# Keylogger Detection using Machine learning approaches

# Background and Literature Survey

**Nature of Research and Significance**

Keyloggers present a hazardous security threat, and the support of this research work is based on an IoT network which has a weak link in their security protocols. The aim of this project is to detect key logger malware through network traffic analysis using machine learning models, with which attack patterns can be built.

**Problem statement**

Most traditional methods of keyloggers' detection perform inefficiently in the context of large-scale IoT networks. The aim of this project is to implement machine learning to solve the challenges of keyloggers' detection using network flow data.

**Objectives**

* Develop classification Model to detect Keylogger attacks from IoT network traffic using Machine Learning.
* Feature selection and random sampling shall be used to handle large and imbalanced datasets.
* Apply advanced algorithms like XGBoost, Gradient Boosting and MLP classifier to improve the accuracy of detection.
* Evaluate the performance in test data using metrics such as ROC score, F1 score, confusion matrix, precision and recall score, balanced accuracy, etc.

**Significance**

The high accuracy of the detection of keyloggers improves the security in the IoT environment further preventing data breaches and losses. In this respect, automation of detection with machine learning could turn out to be efficient and scalable.

**Related Research**

Some intrusion detection studies in the past have used machine learning models like Random Forest and Gradient Boosting. Detection of keyloggers in IoT networks is relatively less explored which is why this research is important to fill that gap.

(Bayzid et al., 2019)addressed underground insect trapped shield that depends on hailing program that underground insect trapped system plans that are always targeted by Keyloggers; It is, however in easy for Keylogger developers to evade this position strategy by using diverse ways to log the customer actions besides using Set Windows Hooked anyway.

There is a large proportion of the powerful detection tool being implemented and researched but it is quite challenging to detect Keyloggers accurately. (Kazi et al., 2023) detected the Keylogger by using TAKD calculations that can be integrated with day-to-day devices such as a switch, door, firewall, IDS etc. to enhance its key logging detection. The proposed TAKD calculation combined the anomaly-based detection feature and log-based process for overcoming the problem of signature-based detection.

# Materials and Methods

**Data Source**

The NF-BoT-IoT-v2 dataset will be used which is made up of 37,763,497 samples of which 99.64% are attack samples. It contains NetFlow features indicative of the network behaviors which will be used for the training of machine learning models against the corresponding keyloggers.

**Methods**

1. **Data Pre-processing**

* The data set is large where random sampling will be employed to draw a representative subset of the data.
* Handle the class imbalances by using a random sample that includes attack and benign samples to retain diversity.

1. **Feature Selection**

The selection of the most relevant features from the NetFlow data shall be done based on Mutual Information Score which computes the dependency between features and target variable for selecting those features that may be most informative about keyloggers.

1. **Model Training**

* Use the selected features to train different machine learning models such as Random Forest, Gradient Boosting and XGBoost.
* As above, the model will be nicely regularized with appropriate cross-validation.

1. **Testing Strategy**

* The data will be divided into training and test sets. Models that show the best performance according to accuracy, precision, recall and F1-score will be evaluated.
* The error rate of the models in detecting keylogger attacks will be measured using confusion matrices.

1. **Reasonable potential**

Feature selection with Mutual Information score guarantees that the most relevant features are chosen while dimensionality and computation are minimized. Random sampling enables one to process data in an efficient way, while the advanced models of machine learning have proved quite fitted for complex network traffic data to perform an accurate detection of keyloggers.

# Facilities and timetable

**Resources**

* **Hardware:** A high-performance computer with a GPU will be efficient for handling the big dataset and training machine learning models quickly.
* **Software:** Python, pandas, scikit-learn, libraries of feature selection and model training such as Random Forest Classifier, XGBoost, MLP classifier.

**Timeline**

|  |  |
| --- | --- |
| **Week** | **Steps** |
| **Week 1-2** | Data sampling, preprocessing, and feature selection using Mutual Information Score. |
| **Week 3-4** | Train machine learning models and validate results. |
| **Week 5-6** | Fine-tune models and perform cross-validation. |
| **Week 7-8** | Analyze results, write report, and prepare the final presentation. |

## Expected Results

The project is expected to

* Detection of keylogger attacks using Random Forest, Gradient Boosting and XGBoost models.
* Demonstration of feature selection with Mutual Information Score which improves the performance of detection by reducing the influence of irrelevant features and giving importance to the most discriminative features.
* Provide relevant information on the key features of network traffic related to the activity of keyloggers.

**Data Analysis Approaches**

The emphasis will be on model performance based on accuracy, precision, recall, and the F1-score. Confusion matrices and ROC curves will also be developed to report model effectiveness, with a focus on imbalance issues between attack and benign samples.

# References

Bayzid, M., Shoikot, M., Hossain, J., & Rahman, A. (2019). Keylogger Detection using Memory Forensic and Network Monitoring. *International Journal of Computer Applications*, *177*(11), 17–21. https://doi.org/10.5120/ijca2019919483

Kazi, A., Manthan Mungekar, Sawant, D., & Pankaj Mirashi. (2023). KEYLOGGER DETECTION. *International Research Journal of Modernization in Engineering Technology and Science*. https://doi.org/10.56726/irjmets37020

(2024). Uq.edu.au. https://rdm.uq.edu.au/files/b97fa3c0-ef9c-11ed-8007-832d3bc76ba8