta and Chi-square test). The proposed research structure entails the process of feature selection which is undertaken during the preprocessing phase that is depicted in Chapter 6. Depending on the significance of features, the Boruta and Chi-square statistical tests are conducted and only the most significant features of zero-day attacks are kept.
c. Rationale behind the Selections: The features are used to detect/classify attacks, such as zero-day signature within anomalous patterns.

Table 4.3 Feature Table (Attack Dataset)

Feature No.	Feature Name	Description	Zero-Day Contribution
1	Src IP	Source address	Identifies abnormal origins
2	Dst IP	Destination address	Reveals targeted systems
3	Src Port	Originating port	Detects suspicious services
4	Dst Port	Target port	Key for exploitation attempts
5	Protocol	TCP/UDP/ICMP	Identifies control channels
6	Packet Size	Avg. packet length	Anomalous in exploits

Feature No.	Feature Name	Description	Zero-Day Contribution
7	Flow Duration	Start–end time	Short bursts indicate scans
8	Flow Count	No. of packets	Overflows suggest DoS/Exploit floods
9	Payload Bytes	Data carried	Encoded exploit payload detection
10	Header Flags	SYN/ACK/RST/FIN	Abnormal flag combos → attacks
11	Connection State	Normal/Aborted	Failed attempts can be attack indicators
12	Session Rate	Packets/sec	Detects brute-force attempts
13	Entropy Score	Payload randomness	High = encrypted zero-day payloads
14	Label	Benign/Attack	Ground-truth for training
15	Attack Type	Exploit/DoS/Scan/Zero-day	Used for multi-class classification

The most significant are Zero-Day Characteristics, Payload Bytes, Entropy Score, Header Flags and Dst Port because the zero-day exploits are usually correlated with crafted payloads, obfuscated data and out-of-character protocol/port behavior which is not represented in signature-based data sets.

4.3.2 Dataset Generation Process

To strictly test the recommended multi-phase, zero-day attack detection framework, two heterogeneous data sets are used to test the framework, including, Dataset D1, which is a simulated cloud-based dataset, and Dataset D2, which is a publicly available benchmark dataset. This is due to the fact that the difference between the simulated and the benchmark data not only offers a controlled attack-path analysis but also offers the generalization of the real world.

Dataset D1: Simulated Zero-Day Attack Path Dataset (PATH)

Dataset D1 is a simulation dataset that is run on a cloud simulation system which is utilized to simulate the behaviour of realistic cloud infrastructure. The simulation will be emulating the existence of numerous virtual machines, cloud services, user workload, and adversarial activities to determine the paths of attacks in a cloud network. Any benign and malicious activity is modelled with special emphasis on the zero-day attack paths, where the exploit behavior is not familiar to the signature-based systems.

The simulated environment generates structured records of events and logs of the traffic flows that represent the interaction between cloud nodes, services, and attackers. The specified records are run through with the purpose of extracting attack-path-oriented features and standardizing and transforming into ML and DL frameworks. Phase 1 and Phase 2 of the proposed framework are largely based on Dataset D1 to determine the attack path and formulate strategic predictions in a zero-day attack.

Dataset D2: Benchmark Zero-Day Attack Dataset (Celosia)

One of the publicly available datasets, Celosia Zero-Day Attack Dataset (dataset D2), was created based on the traffic of the IoT and cloud support networks. It is composed of time-series, flow level logs, which are in CSV form, and it also contains the element of packet length, connection length, traffic volume, and statistical flow characteristics. The cases are divided into benign and malicious cases and as a result supervised learning can be performed.

In contrast to Dataset D1, Dataset D2 is not simulated within the framework of the current research, on the contrary, the dataset is borrowed as it is to give one a benchmark with the purpose of offering the reproducibility and the unfair performance comparison. This data is preprocessed (data cleaning, data normalization, feature selection and class balancing (smote in case of its usage)) and then the model is trained. Phase 2 and Phase 3 are primarily founded on D2dataset in predicting and detecting zero-day attack.

Although the origin was different, both D1 and D2 have gone through the similar preprocessing pipeline so as to attain comparability. These include feature normalization, feature selection and split the dataset in testing and training. This coherent processing applied to do just evaluation of the developed models on the simulated attack routes and real world benchmark traffic.

Train-Test Split:

i. Training: 70%
ii. Testing: 30%
iii. Balanced using SMOTE

To rectify the scenario of class imbalance, when the number of examples of a zero-day attack is lower, Synthetic Minority Over-sampling Technique (SMOTE) applies to the

training set only during the preprocessing step as described in Chapter 6 (Experimental Setup). The SMOTE can generate samples of minority classes artificially and thus not bias those models towards the majority (benign or known-attack) classes and it does not include information leakage to the test set.

Classification Achieved: Data is used to train the proposed EBPNN + graph model, which reached high detection rates and strength in the classification of zero-days.

4.4 Justification for Using Both Datasets

Dataset 1 (Path Dataset) contains structural evidence of the attack propagation paths at the vantage point of the system, whereas Dataset 2 (Attack Dataset) contains behavioral evidence of the malicious traffic or zero-day exploits at the vantage point of network. The latter, though essential to the research since under the network perspective, behavior patterns of a zero-day attack can typically be determined by using dynamic patterns (real-time) of behavior as opposed to structural patterns (credential patterns). With the inclusion of Dataset 2, the model is better trained to help monitor real-world exploit payloads and network-based actions occurring during an attack and thereby enhances the generalizability of the model. With both datasets, the model is kept up to date and the research is justified, as it shows propagation paths at path-level and behavior patterns at traffic-level - enhancing the reliability and general robustness.

The provided research structure presupposes four steps, and the suitability of Dataset D1 (PATH) and Dataset D2 (Celosia) justified in terms of the objectives and procedures in each of the steps.

Phase 1: Zero-Day Attack Path Identification

The dataset D1 (Zero-Day Attack Path Dataset - PATH) is focused on the analysis of the attack path in the cloud environment, and, thus, is extremely suitable to this step. It logs multi-stage interactions among cloud nodes, services and attackers which enable the proper modelling of path of attack propagation. In this phase, one uses the Enhanced Back Propagation Neural Network (EBPNN) model and attack modeled by graph-based to identify the origin and direction of attack paths in zero-days attack. Data set D2 cannot be used in this step, as it does not contain explicit attack path information.

The enhancements proposed in the BPNN, based on a probabilistic graph approach, are not limited to the simulated cloud environment. This technique is flexible and can be applied to other applications and target systems such as enterprise networks, IoT devices, and industrial control systems. The method remains general and effective across different environments, with trivial adaptations to data format and system topology.

Phase 2: Zero-Day Attack Prediction Using Game Theory

Phase 2 is the utilization of Dataset D1 and Dataset D2 in the prediction of a zero-day attack. The Game Theory can be used in dataset D1 to predict the strategies, in which the interactions between the attackers and the defenders are developed with regard to identified attack paths. This is supported by Dataset D2 that enables prediction of behavior based on Improved Decision Tree, Random Forest, Logistic Regression, Auto Encoder, and Modified Bi-LSTM models. The combination of the two datasets will ensure the strength of the simulated and benchmark environment.

Phase 3: Zero-Day Attack Prediction and Detection Using Deep and Transfer Learning

Phase 3 focuses on appropriate predicting and detection of the zero-day attacks using the deep and transfer learning techniques. The key factor that causes Dataset D2 (Celosia) to be used is due to the fact that it contains the labeled real-time flow-level characteristics, which can be readily adapted to deep learning model. The ResNet50 architectures based on Bi-LSTM methods are stacked with a method of ensemble learning to provide the spatial, temporal and contextual characteristics of the attacks. Dataset D1 could be considered a validation dataset but not the very significant detection dataset.

Phase 4: Comparative Analysis and Optimization of Prediction and Detection Models

The phase 4 is dedicated to the racial performance of all the preceding phases and optimization. This step compares in a systematic way the existing machine learning, deep learning and hybrid ensemble models using Dataset D1 and Dataset D2. The Optimized Levy Flight Firefly Optimization Algorithm (OLFOA) is used in optimization of hyperparameters, ensemble fusion strategies and gives a fair comparison in terms of accuracy, precision, recall, F-measure, false alarm rate and computational efficiency. The

step ensures the accuracy, generalization ability, and uniformity of the heterogeneous datasets of the proposed framework.