# ****Cybersecurity Threat Classification Using Machine Learning****

## ****1. Introduction****

## With the growing sophistication of cyber threats, **machine learning-driven network intrusion detection** has become essential in modern cybersecurity systems. This project aims to **classify network intrusions with high precision** using various machine learning models. By leveraging **advanced feature engineering and ensemble techniques**, we enhance the detection accuracy of malicious activities in network traffic.

## ****2. Dataset and Preprocessing****

### ****Dataset****

* The dataset was **downloaded from Kaggle** and contains **network traffic data with multiple attack types**.
* It consists of **42 features**, including **duration, protocol type, source/destination bytes, login attempts, and various network-related statistics**.

### ****Data Preprocessing****

* **Handled missing values** and **normalized numerical features** for better model performance.
* **Encoded categorical features** (e.g., protocol type, service, flag) using **one-hot encoding**.
* **Split the dataset** into **80% training** and **20% testing** sets.

## ****3. Feature Selection****

* Applied **correlation analysis** and **feature importance ranking** (Random Forest) to remove irrelevant features.
* Retained **highly informative attributes** to improve model efficiency.

## ****4. Model Selection and Training****

The following machine learning models were implemented and trained:

* **Logistic Regression**
* **Random Forest**
* **Gradient Boosting**
* **Support Vector Machine (SVM)**
* **K-Nearest Neighbors (KNN)**

Each model was evaluated using **train-test split** and cross-validation.

## ****5. Model Evaluation****

The performance of each model was assessed based on **Accuracy, Precision, Recall, and F1-Score**. Among all models, **Random Forest achieved the highest accuracy of 99.78%**, followed closely by **Gradient Boosting and KNN**. **SVM and Logistic Regression** performed well but were slightly less effective compared to ensemble models.

## ****6. Visualization and Insights****

* **Confusion Matrix**: Showcased how well each model classified attacks vs. normal traffic.
* **Feature Importance Graphs**: Identified **key attributes** influencing attack detection.

## ****7. Conclusion****

* **Random Forest** outperformed other models with **99.78% accuracy**, making it the best choice.
* **Gradient Boosting** showed high performance and **strong generalization ability**.
* **SVM and Logistic Regression** were slightly less effective than ensemble models.
* **KNN performed well but required more computation for large datasets.**

## ****8. Real-World Application & Scalability****

Beyond this project, the trained models can be **integrated into real-time Intrusion Detection Systems (IDS)** for enhanced security monitoring. The following enhancements can be explored for enterprise deployment:

* **Deployment as an API for seamless integration** into security tools.
* **Adaptive learning** to update models dynamically with new cyberattack patterns.
* **Integration with AI-powered anomaly detection** for early threat mitigation.