WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Integrated Intrusion Detection and Prevention System using Honeypot in Cloud Computing

Prof. Jyothis K P[1], Harsha Nayaka K[2], Amar[3], Vinayak[4]

[1]Assistant Professor Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management Bengaluru, India
Prof_jyothis-cse@dsatm.edu.in,
[2]Student, 3th Year, B.E Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bengaluru, India
1dt23cs406@dsatm.edu.in
[3]Student, 3th Year, B.E Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bengaluru, India
1dt23cs401@dsatm.edu.in
[4]Student, 3th Year, B.E Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management Bengaluru, India
1dt23cs416@dsatm.edu.in

ABSTRACT-

Cyber threats have emerged as the most significant concern for cloud-based systems, organizations, and digital infrastructure ever since the widespread adoption of cloud computing. Our objective here is to use honeypot-based deception and machine learning analytics to detect, analyze, and prevent network intrusions in real time. In this paper, we present an integrated Intrusion Detection and Prevention System (IDPS) using virtual honeypots to track, log, and study attacker behavior within cloud environments.

By deploying decoy systems that mimic real services in cloud platforms and using intelligent models that analyze incoming traffic, this approach predicts potential intrusions and takes preventive measures proactively. Just as wearable devices revolutionized sports injury prediction, this solution revolutionizes cloud defense — identifying attack risks before they manifest, thus allowing administrators to strategize cloud security better and take defensive actions in advance.

Keywords — intrusion detection, cloud computing, honeypot, machine learning, IDS/IPS, threat prediction, cybersecurity, anomaly detection, network forensics, cloud security architecture, attacker profiling, log analysis, virtual decoys, automation, cyber defense.

I. INTRODUCTION-

The global cloud computing ecosystem is a multi-billion dollar industry, with increasing emphasis placed on securing cloud environments from evolving cyber threats. A multitude of factors influence the risk of intrusion or compromise in such systems. This research focuses on analyzing parameters like network traffic patterns, protocol anomalies, access behavior, and geographical data to assess the vulnerability of cloud-hosted services. Our system integrates honeypot deployments with intelligent classifiers that detect and predict malicious activity based on these contextual signals.

The detection model identifies the likelihood of an intrusion attempt by analyzing real-time and historical data collected from strategically placed honeypots that mimic legitimate cloud services. These decoys attract attackers, allowing the system to capture rich threat intelligence including source IP, attack vectors, payloads, and timing. The classifier model uses this to estimate intrusion probability, while a parallel regressor model evaluates the severity of the threat — minor, moderate, or critical — based on system exposure, attack sophistication, and payload complexity.

Furthermore, our system categorizes threats, triggering appropriate prevention mechanisms ranging from logging and IP blocking to automated isolation of virtual instances. These insights help administrators implement predictive defense strategies, prioritize patching schedules, and optimize incident response plans. In turn, this leads to reduced downtime, minimized data loss, and strengthened trust in cloud platforms. The approach also supports proactive policy enforcement and scalable threat modeling, ultimately building a secure, self-defending cloud ecosystem capable of adapting to modern cybersecurity challenges.

II. LITERATURE REVIEW

Thesis on Predictive Intrusion Modeling in Cloud Computing

The goal of this thesis is to investigate the potential of predictive modeling for cybersecurity threats in cloud environments. This work was inspired by the rapid growth of cloud services and the increasing frequency of sophisticated attacks targeting cloud infrastructure. In this review, three main investigations were conducted:

Predicting Intrusion Impact and Response Time Using Attack Records

For this investigation, three datasets comprising known cloud attack records (including IDS/IPS logs, honeypot event captures, and incident response timelines) were analyzed using different machine learning algorithms to build predictive models. These models estimate how long it takes to detect and respond to attacks based on threat severity and system exposure.

Predicting Intrusions in Virtual Cloud Environments Using Resource Exposure Records

This study examined the relationship between cloud resource exposure (e.g., public IP usage, open ports, virtual machine uptime) and intrusion likelihood. The primary goal was to predict how long a virtual server or container could remain online before becoming a target for scanning or exploitation attempts.

Predicting Intrinsic Attack Risk Using Network Telemetry and Traffic Signatures

A significant percentage of attacks in cloud systems stem from automated bots and adversarial scripts probing for vulnerabilities. These patterns can be detected via live traffic inspection and network telemetry. This research aimed to predict when and where an intrusion is likely to occur using data from honeypots, traffic flow records, and system logs.

A Narrative Review of Predictive Techniques in Cybersecurity

Security breaches are a common and costly occurrence in cloud-based platforms, with substantial financial, operational, and reputational consequences for businesses and users. There are numerous methods for identifying risks and vulnerabilities, but selecting the correct statistical and AI-based approaches is crucial, as poor decisions can lead to system compromise.

This review aims to:

- Outline commonly implemented methods for identifying intrusion risk factors.
- Encourage researchers to carefully evaluate traffic features, system metadata, and access patterns in relation to potential attacks.
- Describe advances in ML-based threat modeling and real-world applications of predictive intrusion detection.

Dataset Details (Cybersecurity Equivalent)

The dataset comprises two primary components: the **Attack Record File** and the **Session Log File**.

- **Attack Record File** includes detailed information about known attacks on cloud-based systems over two years. Key fields include:
    - **Source IP**, **Destination IP**, **Timestamp**
    - **Payload Type**, **Protocol Used**, **Attack Severity**
    - **Detection Method** (e.g., honeypot, IDS rule, anomaly score)
    - **Mitigation Action** and response latency
- **Session Log File** documents regular activity from both benign and suspicious users. Indexed by:
    - **VM ID**, **Port**, **Session ID**
    - **Geolocation**, **User-Agent**, **Protocol Details**
    - **Service Type** (e.g., SSH, HTTP, FTP), **Usage Duration**

This structure enables granular analysis of attack vectors in their real network and system context.

Data Integration and Enrichment

Fields such as IP address, timestamp, and VM ID serve as primary relational identifiers, linking the activity logs with honeypot captures. This linkage facilitates advanced detection of stealthy intrusions, lateral movements, and repeat offenders.

Additionally, telemetry from firewalls, virtual switches, and containers includes resource metrics like CPU/memory usage spikes, abnormal I/O operations, and port scans. These are used to infer system stress and potential breaches.

Smart Monitoring Using Cloud-native Telemetry

Cloud monitoring tools (e.g., AWS CloudTrail, Azure Monitor) and custom agents simulate the role of "smartwatches" in cybersecurity by continuously logging operational metrics such as:

- **Process Anomalies**

- **Unusual Login Patterns**
- **Data Exfiltration Indicators**
- **File Integrity Changes**

Machine learning models analyze this rich, continuous stream to anticipate zero-day threats and suggest early mitigations.

Behavioral Analytics for Intrusion Prediction

As with injury prediction in athletes using biometric and environmental data, intrusion prediction uses a mix of:

- **Historical attack data** (known threats)
- **Live system behavior** (like GPS/biometrics for players)
- **Environmental conditions** (open ports, public IPs, weak configurations)

This supports proactive response strategies, such as:

- Preemptive VM isolation
- Auto-scaling of firewalls
- Dynamic reconfiguration of access control lists (ACLs)

Application of the Framework

The proposed predictive IDPS model contributes to:

- Reducing intrusion attempts via honeypot deception
- Prioritizing vulnerabilities using machine learning severity scores
- Enhancing cloud monitoring with context-aware alerting
- Improving incident response workflows by anticipating attack types

III. Problem statement

Task 1

Goal: Develop a machine learning–based IDPS model that analyses key factors such as:

- network traffic behavior,
- user access logs,
- honeypot-captured data,
- attack history,

to predict the **probability of intrusion** or malicious activity in a cloud computing environment.

Just like tracking athlete metrics, the system evaluates protocol behavior, packet frequency, port access attempts, login irregularities, and origin patterns to assign a risk score to incoming traffic.

Task 2

Goal: Create a reliable alert system that supports **early intervention and automated defense** to improve cloud infrastructure safety, using **visualizations such as:**

- intrusion graphs,
- IP heat maps,
- attack severity charts, and
- behavioral anomaly timelines.

These visual tools help administrators assess threat dynamics over time and make proactive adjustments to access control and firewall rules.

Task 3

Goal: Assess the **severity of detected threats**, categorizing each incident as:

- **Minor (benign or failed attempt)**
- **Moderate (reconnaissance or partial intrusion)**
- **Critical (successful data breach or sustained attack)**

The model also identifies whether:

- **one node (VM/container) is targeted** or
- **multiple resources are under coordinated attack**

This allows the system to recommend if isolation, escalation, or network-wide lockdown is required.

Dataset Integration (Cybersecurity Perspective)

To ensure a holistic understanding of intrusions and attack behavior, the dataset integrates data from multiple sources:

Data Sources:

- **Honeypot Logs**: Simulated vulnerable services capturing attacker commands, IPs, payloads, and access attempts.
- **Cloud Monitoring Tools**: Collect real-time metrics such as:
    - request frequency,
    - login attempts,
    - port access,
    - abnormal process creation,
    - and protocol misuse.
- **System Telemetry**: CPU/memory usage, packet rates, bandwidth spikes indicating exploitation or brute-force attacks.
- **Threat Intelligence Feeds**: Include blacklisted IPs, known botnets, and malware signatures.
- **Incident Records**: Historical data from prior breaches including:
    - attack vectors,
    - impact level,
    - duration,
    - mitigation steps,
    - and system downtime.

Behavioral Profiling for Intrusion Detection

In the same way that sports systems analyze physiological and environmental stressors, this cybersecurity model evaluates:

- attacker “fingerprints” through traffic signatures,
- environmental conditions (open ports, public-facing APIs),
- and usage baselines to detect subtle deviations.

This approach enables the system to **predict and prevent** future intrusions by correlating known behaviors with real-time events — similar to predicting athlete injuries by recognizing fatigue or overtraining symptoms.

IV. Methodology

Predictive Intrusion Detection Using Machine Learning in Cloud Computing

This process involves comparing data collected during **intrusion events** and **normal network behavior** to identify **key indicators of potential threats**. By analyzing a **correlation map of traffic features**, we can pinpoint those most strongly associated with malicious activity, such as abnormal port access frequency, rapid changes in packet flow, or unexpected protocol usage. These insights allow for the creation of predictive models that classify cloud traffic patterns in real time.

Machine learning algorithms are leveraged to **improve the detection accuracy**, utilizing labeled datasets (normal vs. attack traffic) to train classifiers capable of identifying subtle deviations and behavioral anomalies. These models **become more accurate over time** as they are exposed to a broader range of threat signatures and attack behaviors.

This predictive approach supports both **detection and prevention** by flagging high-risk traffic and triggering automated countermeasures. For example, abnormal access to a honeypot or traffic spikes from known blacklisted IPs can initiate alerts and enforce blocking rules.

Final Features Selected for Intrusion Prediction:

- **Frequency of TCP/UDP packets**
- **Unusual protocol usage (e.g., Telnet/FTP over HTTPS)**
- **Inbound/outbound traffic ratio**
- **Geolocation of IP addresses**
- **Time-based anomalies (nighttime scans, burst access)**

Data Preprocessing and Feature Engineering

- **Handling Missing Values**: Null values in logs or metadata are cleaned.
- **Normalization**: Feature scaling ensures consistent treatment of large and small numeric values (e.g., packet size, access counts).
- **Outlier Detection**: Helps to isolate spike-based anomalies that may indicate scanning or DDoS attempts.

Feature engineering transforms raw data into actionable intelligence, including:

- Sudden spikes in session attempts
- Rapid failed login bursts
- Abnormal source-destination communication pairs

Machine Learning Models Applied

Binary Classification (Is this traffic malicious?)

- **Random Forest Classifier**
- **Gradient Boosting (XGBoost)**: Best performing model due to ensemble architecture
- **Logistic Regression**: Performed poorly due to non-linearity and high sparsity
- **Decision Tree**: Moderate performance with clear rule-based interpretability

Severity Estimation (How serious is the threat?)

- **Regression Models** used to predict breach impact:
    - **Linear Regression**: Used as a baseline (performed poorly due to dimensionality)
    - **Support Vector Regressor (SVR)**: Better than Linear but not optimal
    - **Random Forest Regressor**: Best results; handled complex non-linear relationships

Multi-label Classification (Which systems/components are at risk?)

- Predicts:
    - Whether web services, databases, or authentication systems are targeted.
    - Potential attacker intentions (e.g., data theft, service disruption, lateral movement)

Evaluation Metrics

Models are validated using standard performance metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**

- **Mean Squared Error** (for regression)
- **ROC-AUC** (for classification confidence)

XGBoost outperformed other classifiers in identifying attack behavior, while **Random Forest Regressor** delivered the lowest error rate in estimating severity.

Real-Time Honeypot Data Integration

Live traffic from honeypots is processed continuously. Alerts are generated in real time when:

- Access patterns deviate from historical baselines
- Known malicious signatures are detected
- Anomalous behavior correlates strongly with past breach events

These alerts feed into a **decision support framework** that can:

- Automatically block traffic
- Notify administrators
- Launch forensic data collection

V. Conclusion

Intrusion prediction using machine learning offers significant advancements in the realm of cybersecurity, providing a data-driven approach to enhance cloud infrastructure resilience and data protection. By analyzing key network and behavioral metrics collected from honeypot systems and monitoring tools — such as connection frequency, traffic anomalies, geographic origin, and protocol misuse — this research demonstrates how machine learning can be effectively deployed to predict potential intrusions. The use of real-time telemetry allows for the identification of attack patterns that indicate elevated threat risks, enabling **timely interventions** to prevent system compromise.

This study highlights the value of predictive models in cybersecurity. Not only can they detect and classify malicious traffic, but they also help estimate threat severity and identify specific assets (e.g., servers, databases, APIs) most vulnerable to attack. The findings confirm that combining honeypot-derived deception data with advanced machine learning algorithms provides crucial insights into network vulnerability and system exposure.

Furthermore, the approach can be expanded to develop **adaptive prevention strategies**, tailored to the unique configuration and usage patterns of individual cloud deployments. These dynamic models can continuously update based on new threats, ensuring that security measures evolve in sync with the rapidly changing threat landscape. As cybersecurity tools and monitoring systems advance, the potential for **real-time threat modeling and automated response** will further mature, offering robust, scalable, and intelligent defense mechanisms.

In conclusion, the application of machine learning to intrusion detection and prevention represents a **paradigm shift** in how cloud security is approached. It empowers system administrators, security analysts, and organizations to make informed, proactive decisions and contributes to the broader goal of building **self-healing, threat-resilient cloud environments**.

10. Individual Contribution:

1. **Data Collection and Organizing:** Harsha
2. **Literature Review:** Vinayak
3. **Exploratory Analysis:** Harsha Nayaka,Amar
4. **Creating Learning Models:**

IDP: Harsha Nayaka,Amar,Vinayak

5. **Analyzing Accuracy Among All codes and Reasoning for the Best Output:**

Data Classification Problem : Amar

Writing code for prediction : Harsha Nayaka

VI. References

1.**Kampakis, S.** (2016). *Predictive Modelling of Football Injuries*. Doctoral Thesis, University College London (UCL).

▶ Adapted for predictive modeling of intrusion risk using historical incident data and behavioral indicators.

2.**Ruddy, J. D., & Cormack, S. J.** *Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches*. [NCBI]

▶ Inspires comparative statistical analysis of threat risk modeling techniques such as logistic regression, decision trees, and ensemble learning in cybersecurity.

3.**ResearchGate.** *A Machine Learning Approach to Assess Injury Risk in Elite Youth Football Players*.

▶ Adapted as a foundation for applying ML classifiers to assess risk of cyber-intrusions using behavioral telemetry and honeypot logs.

4.**Akobeng, A. K.** (2007). *Understanding Diagnostic Tests 1: Sensitivity, Specificity, and Predictive Values*. Acta Paediatr, 96, 338–341.

▶ Used to evaluate cybersecurity model effectiveness — especially confusion matrix metrics like precision, recall, and F1-score.

5.**Altman, N., & Krzywinski, M.** (2015). *Association, Correlation and Causation*. Nature Methods, 12, 899–900.

▶ Referenced for proper interpretation of feature correlation analysis in intrusion detection systems.

6.**Chen, T., & Guestrin, C.** (2002). *XGBoost: A Scalable Tree Boosting System*.

▶ Core reference for implementing XGBoost as an ensemble method to classify and rank threat likelihood based on honeypot data.

7.**Bekkerman, R.** *The Present and Future of the KDD Cup Competition: An Outsider's Perspective*.

▶ Serves as an insight into practical applications and future directions of large-scale data mining for threat intelligence and intrusion prediction.