AI-POWERED REAL-TIME INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

A Project in Cybersecurity and Artificial Intelligence

[1]Saar Patwari[1], [2]Vraj Purohit[2], [3]Yogita Sable[3], Divya Salve

[1]BE Student (Final Year), [2]BE Student (Final Year), [3]BE Student (Final Year), BE Student (Final Year)
[1]Department of Computer Engineering[st] Author,
[1]Name of organization of 1[st] Pillai College of Engineering, New Panvel, India

Abstract : The rapid growth of cyber threats has exposed critical limitations in traditional signature-based Intrusion Detection Systems (IDS), which fail against zero-day attacks, generate high false positives, and demand extensive manual analysis. This paper proposes an AI-powered real-time IDS leveraging two complementary machine learning models: Random Forest, a supervised ensemble classifier, and Isolation Forest, an unsupervised anomaly detection algorithm. The system is trained and evaluated on the NSL-KDD benchmark dataset. Random Forest achieved 98.2% detection accuracy while Isolation Forest achieved 96.5%, both significantly outperforming conventional rule-based approaches. The proposed system incorporates a real-time network traffic analysis pipeline using Wireshark, a preprocessing module, a graphical user interface, and SIEM-style log collection to simulate Security Operations Center workflows. The integration of machine learning with proactive threat detection marks a significant advancement toward intelligent and autonomous network security.

Index Terms **- Intrusion Detection System, Machine Learning, Random Forest, Isolation Forest, NSL-KDD, Anomaly Detection, Cybersecurity, SIEM.**

I. INTRODUCTION

The rapid advancement of technology has led to an exponential increase in cyber threats targeting individuals, organizations, and critical infrastructure worldwide. As networks grow in complexity and scale, attackers continuously develop sophisticated intrusion techniques that exploit previously unknown vulnerabilities, commonly referred to as zero-day attacks. Traditional Intrusion Detection Systems (IDS) have long served as the first line of defense in network security; however, their reliance on static, signature-based detection mechanisms has rendered them increasingly ineffective in modern threat landscapes.

Signature-based IDS operate by comparing incoming network traffic against a predefined database of known attack patterns. While effective against well-documented threats, these systems are inherently reactive and fail to detect novel or polymorphic attacks. Furthermore, they are notorious for generating high volumes of false positive alerts, which overwhelm security analysts and divert attention from genuine threats. The manual analysis required to process these alerts is time-consuming, error-prone, and unsustainable at scale. There is therefore an urgent need to transition from reactive, rule-based systems toward proactive, intelligent detection mechanisms powered by artificial intelligence and machine learning.

II. TYPE STYLE AND FONTS

Machine learning offers a powerful paradigm shift in intrusion detection. By learning patterns from historical network traffic data, ML models can generalize to detect both known and previously unseen attack types with high accuracy and minimal manual intervention. This paper proposes an AI-powered real-time IDS that combines two complementary machine learning approaches: Random Forest, a supervised ensemble learning algorithm for classification, and Isolation Forest, an unsupervised algorithm designed specifically for anomaly detection. Together, these models form a robust detection framework capable of identifying a wide variety of cyber threats in real time.

III. LITERATURE REVIEW

Intrusion Detection Systems have been an active area of research for several decades. Early IDS implementations relied exclusively on signature-based detection methods, which compare network packets against a database of known malicious patterns [1]. While effective for catalogued threats, these systems are inherently limited by their inability to detect zero-day exploits and evolving attack vectors. Network-Based IDS (NIDS) monitor traffic across an entire network segment, whereas Host-Based IDS (HIDS) focus on individual system logs and process activity. Hybrid systems combine both approaches to achieve broader coverage [2].

IV. PREPARE YOUR PAPER BEFORE STYLING

The adoption of machine learning for intrusion detection has gained significant momentum over the past decade. The NSL-KDD dataset, introduced by Tavallaee et al. [3], has become a widely accepted benchmark for evaluating ML-based IDS due to its improved distribution of attack categories over the older KDD'99 dataset. Researchers have demonstrated that ensemble methods such as Random Forest achieve high classification accuracy on this dataset [4]. Isolation Forest, proposed by Liu et al. [5], has shown strong performance in unsupervised anomaly detection scenarios where labeled training data is scarce or unavailable.

Despite these advances, most existing ML-based IDS implementations lack true real-time capability and do not integrate with Security Information and Event Management (SIEM) workflows. Furthermore, many studies rely solely on offline datasets without

capturing live network traffic. This paper addresses these gaps by implementing a complete end-to-end pipeline that captures live traffic via Wireshark, processes and classifies it in real time, and logs all detection events in a SIEM-compatible format for downstream analysis.

Dataset: NSL-KDD

The NSL-KDD dataset was used to train and evaluate the proposed machine learning models. It is a refined version of the original KDD'99 dataset and addresses two key limitations: the removal of redundant records that would otherwise bias classifiers toward majority classes, and a more realistic distribution of attack types that better reflects real-world network conditions [3]. The dataset comprises 41 features per network connection record, including duration, protocol type (TCP, UDP, ICMP), service type (HTTP, FTP, SMTP), connection flag, login status, and connection count to the same host. Attack categories in the dataset include Normal traffic, Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R) attacks.

A. Data Preprocessing

Raw network data collected via Wireshark was first converted to CSV format and subjected to a rigorous preprocessing pipeline. Categorical features such as protocol type, service, and connection flag were encoded using One-Hot Encoding to produce numerical representations suitable for machine learning. Missing values were identified and removed, and inconsistent records were discarded. Numerical features were normalized using Min-Max Scaling to ensure all values fell within a uniform range, preventing any single feature from dominating the model. Finally, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was applied to balance the class distribution and prevent model bias toward the majority normal traffic class.

B. Machine Learning Models

is collected on stock prices for sample firms

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions. It is robust against overfitting, handles high-dimensional feature spaces effectively, and provides feature importance scores that highlight the most discriminative network attributes. These properties make it well-suited for multi-class network intrusion detection tasks.

Isolation Forest is an unsupervised anomaly detection algorithm that isolates observations by randomly selecting a feature and splitting between the maximum and minimum values. Anomalous data points require fewer random partitions to isolate, making the algorithm highly efficient and scalable for real-time network traffic analysis without requiring labeled anomaly data.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to computeusualvariation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at lastturned into nominal interest rate andchange in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rateleads todecreasethe cash flow's present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).

I. SYSTEM ARCHITECTURE AND METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study,Data and Sources of Data, study's variables and analytical framework. The detailsare as follows;

3.1Population and Sample

KSE-100 index is an index of 100 companies selected from 580 companies on the basis of sector leading and market capitalization. It represents almost 80% weight of the total market capitalization of KSE. It reflects different sector company's performance and productivity. It is the performance indicator or benchmark of all listed companies of KSE. So it can be regarded as universe of the study.Non-financial firms listed at KSE-100 Index (74 companies according to the page of KSE visited on 20.5.2015) are treated as universe of the study and the study have selected sample from these companies.

is collected on stock prices for sample firms

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions. It is robust against overfitting, handles high-dimensional feature spaces effectively, and provides feature importance scores that highlight the most discriminative network attributes. These properties make it well-suited for multi-class network intrusion detection tasks.

Isolation Forest is an unsupervised anomaly detection algorithm that isolates observations by randomly selecting a feature and splitting between the maximum and minimum values. Anomalous data points require fewer random partitions to isolate, making the algorithm highly efficient and scalable for real-time network traffic analysis without requiring labeled anomaly data.

Consumer Price Index (CPI) is used as a proxy in this study for inflation rate. CPI is a wide basic measure to computeusualvariation in prices of goods and services throughout a particular time period. It is assumed that arise in inflation is inversely associated to security prices because Inflation is at lastturned into nominal interest rate andchange in nominal interest rates caused change in discount rate so discount rate increase due to increase in inflation rate and increase in discount rateleads todecreasethe cash flow's

present value (Jecheche, 2010). The purchasing power of money decreased due to inflation, and due to which the investors demand high rate of return, and the prices decreased with increase in required rate of return (Iqbal et al, 2010).

This is assumed that decrease in the home currency is inverselyassociated to share prices (Jecheche,2010). Pan et al. (2007) studied exchange rate and its dynamic relationship with share prices in seven East Asian Countries and concludethat relationshipof exchange rate and share prices varies across economies of different countries. So there may be both possibility of either exchange rate directly or inverselyrelated with stock prices.Oil prices are positively related with share prices if oil prices increase stock prices also increase (Iqbal et al, 1012).Ataullah (2001) suggested that oil prices cause positive change in the movement of stock prices. The oil price has no significant effect on stock prices (Dash & Rishika, 2011).Six month T-bills rate is used as proxy of interest rate. As investors arevery sensitive about profit and where the signals turn into red they definitely sell the shares. And this sensitivity of the investors towards profit effects the relationship of the stock prices and interest rate, so the more volatility will be there in the market if the behaviors of the investors are more sensitive. Plethora (2002)has tested interest rate sensitivity to stock market returns, and concluded an inverse relationship between interest rate and stock returns. Nguyen (2010) studies Thailand market and found thatInterest rate has aninverse relationship with stock prices.

The study follow Fama and McBeth two pass regressionto test these asset pricing models.The Durbin Watson is used to check serial correlation and measures the linear association between adjacent residuals from a regression model. If there is no serial correlation, the DW statistic will be around 2. The DW statistic will fall if there is positive serial correlation (in worst case, it will be near zero). If there is a negative correlation, thestatistic will lie somewhere between 2 and 4. Usually the limit for non-serial correlation is considered to be DW is from 1.8 to 2.2. A very strong positive serial correlation is considered at DW lower than 1.5 (Richardson and smith, 1993).

According to Richardson and smith(1993) to make the model more effective and efficient the selection criteria for the shares in the period are: Shares with no missing values in the period, Shares with adjusted $R^2 < 0$ or F significant (p-value) >0.05of the first pass regression of the excess returns on the market risk premium are excluded. And Shares are grouped by alphabetic order into group of 30 individual securities (Roll and Ross, 1980).

$$R_i - R_f = (R_m - R_f)\beta \qquad (3.1)$$

$_{i\ f}$ isMonthly risk free rate, R_m isMonthly return of market and βis systematic risk (market risk).

$_{i}$ - R_f of each security is estimated from a time series share prices of KSE-100 index listed shares for each period under consideration. And for the same periodthe market Premium R_m - $R_{f\ i}$ - R_f on the market premium R_m - R_f

$$\hat{R}_i = \gamma_0 + \gamma_1\beta_1 + \epsilon \qquad (3.2)$$

Where λ_0= intercept, $\hat{R}_I$is average excess returns of security i,$\beta_I$isestimated be coefficient of security I and Є is error term.

$$R_i - R_f = \beta_i f_1 + \beta_{i2} f_2 + \beta_{i3} f_3 + \beta_{i4} f_4 + \epsilon \qquad (3.3)$$

$_{f}$ is risk free rate, β_i is the sensitivity of stock i with factors and ϵ is the error term.

$$\hat{R} = \gamma_0 + \gamma_1\beta_1 + \gamma_2\beta_2 + \gamma_3\beta_3 + \gamma_4\beta_4 + \epsilon_i \qquad (3.4)$$

$_{1\ \ 4}$ are the factors scores and ε_i is the error term.

$$R_i = \alpha R_{APT} + (1-\alpha) R_{CAPM} + e_i \qquad (3.5)$$

$_{i=}$ the average monthly excess returns of the stock i, $R_{APT=\ CAPM=}$ expected excess returns estimated by CAPM and α measure the effectiveness of the models. The APT is the accurate model to forecast the returns of the stocks as compare to CAPMif α is close to 1.

$$R = [ESS_0/ESS_1]^{N/2} N^{K_0 - K_1/2} \qquad (3.6)$$

$_{0\ 1\ 0\ 1}$

IV. RESULTS AND DISCUSSION

4.1 Model Performance Comparison

Table 1: Model Performance on NSL-KDD Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)	Detection Speed
Random Forest	98.2	97.8	98.5	0.047	1.2	Real-time
Isolation Forest	-0.01	96.1	97.0	96.5	2.8	Real-time
Traditional Signature-Based IDS	76.0	75.2	77.1	76.1	18.5	Offline only
Decision Tree	93.4	92.8	94.0	93.4	5.3	Near Real-time
Naive Bayes	84.7	84.1	98.1	84.7	10.2	Near Real-time

Table 1 summarizes the performance of all evaluated models on the NSL-KDD test set. Random Forest achieved the highest overall accuracy of 98.2%, with a precision of 97.8%, recall of 98.5%, and F1-score of 98.1%, while maintaining a false positive rate of only 1.2%. Its real-time processing capability makes it the most suitable model for deployment in live network environments.
Isolation Forest, operating in unsupervised mode, achieved 96.5% accuracy and demonstrated a low false positive rate of 2.8%, making it highly effective for detecting zero-day attacks where no labeled training data exists for the threat type.
Traditional signature-based IDS achieved only 76% accuracy with an 18.5% false positive rate, confirming the limitations highlighted in the literature review. Decision Tree and Naive Bayes classifiers performed moderately, achieving 93.4% and 84.7% accuracy respectively, but neither matched the performance nor real-time capability of the proposed system. These results validate the superiority of ensemble-based and anomaly detection approaches for modern intrusion detection.

0

1 :

The SIEM-compatible log output produced by the system captures each detection event with full metadata including timestamp, source and destination IP, port, protocol, predicted attack label, and action taken. These logs were validated against representative intrusion scenarios and confirmed to accurately reflect both detected and blocked events, demonstrating the system's suitability for integration with enterprise SOC workflows.
The graphical user interface (GUI) developed for the system provides real-time visualization of detected threats, displaying attack type, severity, source, and timestamp in an intuitive dashboard. Security analysts can view live alerts and historical log data side by side, enabling rapid triage and response. The combination of high detection accuracy, real-time processing, and SIEM-integrated logging positions the proposed system as a significant advancement over conventional IDS solutions currently deployed in network security environments.

Conclusion and Future Work

This paper presented an AI-powered real-time Intrusion Detection System combining Random Forest and Isolation Forest machine learning models, trained and evaluated on the NSL-KDD benchmark dataset. The system successfully detects five major attack categories — Brute Force, DoS, Port Scanning, Malware Propagation, and Impersonation — with accuracies of 98.2% and 96.5% respectively. By integrating live traffic capture via Wireshark, real-time classification, a graphical interface, and SIEM-compatible logging, the system bridges the gap between academic machine learning research and practical cybersecurity deployment. The proposed system significantly outperforms traditional signature-based IDS across all evaluated metrics.**“ ”** in the text, and “ ”

Future work will explore the integration of deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, for sequential traffic pattern analysis. The authors also plan to expand the dataset to include newer attack types such as Advanced Persistent Threats (APTs) and encrypted traffic attacks, and to deploy the system as a containerized microservice compatible with enterprise-scale cloud environments. Integration with commercial SIEM platforms such as Splunk and the ELK Stack is also a planned direction for future development. ”,notjust “M”.

V. ACKNOWLEDGMENT

The authors sincerely thank Dr. Suvarna Pansambal (Project Guide), Dr. Shweta Sharma (Project Coordinator), and Prof. Mahendra Patil (Head of Department) at Atharva College of Engineering, Malad., for their invaluable guidance and support throughout this project. The authors also acknowledge the open-source communities behind scikit-learn, Wireshark, and the NSL-KDD dataset for making their resources freely available.

REFERENCES

[1] S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy," Technical Report, Chalmers University of Technology, 2000.
[2] V. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Computers & Security, vol. 28, no. 1-2, pp. 18-28, 2009.

[3] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.
[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. IEEE International Conference on Data Mining (ICDM), pp. 413-422, 2008.
[6] S. Revathi and A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection," International Journal of Engineering Research and Technology, vol. 2, no. 12, 2013.
[7] M. A. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, and H. Janicke, "RDTIDS: Rules and Decision Tree-Based Intrusion Detection System for Internet-of-Things Networks," Future Internet, vol. 12, no. 3, p. 44, 2020.
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.
[9] A. Divekar, M. Parekh, V. Savla, R. Mishra, and M. Shirole, "Benchmarking Datasets for Anomaly-Based Network Intrusion Detection: KDD CUP 99 Alternatives," in Proc. IEEE International Conference on Computer Communication and Informatics (ICCCI), 2018.
[10] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," in Proc. Network and Distributed Systems Security Symposium (NDSS), 2018.