Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Security Information and Event Management (SIEM) Open Source Solution

Anuja Chincholkar, Armaan Bansal, Samyak Jain, Shubham Shah
Computer Science Engineering (Cyber Security and Forensics)
MIT ADT University, Pune, India

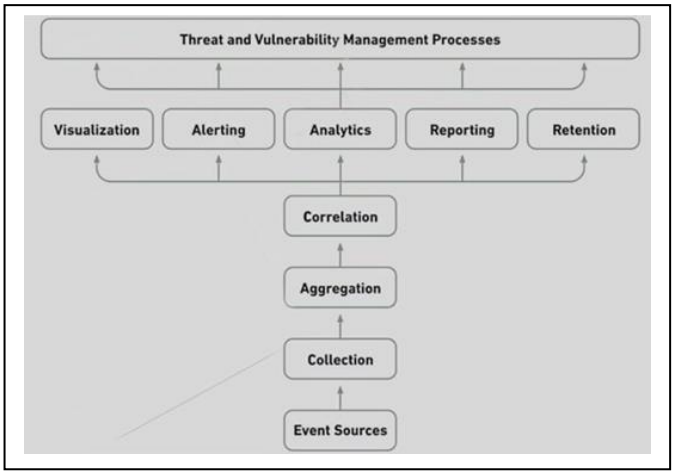
***Abstract*— In today's digital landscape, small and medium-sized businesses (SMBs) face increasing cyber threats but often lack the resources to deploy comprehensive security measures. This paper introduces a lightweight, scalable, and cost-effective SIEM toolkit tailored specifically for small B2B enterprises. The proposed solution simplifies deployment, manages security logs efficiently, and provides essential threat detection and incident response features. Leveraging built-in integrations, preconfigured rules, and intuitive dashboards, this toolkit aims to enhance security posture without requiring extensive technical expertise. By integrating automated alerts and responses, it empowers SMBs to quickly identify and mitigate emerging threats, ensuring compliance and operational continuity. This research demonstrates that a targeted, easy-to-use SIEM toolkit can significantly bolster the cybersecurity defenses of small businesses, enabling them to operate securely in an increasingly hostile environment.**

Keywords—template, Scribbr, IEEE, format

I. INTRODUCTION

In the rapidly evolving digital era, cybersecurity has emerged as a crucial concern for organizations of all sizes. Cyber threats have become increasingly sophisticated, frequent, and damaging, necessitating robust and proactive security measures. While large enterprises frequently deploy advanced Security Information and Event Management (SIEM) systems to monitor and respond to these threats, small and medium-sized businesses (SMBs), especially those operating in business-to-business (B2B) markets, face unique challenges in implementing adequate cybersecurity programs. These challenges stem primarily from limited financial resources, constrained IT personnel expertise, and the often prohibitive complexity and cost of commercial SIEM solutions. Despite their smaller scale, B2B SMBs remain attractive targets for cybercriminals due to their role as supply chain partners and custodians of critical data.

Security Information and Event Management (SIEM) is a cornerstone technology in modern cybersecurity frameworks, enabling centralized collection, normalization, correlation, and analysis of security event data from diverse IT assets. SIEM platforms provide real-time situational awareness by



aggregating logs and telemetry from network devices, servers, endpoints, applications, and increasingly cloud environments. This consolidated security intelligence supports rapid detection of cyberattacks, suspicious behaviors, policy violations, and compliance breaches. SIEM systems also facilitate incident response processes, forensic investigations, and audit reporting.

Historically, SIEM solutions were developed with the challenges of large, complex enterprises in mind, reflecting their extensive infrastructure and security team requirements. Consequently, these systems tend to be elaborate, expensive, and resource-intensive, demanding ongoing tuning of correlation rules, management of voluminous logging data, and significant manual analyst involvement. For SMBs, these attributes can render SIEM adoption impractical or inaccessible. The growing threat landscape, however, necessitates that small B2B businesses similarly achieve heightened security visibility to protect against ransomware, data leaks, insider threats, and advanced persistent threats (APTs).

There exists an urgent need to democratize SIEM capabilities, designing scaled, cost-effective toolkits that match the operational realities and security needs of small B2B businesses. A tailored SIEM toolkit must lower the barrier of entry by simplifying deployment, reducing the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

requirement for specialized cybersecurity expertise, and focusing on core use cases relevant to smaller enterprises. Essential features include seamless log collection from critical IT assets, effective normalization to handle diverse data types, lightweight correlation engines that identify relevant threat patterns, and user-friendly alerting mechanisms. Automation features and machine learning-driven anomaly detection can further enhance the toolkit's efficacy by reducing false positives and enabling swift incident response, even in organizations lacking dedicated security operation centers (SOCs).

Moreover, the modern technological environment introduces new complexities and opportunities for SMBs. Increasing adoption of cloud computing services, virtualization, and remote work arrangements expands the attack surface and complicates security monitoring. The SIEM toolkit must adapt to these trends, supporting hybrid on-premise and cloud environments, and providing comprehensive visibility regardless of infrastructure distribution. In parallel, evolving regulatory frameworks such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the Payment Card Industry Data Security Standard (PCI-DSS) impose stringent requirements on how businesses collect, store, and report security-related data. Compliance capabilities are thus integral to any practical SIEM solution for SMBs.

This paper proposes a novel SIEM toolkit designed specifically for small B2B enterprises, synthesizing simplicity, scalability, and affordability. By leveraging modular architecture and incorporating open-source components where suitable, the toolkit minimizes total cost of ownership while offering flexibility for customization. The design prioritizes streamlined workflows and intuitive user interfaces, enabling business owners and IT personnel with limited cybersecurity training to maintain robust defenses and meet audit obligations. The inclusion of automated threat intelligence feeds and integration with security orchestration, automation, and response (SOAR) platforms further extends the toolkit's capabilities, empowering small organizations to react promptly and effectively to emerging threats.

The methodology behind this research includes the systematic design of toolkit architecture, implementation of prototype components, and rigorous evaluation using realistic cyberattack scenarios and operational benchmarks. Performance metrics such as detection accuracy, false positive rates, processing latency, and resource consumption are analyzed to validate feasibility in constrained environments typical of small B2B businesses.

In summary, this work addresses a critical gap in cybersecurity tooling—a scalable, user-centric SIEM solution adaptable to the needs and capacities of small B2B enterprises. Our contributions include a detailed system architecture, implementation insights, and empirical results demonstrating the toolkit's practical benefits. By facilitating accessible, effective cybersecurity monitoring and response, this research aims to enhance the security posture of an often underserved but vital business segment.

II. LITERATURE REVIEW

A. Overview of SIEM Systems

Security Information and Event Management (SIEM) systems amalgamate two critical cybersecurity functions: Security Information Management (SIM) and Security Event Management (SEM). SIM focuses on the long-term storage, analysis, and reporting of log data, while SEM addresses real-time monitoring, correlation, and alerting of security events. The consolidation of these functions into SIEM platforms enables organizations to gain comprehensive situational awareness, improve threat detection efficiency, and facilitate compliance with regulatory mandates.

Over the past decade, SIEM technology has evolved from basic log management tools into sophisticated cybersecurity platforms incorporating advanced analytics, machine learning, and automation capabilities. Early SIEM solutions relied primarily on manual rule creation and static correlation methods, which limited their ability to identify complex attack patterns or adapt to new threat vectors. Recent advancements have integrated behavioral analytics, anomaly detection, and threat intelligence feeds, enabling dynamic and context-aware incident detection.

- SIEM Architecture and Core Components

SIEM architecture typically comprises several core components: data sources, data collection and normalization, correlation engine, alerting and visualization, data storage, and integration with other security tools.

- Data Sources: These include logs from firewalls, intrusion detection/prevention systems (IDS/IPS), servers, endpoints, cloud services, and applications. The breadth and diversity of data inputs substantially affect SIEM efficacy.
- Data Collection and Normalization: Efficient SIEM solutions implement secure, scalable mechanisms to collect and standardize logs from disparate sources into consistent formats for analytical processing. Normalization involves parsing raw logs, extracting key attributes, and mapping heterogeneous data into a unified schema.
- Correlation Engine: This component applies rules, algorithms, or machine learning models to detect suspicious patterns or anomalies across aggregated data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Classical correlation methods involve Boolean logic and state machine detection; modern engines leverage statistical and AI-driven approaches to discern subtle attack indicators.

- Alerting and Visualization: Real-time dashboards, reports, and alert notifications empower security analysts to monitor ongoing threats and prioritize response actions. Usability of these interfaces significantly impacts SOC efficiency.
- Data Storage: Persistent storage supports forensic investigations, compliance audits, and threat hunting. Scalable storage architectures, including cloud-based and elastic solutions, address growing data volumes.
- Integration and Automation: SIEMs increasingly integrate with Security Orchestration, Automation, and Response (SOAR) systems, threat intelligence platforms, and vulnerability scanners to automate workflows and enrich detection accuracy.

B. Market Landscape and SIEM Vendors

- The SIEM market exhibits a dynamic vendor landscape, ranging from established commercial providers to emerging open-source projects. Leading commercial platforms include IBM QRadar, Splunk, Micro Focus ArcSight, LogRhythm, and McAfee Enterprise Security Manager. These products offer comprehensive feature sets, scalability, and enterprise support but often entail high acquisition and operational costs.
- Open-source SIEM tools such as OSSIM, Wazuh, and Graylog gained traction for smaller organizations or research purposes due to their flexibility and lower cost. However, they typically require more technical expertise and lack advanced features inherent in commercial offerings.
- Technology analysts classify SIEM vendors into categories such as leaders, challengers, visionaries, and niche players based on innovation, market presence, and usability criteria. User reviews emphasize factors like ease of deployment, detection capabilities, scalability, and vendor support as critical considerations when selecting SIEM solutions.

C. Strengths of SIEM Systems

- SIEM systems provide several key advantages in organizational cybersecurity posture:
- Comprehensive Log Aggregation: Centralization of logs ensures a single source of truth aiding effective security monitoring.
- Real-Time Threat Detection: Correlation engines enable early identification of known attack signatures and anomalies.
- Compliance Support: SIEMs facilitate adherence to regulations such as GDPR, HIPAA, PCI-DSS by maintaining audit trails and generating compliance reports.
- Forensics and Incident Response: By storing historical logs, SIEMs enable investigation post-breach, supporting root cause analysis and remediation.
- Challenges and Limitations
- Despite their strengths, SIEM systems face several persistent challenges limiting optimal effectiveness:
- Incomplete Data Coverage: Many SIEMs lack comprehensive access to certain data types like endpoint telemetry, DNS logs, or cloud application events, creating blind spots.
- Complexity and Cost: High deployment complexity, configuration overhead, and license fees pose obstacles, particularly for SMBs.
- False Positive Rates: Static correlation rules often generate excessive false alarms, burdening security teams. Advanced machine learning integration is still maturing.
- Scalability Issues: Handling the exponential growth in log volumes strains storage and processing resources.
- Manual Response Dependence: Despite SOAR advancements, many environments still rely heavily on manual analyst intervention.
- Advances in SIEM Technology
- The cybersecurity community and vendors continue research and development to address SIEM limitations:
- Machine Learning & AI: Adoption of unsupervised learning, behavioral analytics, and anomaly detection reduces false positives and adapts to novel threats.
- Big Data Analytics: Utilizing distributed processing frameworks like Apache Hadoop and Spark enables handling vast log datasets efficiently.
- Cloud-Native SIEM: Solutions optimized for cloud infrastructure provide elasticity and native integration with cloud services and workloads.
- Enhanced Visualization: Interactive, real-time dashboards with context-aware visual analytics improve analyst situational awareness.
- Integrated SOAR: Automated playbooks and incident response accelerate mitigation workflows.
- Privacy-Preserving Monitoring: Techniques ensuring compliance with privacy laws while collecting necessary security data are emerging.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Table 1: General information of SIEM software tools.

s.no	SIEM Tools	On-prem	Cloud SIEM	OS Platform	Pricing
01	Splunk	Y	Y	Windows, mac, Linux, Solaris	Quote based
02	Sumo logic	Y	Y	Windows, Linux, mac	$99 to $165/month
03	Manage Engine log 360	Y	-	Windows, website, mac	Quote based
04	Alien vault OTX	Y	Y	Windows, Linux, mac	Open source(on-prem)
05	IBM Qradar	Y	Y	Windows, mac, Linux	Quote based
06	Varonis	Y	Y	Windows, Linux, mac	$17,000/license
07	LogRhythm	Y	Y	Windows, CentOS, Linux	Quote based
08	Rapid 7 insight	-	Y	Windows, Linux, website	$19 to $2000
09	Solar winds	Y	Y	Windows, Linux, mac, Solaris, website	$4665
10	Salesforce	-	Y	Windows, mac, Linux, android, iOS	$25/user/month

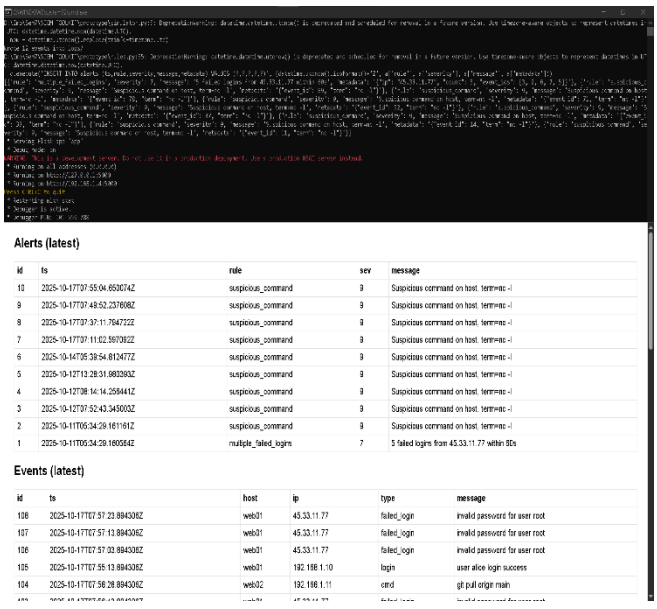
D. SIEM in Small and Medium-Sized Businesses

A major focus area is adapting SIEM technology to SMBs, especially small B2B companies. Research emphasizes the need for SIEMs that balance feature richness with accessibility, affordability, and simplicity. Various lightweight SIEM solutions and managed SIEM services have been proposed to cater to this segment.

Recent studies explore modular architectures, cloud-based SaaS SIEMs, and preconfigured rule sets tailored for smaller infrastructures. Effective threat detection with minimal configuration overhead and support for compliance are noted as critical success factors.

III. METHODOLOGY

This section serves as the blueprint for our investigative journey, outlining the systematic approach employed to achieve our research objectives. In this section, we detail the carefully crafted steps, techniques, and tools used to gather, process, and analyze data. It provides a clear roadmap for how we explore the research questions, test hypotheses, and derive meaningful insights. Our methodology encompasses data acquisition, preprocessing, model selection, and evaluation procedures, ensuring rigor and reproducibility in our study's findings.

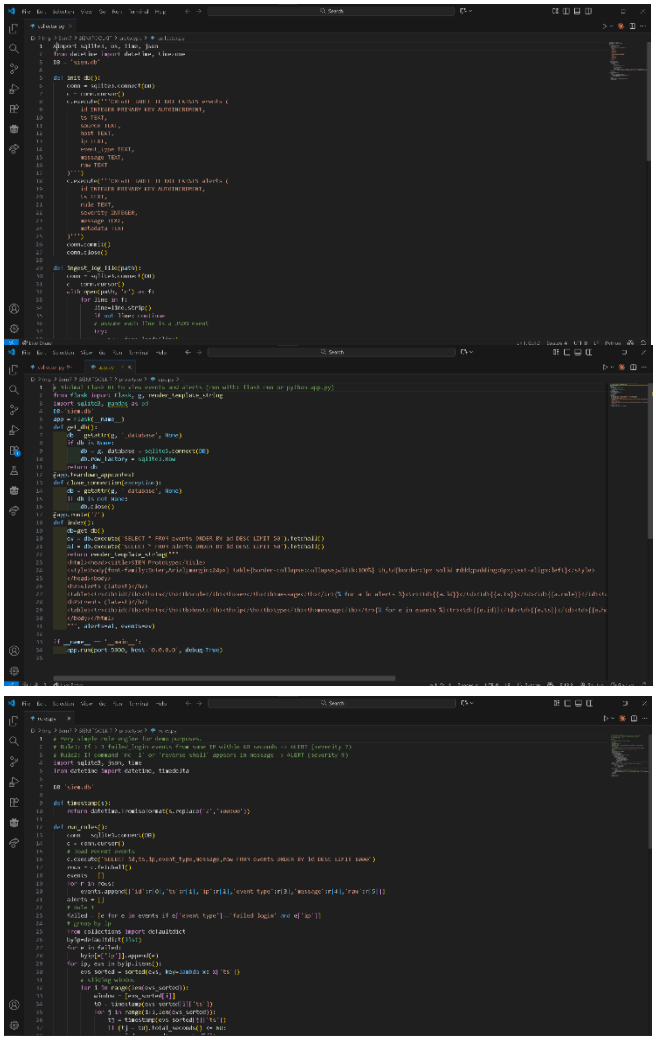


Alerts (latest)

id	ts	rule	sev	message
10	2025-10-17T07:55:04.650307Z	suspicious_command	9	Suspicious command on host. terminal
9	2025-10-17T07:49:52.237608Z	suspicious_command	9	Suspicious command on host. terminal
8	2025-10-17T07:37:11.764722Z	suspicious_command	9	Suspicious command on host. terminal
7	2025-10-17T07:11:02.597092Z	suspicious_command	9	Suspicious command on host. terminal
6	2025-10-14T05:38:54.812477Z	suspicious_command	9	Suspicious command on host. terminal
5	2025-10-12T13:28:31.983239Z	suspicious_command	9	Suspicious command on host. terminal
4	2025-10-12T08:14:14.256441Z	suspicious_command	9	Suspicious command on host. terminal
3	2025-10-12T07:52:43.345003Z	suspicious_command	9	Suspicious command on host. terminal
2	2025-10-11T05:34:29.161161Z	suspicious_command	9	Suspicious command on host. terminal
1	2025-10-11T05:34:29.160584Z	multiple_failed_logins	7	5 failed logins from 45.33.11.77 within 80s

Events (latest)

id	ts	host	ip	type	message
108	2025-10-17T07:57:23.894306Z	web01	45.33.11.77	failed_login	invalid password for user root
107	2025-10-17T07:57:13.894306Z	web01	45.33.11.77	failed_login	invalid password for user root
106	2025-10-17T07:57:03.894306Z	web01	45.33.11.77	failed_login	invalid password for user root
105	2025-10-17T07:55:13.894306Z	web01	192.168.1.10	login	user alice login success
104	2025-10-17T07:56:28.894306Z	web02	192.168.1.11	cmd	git pull origin main



3.1. System overview

The current project system consists of two primary models: text information clustering based on TF-IDF and K-means, and RFC intrusion prediction. This system incorporates the powerful techniques of TF-IDF and K-means clustering as integral components for the initial phase of threat detection. TF-IDF enables us to effectively process and represent textual security logs, while K-means clustering helps in grouping these logs into meaningful clusters based on their semantic similarities.

First of all, network data is collected from a test virtual machine on which a SIEM system is installed to monitor traffic and attempts of brute force attacks. Then, this data is preprocessed to locate null or zero values. It is combined with the CICIDS2017 dataset of web attacks and brute force history penetrations. Next, brute-force textual information is clustered into two main clusters: information technologies-related words and incoherent data. At this stage, visualization of additional information is demonstrated in 3D, 2D, or map-like plots (Fig. 1). Feature importance findings were

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

implemented, focusing on ten main features. Finally, the model is tuned using hyperparameter optimization, and the results of the trials determine how well the model applies to a test dataset.

![Fig. 1: Overview of Model Steps](begins with several virtual machines based on Microsoft Azure, serving as the foundation for data collection. Security events and attempts of brute force attacks are monitored and logged within the Microsoft Azure Sentinel SIEM. The data collected, which includes both attempts of brute force attacks data and Wireshark network traffic data combined with the CICIDS2017 dataset, undergoes a crucial phase of analysis and preprocessing denoted as "EDA Analysis and datasets preprocessing." During this stage, the data is cleansed, transformed, and prepared for subsequent analysis.

The "TF-IDF vectorization for textual data" step involves converting textual data into numerical vectors, making it suitable for machine learning. This vectorization process enables the system to work effectively with textual features. Moving forward, in the models section and hyperparameter tuning, various machine learning models are trained and fine-tuned to optimize their performance. This step ensures the models are well prepared to handle the incoming data.

The "K-means clustering and PCA/t-SNE visualization" phase employs clustering techniques for data segmentation and dimensionality reduction techniques for visual representation, providing a deeper understanding of data patterns and relationships.

Finally, the RFC model emerges as the top-performing model, demonstrating its capability to effectively classify and interpret the data, making it a valuable asset in the system's overall cybersecurity strategy.

3.2. Hosting VM and data acquisition

For this experiment, several test virtual machines, such as honeypot devices, were prepared. A honeypot is a cybersecurity mechanism designed to deceive and lure attackers into a controlled environment, where their actions and tactics can be monitored, analyzed, and studied without putting real systems or data at risk. Honeypots are important tools in cybersecurity for understanding and defending against malicious activities such as hacking, malware, and intrusion attempts.

Additionally, honeypots are intentionally set up to appear as attractive targets to cyber attackers by mimicking legitimate systems, applications, or services. They lure malicious actors into interacting with them, providing critical insights into attacker tactics and techniques. This information helps organizations improve their security posture by allowing them to better understand threats, vulnerabilities, and potential attack vectors. However, deploying and managing honeypots require careful planning and expertise to ensure they are effective and do not inadvertently create security risks.

Data acquisition involves collecting and gathering data from various sources and making it available for analysis or storage. This process was used in our study to collect data from the honeypots, the internet, and virtual machines. Identifying and accessing relevant data sources is the first step. Once identified, data collection methods are employed to retrieve data. Data is often noisy, incomplete, or inconsistent, so data preprocessing is performed to clean, transform, and format the data into a usable and structured format, including data cleaning, missing value imputation, and normalization (see section 3.3).

In many cases, data from multiple sources needs to be integrated into a single dataset. In this study, the CICIDS2017 dataset was merged with the collected data to form a comprehensive dataset. An attractive target for intruders wanting unauthorized access was created. Honeypots simultaneously collected internet traffic data into a single dataset, allowing qualitative analysis and model compilation based on attack types, methods, and tools.

A group of virtual machines was deployed on Microsoft Azure hosting, which provides a free 30-day trial for all Microsoft account holders and offers a wide range of SIEM application services. The Azure logical group included:

Test virtual machines based on Windows 10 operating system

Azure Sentinel SIEM system for collecting internet traffic data

A workspace for visualizing temporarily incoming information from the system logs

A self-written script was run on virtual machines to save the history of login attempts and associated actions to check for brute force attacks. The collected information was converted into two new independent datasets: The first combined CICIDS2017 (Canadian Intrusion Detection Evaluation Dataset of Thursday attacks) with Wireshark traffic data from VMs to create an intrusion detection model. The second dataset was directly delivered from Azure SIEM for clustering analysis.

3.3. Datasets preprocessing

The Azure SIEM dataset consists of various columns such as destinationhost_CF and sourcehost_CF (representing the name of the VM), label_CF (IP addresses of intruders), county_CF (country of intruders), state_CF (state or region of intruders), latitude_CF and longitude_CF (geographical coordinates), timestamp_CF (timestamp of intrusion), and username_CF (usernames that intruders used for brute forcing).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Columns destinationhost_CF and sourcehost_CF were deemed irrelevant or less valuable and dropped from the data frame. Due to 3069 empty lines in the state_CF column, missing data was replaced with "N/A". Proper handling of missing data was crucial to avoid biased or inaccurate results. The date format was converted from ISO 8601 to datetime format to facilitate easier manipulation and analysis. ISO 8601 is a standard date and time representation widely used in Microsoft Azure and data interchange formats.

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic widely used in natural language processing and information retrieval to evaluate the importance of a term within a document relative to a collection of documents or corpus. TF-IDF was applied in our model to process textual security-related data.

For combined Wireshark and CICIDS2017 log files, blank records were removed, and string characters were converted to numerical features using one-hot encoding and TF-IDF. Both techniques were used in the machine learning pipeline due to the presence of categorical and textual features. After preparing this combined feature set, the machine learning model was trained.

3.4. Feature selection

Decision tree visualization and single-tree evaluation of feature significance using DecisionTreeClassifier were utilized to simplify classifier presentation. Feature selection played a key role in providing the classifier with the most relevant features while discarding redundant ones.

Several methods were employed for feature importance evaluation:

Recursive Feature Elimination (RFE) , which recursively fits the classifier and ranks features by importance, removing the least significant until the desired feature subset size is reached

Ensemble methods like Random Forest or Gradient Boosting to assess feature relevance more robustly

Correlation analysis to identify and remove or combine highly correlated features to reduce multicollinearity

Cross-validation methods were used to verify stability and consistency of feature selection. Careful attention was given to avoid overfitting, especially with small model parameters such as five leaves.

3.5. Model selection and performance tuning

Models were evaluated based on accuracy, precision, recall, and F1-measure.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2 displays results of ten classifiers including KNN, SVM, CART, Random Forest Classifier (RFC), AdaBoost, Logistic Regression (LR), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Multi-layer Perceptron (MLP). Models like KNN, CART, RFC, AdaBoost, and LR exhibited the best performance. RFC was selected considering its high accuracy and shorter execution time.

Hyperparameter tuning was done using Grid-Search Cross-Validation (Grid-SearchCV) to systematically search for optimal model parameters, optimizing performance metrics.

For clustering, Silhouette and Elbow methods were explored to find the optimal number of clusters, k. Silhouette scores indicate how well data points fit within their clusters, while Elbow plots examine decrease in within-cluster sum of squares (WCSS). In our case, manual inspection found $k = 2$ as optimal to meaningful separate useful from useless data (see Fig. 2).

IV. RESULTS AND DISCUSSION

Results of our study showed that the feature importance algorithm based on the Decision Tree Classifier identified the most valuable features in the combined CICIDS2017 dataset. After eliminating the less significant features, the feature space was narrowed down to ten key features, including "Average Packet Size" (the average length of TCP or IP packet data), "Flow Bytes/s" (data flow rate), and "Max Packet Length" (maximum packet length), among others. The outcomes of the feature selection process are presented in Fig. 3.

Reducing the feature space to ten features was motivated by multiple considerations. Primarily, it helps address the "curse of dimensionality," whereby high-dimensional data demand larger datasets for effective modeling and lead to increased computational complexity and potential overfitting. By selecting a carefully considered subset of relevant features, we simplified the problem while significantly improving model performance. Feature selection enables the construction of more efficient, interpretable models by discarding irrelevant or redundant information. The choice to limit features to ten was supported by domain knowledge, feature importance analysis, and experimental validation.

Additionally, reducing the feature space facilitated better data visualization, making model outputs more understandable and manageable. Smaller feature sets also mitigate overfitting risks, thereby improving the model's ability to generalize well to unseen data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

Our Random Forest model hyperparameters—number of trees in the forest (n_estimators), the minimum number of objects per tree leaf (min_samples_leaf), maximum tree depth (max_depth), and maximum number of features per tree (max_features)—were optimized using Grid-Search Cross-Validation (Grid-SearchCV), as shown in Fig. 4. The F1-measure was used as the target metric for parameter optimization.

The resulting RandomForestClassifier model exhibited strong performance on the test dataset, achieving a recall score of 0.96 and an F1 score of 0.97. These metrics indicate the model's high effectiveness at correctly identifying true positive instances with a minimal number of false negatives. The high recall reflects a low rate of missed relevant cases, critical in sensitive applications such as medical diagnosis or fraud detection. Moreover, the excellent F1 score signifies a good balance between precision and recall, demonstrating the model's accuracy while minimizing false positives. Overall, the evaluation strongly suggests that the Random Forest classifier is well-suited for the classification task.

For textual data, K-means clustering applied after TF-IDF vectorization yielded positive results, identifying two principal clusters. These clusters divided information into IT-related data and irrelevant ("useless") data. The IT-related cluster contained terms such as "administrator," "VM," "account," and "computer," reflecting typical device names assigned by network users. Such findings illustrate the necessity of fostering personal information security literacy, including avoiding predictable or familiar device naming conventions that could aid attackers. An example visualization of the two primary clusters is displayed in Fig. 5.

V. CONCLUSION

This research paper delved into the dynamic intersection of SIEM systems and machine learning algorithms, highlighting the transformative potential of their synergy in cybersecurity. The integration of ML-driven techniques has ushered in a new era of proactive, adaptive security.

Machine learning algorithms, capable of processing vast and diverse datasets with near-human precision, have significantly enhanced SIEM systems' ability to detect, analyze, and respond to threats. These models identify subtle anomalies, predict emergent risks, and automate incident response tasks, fundamentally altering cybersecurity defenses.

Our study revealed the versatility of ML-driven SIEM approaches across domains such as network intrusion detection, user behavior analytics, and anomaly detection. This adaptability allows organizations to tailor cybersecurity strategies according to specific vulnerabilities, yielding unprecedented customization and accuracy.

However, successful implementation demands robust data management, careful model selection, and ongoing monitoring. The human-machine partnership remains essential, with human expertise complementing automated algorithms to ensure effective cybersecurity outcomes.

The promising results of this study affirm machine learning's critical role in advancing SIEM systems and securing digital environments against evolving cyber threats.

CRediT authorship contribution statement Armaan Bansal: Writing – original draft, Software, Investigation. SamyAK JAIN: Writing – original draft, Software, Methodology.Shubham Shah: Writing – review & editing, Validation, Methodology.

Acknowledgement This work has been researched and conducted under the Guidance of Prof. ANUJA CHINCHOLKAR under MIT ADT University

REFERENCES

[1] R. S and M. Mohankumar, “Security Information and Event Management (SIEM) Performance in on-Premises and Cloud Based SIEM: A Survey,” INSTICC, Jun. 2024, pp. 627–633. doi: 10.5220/0012613800003739.

[2] M. Sheeraz, M. H. Durad, M. A. Paracha, S. M. Mohsin, S. N. Kazmi, and C. Maple, “Revolutionizing SIEM Security: An Innovative Correlation Engine Design for Multi-Layered Attack Detection,” Sensors, vol. 24, no. 15, Aug. 2024, doi: 10.3390/s24154901.

[3] A. Shukla, P. A. Gandhi, Y. Elovici, and A. Shabtai, “RuleGenie: SIEM Detection Rule Set Optimization,” May 2025, [Online]. Available: http://arxiv.org/abs/2505.06701

[4] N. Tendikov et al., “Security Information Event Management data acquisition and analysis methods with machine learning principles,” Results in Engineering, vol. 22, Jun. 2024, doi: 10.1016/j.rineng.2024.102254.

[5] [5]P. Radoglou-Grammatikis et al., “SPEAR SIEM: A Security Information and Event Management system for the Smart Grid,” Computer Networks, vol. 193, Jul. 2021, doi: 10.1016/j.comnet.2021.108008.

[6] P. Artioli, V. Dentamaro, S. Galantucci, A. Magrì, G. Pellegrini, and G. Semeraro, “SIEVE: Generating a cybersecurity log dataset collection for SIEM event classification,” Computer Networks, vol. 266, Jul. 2025, doi: 10.1016/j.comnet.2025.111330.

[7] M. Sheeraz et al., “Effective Security Monitoring Using Efficient SIEM Architecture,” Human-centric Computing and Information Sciences, vol. 13, 2023, doi: 10.22967/HCIS.2023.13.023.

[8] P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, “Securing the Smart Grid: A Comprehensive Compilation of Intrusion Detection and Prevention Systems,” IEEE Access, vol. 7, pp. 46595–46620, 2019, doi: 10.1109/ACCESS.2019.2909807.

[9] T. S. Ustun and S. M. S. Hussain, “IEC 62351-4 Security Implementations for IEC 61850 MMS Messages,” IEEE Access, vol. 8, pp. 123979–123985, 2020, doi: 10.1109/ACCESS.2020.3001926.

[10] J. Manzoor, A. Waleed, A. F. Jamali, and A. Masood, “Cybersecurity on a budget: Evaluating security and performance of open-source SIEM solutions for SMEs,” PLoS ONE, vol. 19, no. 3 March, Mar. 2024, doi: 10.1371/journal.pone.0301183.

[11] N. Mohamed, “Current trends in AI and ML for cybersecurity: A state-of-the-art survey,” Cogent Engineering, vol. 10, no. 2, 2023, doi: 10.1080/23311916.2023.2272358.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11, November 2025

[12] I. Security Agency, "Cloud Security Technical Reference Architecture Federal Risk and Authorization Management Program," 2022.

[13] R. Nazir et al., "A review on machine learning techniques for network security," Journal of Cyber Security Technology. Taylor and Francis Ltd., 2025. doi: 10.1080/23742917.2025.2480730.

[14] S. Kwon, H. Yoo, and T. Shon, "IEEE 1815.1-Based power system security with bidirectional RNN-Based network anomalous attack detection for cyber-physical system," IEEE Access, vol. 8, pp. 77572–77586, 2020, doi: 10.1109/ACCESS.2020.2989770.

[15] R. Daszczyszak, D. Ellis, S. Luke, and S. Whitley, "Sponsor: USCYBERCOM TTP-Based Hunting," 2019.

[16] A. R. Muhammad, P. Sukarno, and A. A. Wardana, "Integrated Security Information and Event Management (SIEM) with Intrusion Detection System (IDS) for Live Analysis based on Machine Learning," in Procedia Computer Science, Elsevier B.V., 2022, pp. 1406–1415. doi: 10.1016/j.procs.2022.12.339.

[17] P. N. Wudali, M. Kravchik, E. Malul, P. A. Gandhi, Y. Elovici, and A. Shabtai, "Rule-ATT&CK Mapper (RAM): Mapping SIEM Rules to TTPs Using LLMs," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.02337

[18] V. Mavroeidis and A. Jøsang, "Data-driven threat hunting using sysmon," in ACM International Conference Proceeding Series, Association for Computing Machinery, Mar. 2018, pp. 82–88. doi: 10.1145/3199478.3199490.

[19] C. L. Martin, "Approved By," 2025.

[20] S. Iglesias Perez and R. Criado, "Increasing the Effectiveness of Network Intrusion Detection Systems (NIDSs) by Using Multiplex Networks and Visibility Graphs," Mathematics, vol. 11, no. 1, Jan. 2023, doi: 10.3390/math11010107.

[21] A. Barabanov and D. Makrushin, "SECURITY AUDIT LOGGING IN MICROSERVICE-BASED SYSTEMS: SURVEY OF ARCHITECTURE PATTERNS." [Online]. Available: https://microservices.io/patterns/deployment/sidecar.html

[22] S. Akhtar, S. Khan, and S. Parkinson, "LLM-based event log analysis techniques: A survey," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.00677

[23] W. Feng, S. Wu, X. Li, and K. Kunkle, "A Deep Belief Network Based Machine Learning System for Risky Host Detection."

[24] A. Virkud, M. A. Inam, A. Riddle, J. Liu, G. Wang, and A. Bates, How does Endpoint Detection use the MITRE ATT&CK Framework? [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/virkud

[25] A. Aldhaheri, F. Alwahedi, M. A. Ferrag, and A. Battah, "Deep learning for cyber threat detection in IoT networks: A review," Internet of Things and Cyber-Physical Systems, vol. 4. KeAi Communications Co., pp. 110–128, Jan. 01, 2024. doi: 10.1016/j.iotcps.2023.09.003.

[26] A. Rocha, F. Ayotunde Alaba, E. J. Ahmed, and A. Ibrahim, "Cybersecurity Model For Intelligent Cloud Computing Systems." [Online]. Available: https://ssrn.com/abstract=4970422

[27] Cisa, "Best Practices for MITRE ATT&CK ® Mapping CHANGE RECORD," 2021.

[28] K. Scarfone and M. Souppaya, "Cybersecurity Log Management Planning Guide," Oct. 2023. doi: 10.6028/NIST.SP.800-92r1.ipd.

[29] M. Song, "A Comprehensive Study of Security Information and Event Management (SIEM) Systems: Architectures, Benefits, and Challenges," 2024. [Online]. Available: https://www.researchgate.net/publication/387948975

[30] C. Panggabean, C. Venkatachalam, P. Shah, S. John, P. Renuka Devi, and S. Venkatachalam, "Intelligent DoS and DDoS Detection: A Hybrid GRU-NTM Approach to Network Security," in Proceedings of the 5th International Conference on Smart Electronics and Communication, ICOSEC 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 658–665. doi: 10.1109/ICOSEC61587.2024.10722438.

[31] "The Path to Choosing a SIEM System-A Systematic Literature Review."

[32] N. Jeffrey, Q. Tan, and J. R. Villar, "A Review of Anomaly Detection Strategies to Detect Threats to Cyber-Physical Systems," Electronics (Switzerland), vol. 12, no. 15, Aug. 2023, doi: 10.3390/electronics12153283.

[33] V. v Gaonkar, "Enhancing Threat Detection: Integrating ELK-Based SIEM with IDS and Pattern Recognition Algorithms," International Journal of Advanced Research in Computer and Communication Engineering Impact Factor, vol. 8, no. 11, 2024, doi: 10.17148/IJARCCE.2024.131127.

[34] M. Wurzenberger, F. Skopik, R. Fiedler, and W. Kastner, "Applying High-Performance Bioinformatics Tools for Outlier Detection in Log Data."

[35] S. A. Hussein and S. R. Répás, "Anomaly Detection in Log Files Based on Machine Learning Techniques," 2024.

[36] H. Maosa, K. Ouazzane, and M. C. Ghanem, "A Hierarchical Security Event Correlation Model for Real-Time Threat Detection and Response," 2024.

[37] M. A. Islam, "Application of artificial intelligence and machine learning in a security operations center," Issues in Information Systems, vol. 24, no. 4, pp. 311–327, 2023, doi: 10.48009/4_iis_2023_124.

[38] M. Vielberth, "Security Information and Event Management (SIEM)," in Encyclopedia of Cryptography, Security and Privacy, Springer Berlin Heidelberg, 2021, pp. 1–3. doi: 10.1007/978-3-642-27739-9_1681-1.