VARDHAMAN
COLLEGE OF ENGINEERING
(AUTONOMOUS)
Affiliated to JNTUH, Approved by AICTE, Accredited by NAAC with A++ Grade, ISO 9001:2015 Certified
Kacharam, Shamshabad, Hyderabad – 501218, Telangana, India
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEEERING (AI&ML)

Mini Project Submission (A. Y. 2025-2026 R-22)

1. Project Title:

Anomaly Detection in Host Systems Using Random Forest-Based HIDS

2. Team Details:

Team Number :22

Team Members:

- Bhavanari Sri Akshitha(23881A6670)
- Rasthapuram Malleshwari(23881A66A9)
- Vislavath Arjun Naik(23881A66C5)
- Guide Name: Ms. Shaista Farhat
- Designation: Assistant professor

3. Problem Statement:

Background

With the increasing reliance on computer systems for critical operations, ensuring the security and integrity of host systems has become a paramount concern. Host-based attacks such as malware infections, unauthorized access, privilege escalation, and system file tampering are growing in both frequency and sophistication. Traditional signature-based intrusion detection systems are effective against known threats but often fail to detect new, unknown, or evolving attack patterns. This has led to a growing interest in anomaly detection techniques that can identify unusual behaviors indicative of potential security breaches. Machine learning, especially ensemble methods like Random Forest, offers powerful capabilities for detecting such anomalies in host systems.

Problem Definition

Current Host-based Intrusion Detection Systems (HIDS) struggle to accurately detect zero-day attacks and unusual behavioral patterns due to their dependency on predefined rules or known attack signatures. This results in high false negative rates and exposes systems to undetected threats. Additionally, many HIDS solutions suffer from high false positives, leading to alert fatigue and inefficient threat response. There is a pressing need for a more intelligent and adaptive approach that can distinguish between normal and malicious behaviors with greater precision. This project aims to address these limitations by developing a Random Forest-based anomaly detection model for HIDS that can effectively identify threats with improved accuracy and reliability.

Need for the Project

- **Security Breach Prevention**: Timely detection of anomalies can prevent data theft, system damage, and service disruption.
- **Adaptability**: Random Forest models can generalize well and adapt to new patterns without requiring predefined attack signatures.
- **Reduced False Positives**: The ensemble nature of Random Forest helps reduce noise and improves the system's trustworthiness.
- **Cost-Effective Defense**: Host-based solutions do not require additional hardware or network

.

Target Users/Beneficiaries

- **IT Administrators & System Security Teams**: They will benefit from early and accurate detection of intrusions, enabling faster response and mitigation.
- **Organizations and Enterprises**: Protecting critical infrastructure and sensitive data from internal and external threats.
- **Government and Defense Agencies**: Securing mission-critical systems where data confidentiality and system integrity are essential.
- **Software Vendors & Cloud Service Providers**: To embed smart security solutions into their platforms and improve user trust.

4. Project Objectives:

Objective 1:

To develop a Host-based Intrusion Detection System (HIDS) using the Random Forest algorithm capable of classifying host system activity as normal or anomalous.

Objective 2:

To implement a lightweight alerting mechanism that notifies administrators upon anomaly detection, ensuring real-time responsiveness, low false positive rates, and the system's adaptability to novel attack patterns

5. Abstract:

The development of a Host-based Intrusion Detection System (HIDS) using the Random Forest algorithm to classify host activities as normal or anomalous. The system analyzes various host-level parameters such as login attempts, file access patterns, and CPU usage to identify potential threats. A key component of this work is introduce a lightweight alerting mechanism that provides real-time notifications to administrators upon detecting anomalies. The system is further designed to be scalable and capable of detecting previously unseen attacks. To validate and train the model, benchmark cybersecurity datasets such as the NSL-KDD and ADFA-LD were used. These datasets contain rich host-based activity logs with labeled instances of both normal operations and preprocessing steps such as feature selection, normalization, and encoding were applied to ensure compatibility with the machine learning pipeline and improve detection accuracy. This feature enhances the system's responsiveness while maintaining low false positive rates. The design is adaptive, scalable, and capable of detecting novel attack patterns and also demonstrates how machine learning can be effectively applied in cybersecurity to improve host protection through intelligent monitoring and prompt alerting.

Keywords—Host-based Intrusion Detection System, Random Forest, Anomaly Detection, Cybersecurity, Machine Learning, Alert Mechanism.

6. Tools and Technologies Used:

Software & Platforms

- **Python**: Primary programming language used for data preprocessing, model development, and evaluation.
- **Jupyter Notebook / Google Colab**: Interactive development environment for writing and running Python code.
- **Anaconda**: Python distribution that includes essential data science libraries and Jupyter Notebook.
- **Scikit-learn**: Machine learning library used for implementing the Random Forest algorithm and evaluation metrics.
- **Pandas & NumPy**: Libraries for data manipulation and numerical operations.
- **Matplotlib & Seaborn**: For data visualization and graphical representation of results.

Datasets

- **NSL-KDD** or **ADFA-LD**: Benchmark datasets commonly used in Host-based Intrusion Detection research.

Operating System

- **Windows/Linux**: For running the HIDS, collecting host system logs, and executing scripts.

Technologies & Concepts

- **Machine Learning (Supervised Learning)**: For training the anomaly detection model.
- **Random Forest Classifier**: Ensemble algorithm used to detect anomalies in host behavior.
- **Anomaly Detection Techniques**: Based on behavioral deviation from historical norms.

7. Expected Deliverables:

- **Trained Machine Learning Model** for anomaly detection
- **Python Scripts** for data preprocessing, model training, and evaluation
- **Performance Report** with accuracy, precision, recall, and F1-score metrics
- **Dataset (Processed & Raw)** along with feature extraction code

8. SDG Mapping:

SDG No	Goal Title	Justification
8	Decent Work and Economic Growth	Enhances job security by protecting digital work environments
9	Industry, Innovation and Infrastructure	The project promotes technological innovation by applying machine learning for cybersecurity, helping to strengthen digital infrastructure resilience.

11	Sustainable Cities and Communities	Protects digital infrastructure of smart cities
16	Peace, Justice and Strong Institutions	By preventing cyber threats, the project supports secure digital environments and strengthens institutional security.

9. OBE Mapping:

Program Outcomes (POs)

PO number	Name of the PO Targeted	Justification
PO1	Engineering Knowledge	Applies machine learning and cybersecurity concepts to solve real-world security problems.
PO2	Problem Analysis	Analyzes host system logs and anomalies to detect cyber threats using statistical ML techniques.
PO3	Design/Development of Solutions	Designs and implements a Host-based IDS solution using Random Forest and Python tools.
PO5	Modern Tool Usage	Utilizes modern tools such as Python, scikit-learn, and Jupyter Notebook for implementation.
PO12	Life-long Learning	Encourages continuous learning of evolving technologies like AI and cybersecurity.

Program Specific Outcomes (PSOs)

PSO number	Name of the PSO Targeted	Justification
PSO1	Apply the knowledge of Artificial Intelligence to design, develop, and evaluate computational	The project applies AI and Random Forest to design an effective Host-based Intrusion Detection

	solutions for complex problems in diverse domains, such as healthcare, finance, and automation.	System for securing complex digital infrastructures
PSO2	Demonstrate expertise in using advanced ML tools, techniques, and frameworks to develop innovative solutions for data analysis, pattern recognition, and intelligent decision-making systems.	The project leverages advanced ML techniques, including Random Forest and feature selection, for intelligent anomaly detection and decision-making.

Course Outcomes (COs):

CO1: Apply fundamental and disciplinary concepts and methods in ways appropriate to their principal areas of study.

CO2: Demonstrate skill and knowledge of current information and technological tools and techniques specific to the professional field of study.

CO3: Identify, analyze, and solve problems creatively through sustained critical investigation.

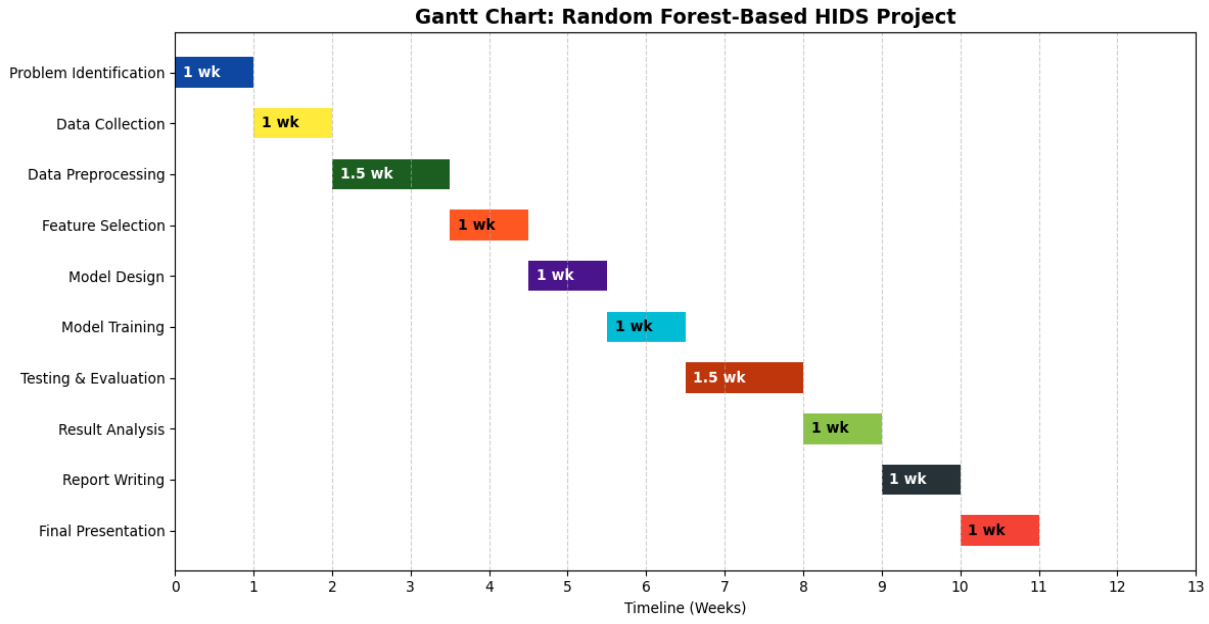
CO4: Demonstrate awareness and application of appropriate personal, societal, and professional ethical standards.

CO5: Design the various compensators and controllers for time invariant systems.

Bloom's Taxonomy Level Focused:

- **Apply** – Implementing machine learning algorithms.
- **Analyze** – Interpreting system logs and identifying anomalies.
- **Evaluate** – Comparing model results and refining the solution.
- **Create** – Designing and developing a complete HIDS solution.

10. Timeline



11. References

[1] B. Awotunde et al., “A Multi-level Random Forest Model-Based Intrusion Detection Using Fuzzy Inference System”, *Int. J. Comput. Intell. Syst.* (2023).
Link: https://link.springer.com/article/10.1007/s44196-023-00205-w
[2] Yukyung Shin1, Kangseok Kim,“Comparison of Anomaly Detection Accuracy of Host-based IDS based on Different ML Algorithms”, International J. Adv. Comp. Sci. Appl. (2020).
Link: https://thesai.org/Publications/ViewPaper?Volume=11&Issue=2&Code=IJACSA&SerialNo=33
[3]Alberto Miguel-Diez, Adri´an Campazas-Vega, Claudia ,Alvarez-Aparicio, Gonzalo,Esteban-Costales, and Angel Manuel Guerrero-Higuer,“ A systematic literature review of methods and datasets for anomaly intrusion detection, *Computers & Security”* (2022).
Link:https://www.researchgate.net/publication/389748138_A_systematic_literature_review_of_unsupervised_learning_algorithms_for_anomalous_traffic_detection_based_on_flows
[4]John Ring,Colin M. Van Oort,Samson Durst and Vanessa White,“Methods for Host-based Intrusion Detection with Deep Learning”, *ACM* (2021)
Link:https://www.researchgate.net/publication/351338993_Methods_for_Host-Based_Intrusion_Detection_with_Deep_Learning
[5] Ehsan Aghaei 1, Gursel Serpen2 ,”Host-based anomaly detection using Eigentraces feature extraction and one-class classification on system call trace data”, *Sensors* (2019).
Link: https://arxiv.org/pdf/1911.11284
[6] Caiwu Lu Yunxiang Cao and Zebin Wang,” Research on Intrusion Detection Based on an Enhanced Random Forest Algorithm”(2024)
Link: https://www.mdpi.com/2076-3417/14/2/714?
[7] Monire Norouzi , Zeynep Gürka¸s-Aydın , Özgür Can Turna , Mehmet Yavuz Ya ˘gci , Muhammed Ali Aydin and Alireza Souri ,” A Hybrid Genetic Algorithm-Based Random Forest Model for Intrusion Detection Approach in Internet of Medical Things”(2023)
Link: https://www.mdpi.com/2076-3417/13/20/11145
[8] V. Priya, I. Sumaiya Thaseen, Thippa Reddy Gadekallu, Mohamed K. Aboudaif, Emad Abouel Nasr, ” Robust Attack Detection Approach for IIoT Using Ensemble Classifie”(2021)
Link: https://arxiv.org/abs/2102.01515
[9] Zhewei Chen, Wenwen Yu, Linyue Zhou,” ADASYN-Random Forest Based Intrusion Detection Model”(2021)
Link: https://arxiv.org/abs/2105.04301
[10] R. Lalduhsaka , Nilutpol Bora , and Ajoy Kumar Khan,” Anomaly-Based Intrusion Detection Using Machine Learning: An Ensemble Approach”(2022)
Link: https://www.igi-global.com/article/anomaly-based-intrusion-detection-using-machine-learning/311466

Signature of the Supervisor **Signature of Coordinator** **Signature of HOD**