Chapter 3

Research Methodology

3.1 Introduction

As it was mentioned earlier, zero-day attack uses an unknown vulnerability in a software application or system. One would have to seek vulnerabilities in software applications or systems before attackers do to detect a zero-day attack. This achieved in different ways including code analysis, vulnerability scanning, penetration testing, as well as threat intelligence.

Threat intelligence is the subject of this research work. As a solution to the limitations mentioned and research gaps found, a unified approach is proposed to detect, anticipate and identify zero-day attacks by learning techniques. The proposed methodology will use four stages, which include the detection of the path of a zero-day attack through an Enhanced BPNN algorithm with Cloud Sim simulation. The next steps are to apply the methods of learning and detect the zero-day attacks, preprocessor of the information, features extraction, model training and evaluation, and results comparison.

The proposed research is relevant in that it attempts to enhance the capability of the organizations to prevent and detect zero-day attacks based on learning techniques. This research can be utilized in creating more effective and efficient tools of identification, prediction and detection of zero-days attacks based on threat intelligence that can assist organizations to enhance assets and information protection against the cybercriminals.

To assess the performance of the proposed models, the research is performed based on the simulation and real-life data. CloudSim is used to perform the simulation, and real-life data is provided by the real zero-day attacks. The outcomes of the comparative analysis have given the ideas about the effectiveness of the proposed methodology and its possibilities to be used practically in cyber security. Chapter 3 is the full methodology of zero-day attack identification, prediction, and detection based on the learning techniques, which is organized into four phases that are marked by subheadings 3.2.1 to 3.2.4. It starts with Phase 1 which utilizes the Enhanced BPNN Algorithm, which is used to determine the paths and then proceeds to Phase 2, which concentrates on the technique of Data Preprocessing and Feature Selection. Phase 3 includes the model Training and Testing techniques of zero-day attack prediction, whereas Phase 4 is a Comparative analysis of the

results to measure the model effectiveness. The chapter covers the research design that will include the methods, tools, and techniques of data collection in the implementation of the proposed methodology. Lastly, the Chapter Summary presents the significance of the methodology in increasing defense mechanisms against zero-day attacks, which will be followed up in other chapters.

The key goal is to cope with the most important problem of managing zero-day attacks on cloud infrastructures. The different stages of the research work are thoroughly structured to enhance the identification, prediction, detection process and comparative analysis of the proposed models. With the incorporation of the current methods in cyber defense systems, the proposed methods offer the following performances:

(i) Better Accuracy, Misclassification rates and better data protection.

(ii) Zero-Day Attacks Predictability, Strong, and Accurate System Communication.

(iii) Improved Zero-Day Attack Detection, Improved generalization capability, Classification Accuracy and High Detection.

(iv) Improved Recognition Rates reduced false positive and complications of detection times.

(v) Zero-Day Attacks Dynamic Cloud Environment Mitigation.

3.2 Steps Involved in the Proposed Methodology

An overall sequential process is designed to propose appropriate learning model to deal with zero-day attacks as indicated in Figure 3.1.

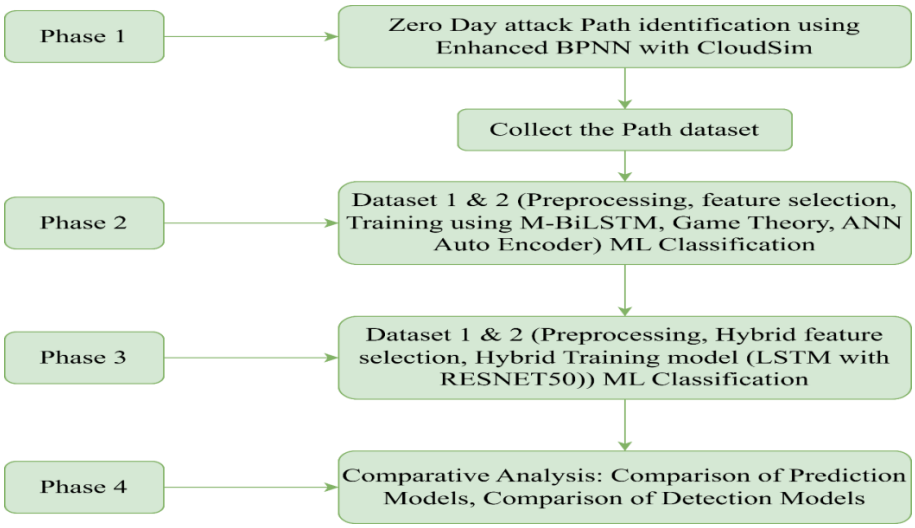

Figure 3.1 Proposed Research Methodology of Zero-Day Attack Handling

Step 1: Zero-Day Attack Path Identification using Enhanced BPNN with CloudSim

Phase 1 involves the step of zero-day attack path identification. The steps followed to determine the source of the attack path include:

i. Collected traffic data of target system.

ii. Standardize the data and prenormalize the data to get the features.

iii. Train Enhanced Back Propagation Neural Network (BPNN) algorithm on the preprocessed data.

iv. This is done with the assistance of CloudSim simulation structure to simulate the network traffic and interactions between the attacker and the target system.

v. Using CloudSim, a dataset of attack paths (Dataset D1, referred to as the Path Dataset) is created that is made out of regular traffic as well as control flows, both being zero-day as well as non-zero-day exploits.

vi. Apply the Enhanced BPNN algorithm to the simulated data and determine the course of the attack.

vii. The conventional BPNN is improved in the following ways:

(i) Adaptation of the learning rate dynamically to accelerate the convergence process and avoid the local minima.

(ii) To stabilize the training process and improve the ability to generalize from training data to future data we apply a technique called weight regularization where we prefer weights with smaller magnitudes on a model to minimize overfitting.

(iii) Probability attacks path statistics on network interaction graph, which enables the model to be sensitive to the complex attack dependencies. These extensions render it more precise and strong as far as classification is concerned in order to identify new attack patterns. Take appropriate measures to predict the attack based on the identified path.

Nonetheless the resolved attack path has zero day exploits as well as non-zero day exploits. As such, only zero day attacks have to be predicted in the identified path.

Step 2: Zero-Day Attack Prediction using Game Theory

The steps involved in predicting a zero-day attack in phase 2 are as follows:

i. Preprocessing: The retrieved information about the attack-associated data of the Dataset D1 (Path Dataset generated with the assistance of CloudSim) and Dataset D2 (Real-World Network Intrusion Dataset) is preprocessed and cleansed with the proposed enhanced Decision Tree-based preprocessing algorithm. This is to remove redundant, noisy and irrelevant features, but important attributes that are essential to attacks necessary in effective prediction of zero-day attacks.

ii. Selection of features: Selecting the most applicable features among the data set through random forest and Logistic regression algorithms is a combination method that would be used to select the features that had the most effect on the performance of the model.

iii. Training and testing: Train and test the model with some machine learning models, such as M-BiLSTM, Game Theory, and Auto Encoder algorithms.

iv. Classification: Train the model on a Stacking Ensemble Classification algorithm, which trains a combination of the predictions made by a number of machine learning algorithms on the final prediction to enhance the accuracy of the result.

Although the zero-day attacks are expected to increase the model reliability and robustness it is extremely vital to research different predictive measures that opens the path towards the effective detection of the zero-day attacks.

Step 3: Zero-Day Attack Prediction and Detection using ResNet

Prediction and Detection with ML and DL is done in Phase 3 under the following steps:

i. Preprocessing: A Decision Tree Regressor-based preprocessing strategy is used to preprocess the dataset on the attack-related information generated with Dataset D1 (Path Dataset created with the help of the CloudSim simulation framework) and Dataset D2 (Real-World Network Intrusion Dataset). The step will eliminate unnecessary and redundant attributes without compromising on discriminative features that will be required to help make efficient predictions and detection of zero-day attacks.

ii. Feature Selection: Select the best features in data set with a combination of Random Forest with Logistic Regression which selects the most important features with regard to how contribute to the performance of the model.

iii. Training and Testing: Train and test the model with various methods of learning such as Long Short-term Memory (LSTM) and Resnet 50 algorithms.

iv. Classification: Classify the model by using Stacking Ensemble Classification algorithm which involves the combination of the predictions of multiple learning algorithms to enhance the accuracy of the final prediction and detection.

Although the outcome indicates an improvement in all measures, extensive scalability and optimization challenges in place during its application in real-time detection at a very large scale.

Step 4: Comparative Analysis for Predicting Zero-Day Attacks Using Optimized Levy Flight-based Fruit Fly Optimization Algorithm

The steps that are followed in phase 4 of comparison of predicting zero-day attacks are as follows:

i. Comparative Analysis: Perform a comparative analysis of the findings of the Phase 2 and Phase 3 depending on such factors as accuracy, time of training and testing the models, and complexity of the models. Report the results in the form of graphs, charts, and tables.

ii. Identify the best method: Compare the results to be aware of which strategy is more efficient in predicting the zero-day attacks in the Optimized Levy Flight based Fruit Fly Optimization Algorithm. The above research has also demonstrated that, ensemble deep learning model combined with the Levy Flight motivated optimizing mechanisms can significantly improve the intrusion detection performance.

3.3 Research Design

The proposed will work on improved Prediction and detection accuracy of the zero-day attacks. The synthesized perspective of the proposed methodology and the methods used and the results are depicted in the figure.3.2.

Figure 3.2 Research Design

3.3.1 Contribution 1: Zero-Day Attack Path Identification using Enhanced BPNN with CloudSim

This step proposed a probabilistic graph-based enhanced neural network based on back propagation with simulation using CloudSim to find zero-day attack paths. EBPNN is based on the concept of modeling network traffic in the form of graphs and training a neural network with a learning rate parameter and momentum. The conclusion of this realism is that the EBPNN is able to identify the path of propagation of attacks with most of the paths being identified correctly. Depending on the EBPNN performance, its accuracy is far greater to identify the propagation route of the attack than the Bayesian-based methods, the numbers of which are 99.0% accuracy, 99.21% precision, 99.34% recall, and 99.54% F-measure improvement and up to 3 percent higher routing path identification in the existing literature.

3.3.2 Contribution 2: Zero-Day Attack Prediction Using Game Theory

The hybrid ML system that is proposed to assist organizations in predicting zero-day attacks incorporates Improved Decision Tree and Random Forests in relation to Logistic Regression, M-BiLSTM, Game Theory, and Auto Encoder. Another type of ensemble used to optimize a classification ensemble is a Stacking Ensemble which is used to enhance reliability through the use of numerous learners that optimize the detection of zero-day attacks. The reliability of the proposed model is also quantified in terms of using a number of evaluation criteria including accuracy, precision, recall, F-Measure, false alarm rate and consistency of the model performance on various datasets. In particular, reliability is calculated as the capability of the model to achieve high recall and low false positives, being able to achieve the same results on Dataset 1 and Dataset 2, and being able to achieve consistent results in terms of F-measure, i.e. the model acts in a balanced way. Ensemble learning is also beneficial in ensuring greater reliability in regard to reducing model variance and also countering bias on the learner level. The model gave very good results with accuracy of 95.4% (Dataset 1) and 95.0% (Dataset 2) with F-measure score of more than 89 indicating a consistent mapping of zero-day attacks behaviors across the datasets.

3.3.3 Contribution 3: Zero-Day Attack Prediction and Detection using ResNet50 with BiLSTM

This phase uses deep and transfer learning framework involving the use of ResNet50 together with LSTM, preprocessing using Decision Tree Regressor, and feature selection using RF + Logistic Regression. The last classification is done by use of a Stacking Ensemble. The model achieves an accuracy of 95.9% and precision and recall values of more than 88% (Dataset 1) and more than 91% (Dataset 2) after the evaluation. The findings prove that the DL + TL technique is excellent in real time detection. The presented model also shows superior generalization to untrained data and resistance to unknown threats.

The fact that the trained ResNet50-BiLSTM model has low inference time and computational efficiency can be used to support the real-time detection in this phase because it is established in the Experimental Set Up and Results sections. Specifically, the experimental evaluation also explains shorter prediction latency, reduced time complexity per instance, which implies that the model can support streams of network traffic within the time constraints of the real life. In addition, transfer learning involves lower training expenses, and the small implementation of the BiLSTM can speedily do sequential inference processes, which enables real-time decision making, unlike offline batch inference.

Besides, the generalization property of the model and its ability to withstand previously unknown attack patterns, which can be confirmed with the overall similarity of performance of two datasets and the similarity of inference performance, also proves that the model is applicable to real-life and dynamic cybersecurity settings.

3.3.4 Contribution 4: Comparative Analysis on Prediction of Zero-Day Attack

In the final stage, the accuracy, precision, recall, F1 score, and misclassification are taken as evaluation metrics to compare the two methods of the ML-based approach (Phase II) and the DL+TL-based one (Phase III). The OLFFOA is used to make the model more stable and eliminate imprecise predictions. According to the results, Phase III model is able to obtain up to 0.9 percent higher accuracy than Phase II model on a consistent basis as well as demonstrates greater generalization. In the end, this validates the fact that the ResNet50

convolutional network, along with LSTM and optimization strategies, is the most efficient among all that can be used to predict and identify zero-day attacks.

3.4 Thesis Order Justification

The proposed structure is based on a four-phase architecture that has a logical sequence and dependencies: Identification, Prediction, Detection, and Optimization. Phase 1 denotes the identification of potential zero-day attack paths by Enhanced BPNN and probabilistic graph, on simulated network activity of CloudSim. This is a key initial step, which reveals the weak points of the system. Phase 2 predicts the behavior of the attacker based on the identified paths and using the Modified Bi-LSTM and Game Theory, allowing the system to predict potential threat behaviors. These guesses are the input of Phase 3, where real-time detection will take place by using the Deep-Convolutional N-Zero Day Adversarial Safety Network (DC-nZDA) with ResNet and LSTM, which will be more or less a classification of threats undergoing. Lastly, Phase 4 uses the Optimized Levy Flight-based Fruit Fly Optimization Algorithm (OLFFOA) to refine the detection outputs, minimize the false positives and optimize the model capability. The phase is based on the previous one, and thus, all the steps have to be completed in order to obtain the accurate, interpretable, and deployable results.

Table 3.1 Phase Order Justification

Phase	Function	Why It Comes First
Phase 1: Identification	Identify attack paths via simulation	No predictions or detections can happen without knowing what/where to monitor
Phase 2: Prediction	Predict attacker behavior	Provides proactive intelligence for real-time classifiers
Phase 3: Detection	Detect ongoing attacks in real time	Requires context from Phase 2 to function meaningfully

3.5 Chapter Summary

This chapter gives the research methodology to determine, anticipate and identify the attacks of the zero days by means of a step-by-step approach consisting of four phases. Phase 1 is used in determining the path by using Enhanced BPNN with CloudSim, then Phase 2 is used to keep on with the work and build on top of this by using Deep learning and transfer learning with LSTM and ResNet50 that exhibited real time detection of the past

and possible zero day exploits, and finally, Phase 4, a comparative analysis is used to analyze, accuracy, efficiency and generalizability. All four phases taken together constitute a framework that does not apply to zero-day attack defense only, but to virtually any sensible cybersecurity operation.

Publications

- Swathy Akshaya, M., and Ganapathi, P. (2020). A Review of Machine Learning Methods Applied for Handling Zero-Day Attacks in the Cloud Environment. Handbook of Research on Machine and Deep Learning Applications for Cyber Security, 364-387. (Scopus)