MODEL RESEARCH

Organization Name: Infosys Springboard Internship

Document Title: Internship Research Documentation

Internship Title: Research and Information Gathering on Machine Learning Models for Cyber Threat Detection

Internship Type: Virtual Internship

Internship Duration: 2 Months

Date: 04/01/2026

Intern Name: KOTAKONDA SAI SANDEEP

1.Logistic Regression:

Model Overview

Logistic Regression is a supervised statistical learning model used for binary and multi-class classification. Despite its name, it is a classification algorithm, not a regression model. It estimates the probability that an input belongs to a particular class using a logistic (sigmoid) function.

Role in Cyber Threat Detection

In cybersecurity, Logistic Regression is primarily used as a baseline threat classifier. It converts raw security features—such as packet counts, login attempts, or URL characteristics—into a probability score representing the likelihood of malicious activity.

Why It Is Used

- Extremely fast to train and deploy
- Highly interpretable (coefficients show feature importance)
- Scales well to large datasets

Security teams prefer it when explainability and auditability are required.

Advantages in Cyber Threat Detection

- Produces clear probabilistic risk scores
- Easy to debug and validate
- Works well for linearly separable security patterns

Where It Is Used

- Phishing email classification
- Fraudulent login detection
- Risk scoring engines in SOC dashboards

Example

A system monitoring login attempts may use Logistic Regression to estimate the probability that a login is malicious based on features like IP reputation, time of access, and failed attempts. If the probability exceeds a threshold, the login is flagged.

2. Decision Tree:

Model Overview

A Decision Tree is a supervised learning model that makes decisions by recursively splitting data based on feature conditions. The result is a tree-like structure of if-then rules.

Role in Cyber Threat Detection

Decision Trees are used to translate complex security data into human-readable rules, making them valuable for intrusion detection and policy enforcement systems.

Why It Is Used

- Intuitive and interpretable
- Converts data into explicit security rules
- Easy to explain to non-ML stakeholders

Advantages in Cyber Threat Detection

- Clear reasoning behind alerts
- Helps analysts understand attack logic
- Useful for rule generation and compliance

Where It Is Used

- Intrusion Detection Systems (IDS)
- Firewall rule generation
- Security policy validation

Example

A decision tree may classify traffic as malicious if:

- Protocol = TCP
- Port = 445
- Packet rate > threshold

This directly mirrors real-world attack rules like SMB exploitation.

3. Random Forest:

Model Overview

Random Forest is an ensemble learning model that combines multiple decision trees to improve accuracy and reduce overfitting. Each tree is trained on a random subset of data and features.

Role in Cyber Threat Detection

It serves as a core detection engine for structured cybersecurity data such as network flows, endpoint telemetry, and malware features.

Why It Is Used

- Strong performance on noisy security data
- Handles high-dimensional features well
- Resistant to overfitting

Advantages in Cyber Threat Detection

- High detection accuracy
- Robust to incomplete or corrupted data
- Captures non-linear attack patterns

Where It Is Used

- Network intrusion detection
- Malware classification
- Insider threat analysis

Example

Random Forest can analyze NetFlow features (bytes, duration, packets, flags) to detect DDoS or scanning behavior with high reliability.

4. Support Vector Machine (SVM):

Model Overview

SVM is a supervised learning algorithm that finds an optimal boundary (hyperplane) separating different classes with maximum margin.

Role in Cyber Threat Detection

SVMs are used when high accuracy is required with limited labeled data, especially in malware and intrusion detection.

Why It Is Used

- Effective in high-dimensional spaces
- Strong theoretical foundations
- Works well with small datasets

Advantages in Cyber Threat Detection

- Precise classification boundaries
- Good generalization ability
- Handles complex attack patterns

Where It Is Used

- Malware family classification
- Binary intrusion detection
- Network traffic classification

Example

An SVM can classify executable files as malicious or benign using opcode frequency features, even when labeled samples are limited.

5. Gradient Boosting (XGBoost / LightGBM):

Model Overview

Gradient Boosting builds models sequentially, where each new model corrects the errors of the previous ones. XGBoost and LightGBM are optimized implementations.

Role in Cyber Threat Detection

This model is widely used as a high-performance classifier for structured security datasets.

Why It Is Used

- Often outperforms deep learning on tabular data
- Highly tunable and efficient
- Handles missing values well

Advantages in Cyber Threat Detection

- Very high detection accuracy
- Strong feature interaction modeling
- Production-ready performance

Where It Is Used

- Endpoint detection and response (EDR)
- Malware detection
- Network traffic analysis

Example

A gradient boosting model can detect malicious URLs by combining features such as domain age, entropy, and lexical patterns.

6. Naive Bayes:

Model Overview

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features.

Role in Cyber Threat Detection

It is mainly used for text-based threat detection, such as spam and phishing.

Why It Is Used

- Extremely fast
- Works well on high-dimensional text data
- Requires minimal training data

Advantages in Cyber Threat Detection

- Low computational cost
- Effective for email and message filtering
- Easy to deploy at scale

Where It Is Used

- Spam filtering
- Phishing email detection
- Malicious message classification

Example

Naive Bayes can classify emails as phishing based on word probabilities like "urgent", "verify", or "account".

7. k-Means Clustering:

Model Overview

k-Means is an unsupervised learning algorithm that groups data into clusters based on similarity.

Role in Cyber Threat Detection

It is used for behavior profiling and anomaly grouping, not direct threat detection.

Why It Is Used

- No labeled data required
- Simple and fast
- Useful for exploratory analysis

Advantages in Cyber Threat Detection

- Identifies unknown behavior patterns
- Helps discover new attack groups
- Supports analyst investigation

Where It Is Used

- Network traffic clustering
- User behavior profiling
- Malware sample grouping

Example

k-Means can cluster network traffic flows to separate normal user behavior from suspicious scanning or beaconing patterns.

8. Isolation Forest:

Model Overview

Isolation Forest is an unsupervised anomaly detection model that isolates anomalies by randomly partitioning data.

Role in Cyber Threat Detection

It is designed to detect rare and abnormal behaviors, making it suitable for zero-day attack detection.

Why It Is Used

- No attack signatures required
- Efficient on large datasets
- Detects novel threats

Advantages in Cyber Threat Detection

- Effective for unknown attacks
- Low training cost
- Scales well

Where It Is Used

- Insider threat detection
- Network anomaly detection
- Zero-day attack monitoring

Example

An Isolation Forest may flag a user who suddenly accesses sensitive systems at unusual hours from new locations.

9. Autoencoders:

Model Overview

Autoencoders are neural networks trained to reconstruct input data. High reconstruction error indicates anomalies.

Role in Cyber Threat Detection

They learn normal system behavior and detect deviations as potential threats.

Why It Is Used

- Powerful for complex patterns
- No labels required
- Captures non-linear relationships

Advantages in Cyber Threat Detection

- Detects stealthy anomalies
- Effective for network and host data
- Learns deep representations

Where It Is Used

- Host-based intrusion detection
- Network traffic anomaly detection
- System call monitoring

Example

An autoencoder trained on normal traffic will produce high reconstruction error when malware-generated traffic appears.

10. Long Short-Term Memory (LSTM):

Model Overview

LSTM is a type of recurrent neural network designed to model sequential and temporal data.

Role in Cyber Threat Detection

LSTMs detect multi-stage and time-dependent attacks, which traditional models miss.

Why It Is Used

- Captures temporal dependencies
- Detects slow, stealthy attacks
- Handles sequential logs and flows

Advantages in Cyber Threat Detection

- Strong against advanced persistent threats
- Models attack progression
- Reduces false positives

Where It Is Used

- Log anomaly detection
- Botnet and C2 detection
- Advanced intrusion detection

Example

An LSTM can detect a sequence of events—port scan → privilege escalation → data exfiltration—as a coordinated attack rather than isolated actions.