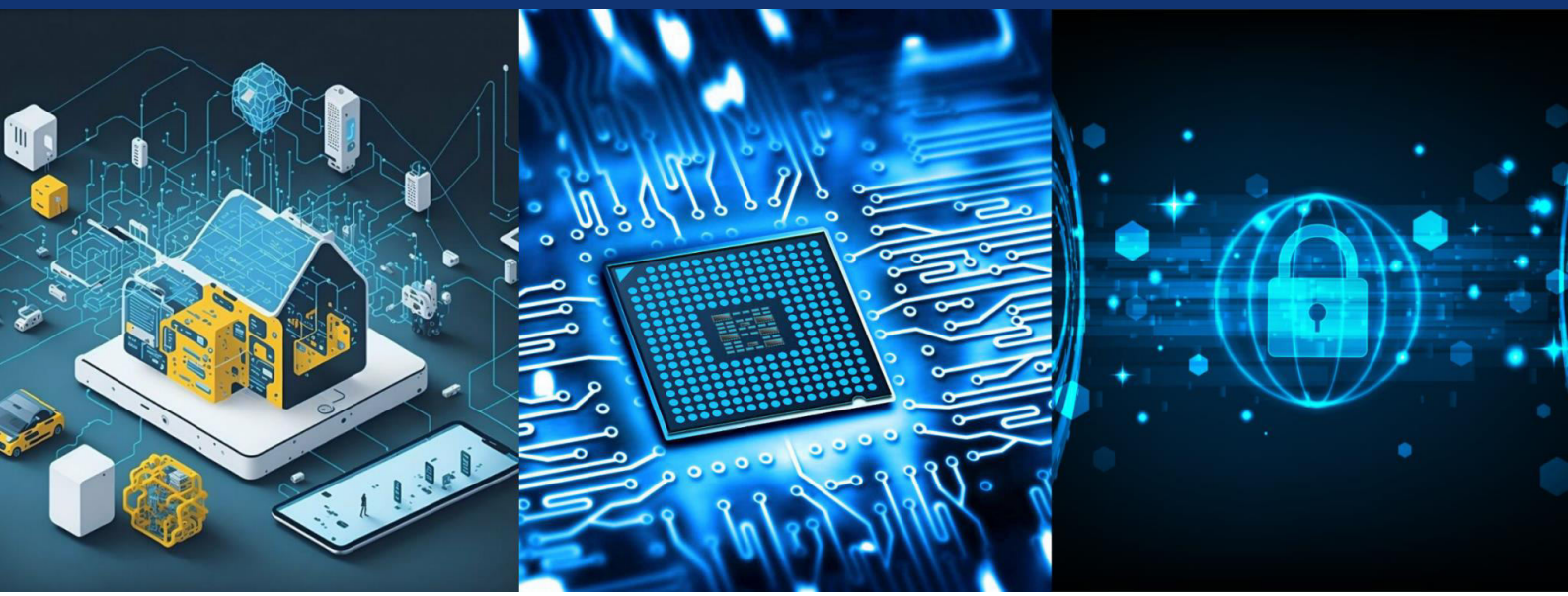
IJIRCCE

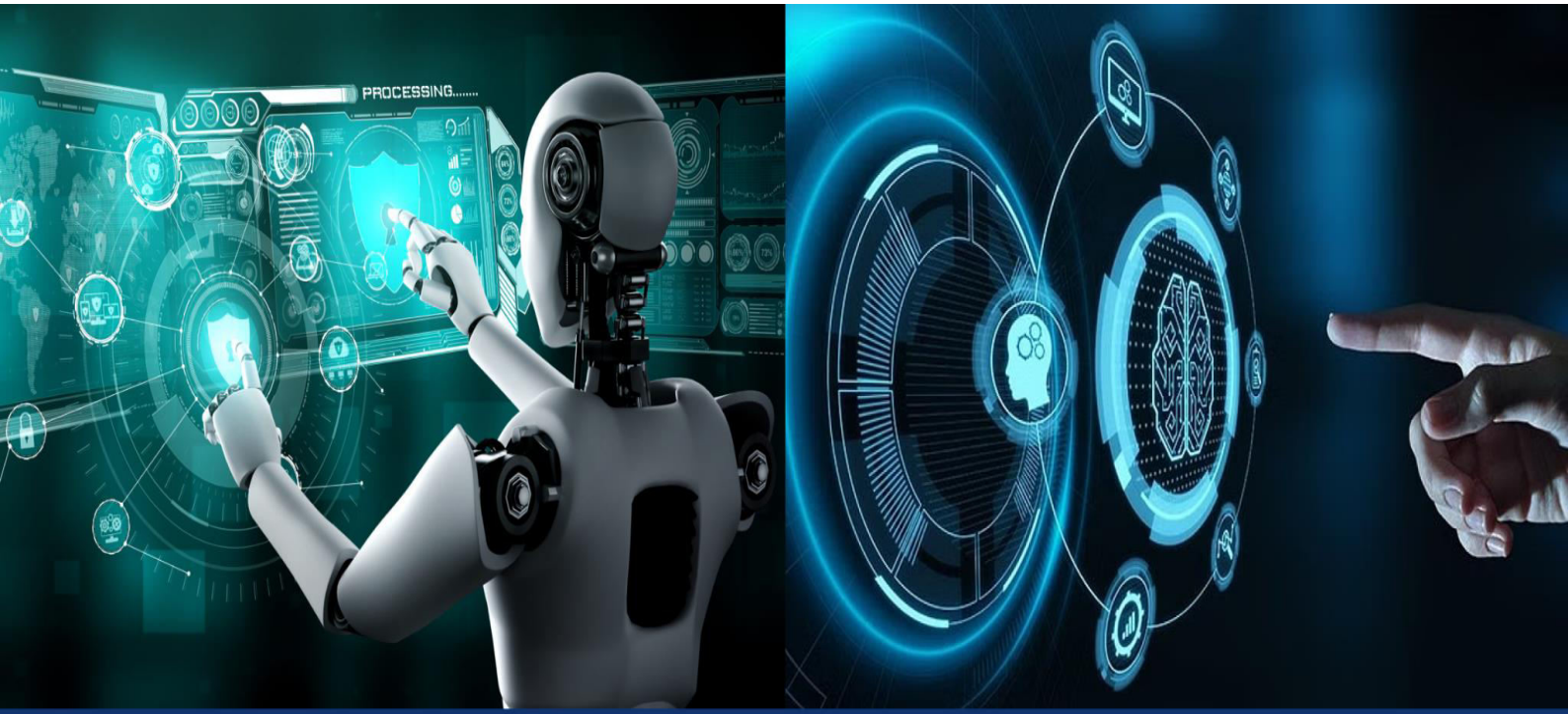
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 14, Issue 4, April 2026

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404034

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

LogSentinel: An Automated Hybrid System for Systematic Log Classification and Threat Identification

Jay Panchal, Sweety Patel

UG Student, Dept. of C.S.E, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

Assistant Professor, Dept. of C.S.E, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

ABSTRACT: Modern computing systems rely on logs to monitoring and security. Due to an increase in volumes of logs and an increased level of complexity, logs need to be analyzed manually which will lead to inefficiencies and errors. Rule-based methods are difficult to apply to unstructured and dynamic log data. Therefore, the purpose of this research paper is to develop LogSentinel as a hybrid system that uses both machine learning and natural language processing (NLP) to automatically classify log entries and detect potential cyber threats. LogSentinel has two primary classifications of log data. First, it classifies log entries quickly using regular expression to find specific patterns. Second, LogSentinel identifies patterns in semi-structured log data with machine learning algorithms. In addition to identifying patterns in log data, LogSentinel applies NLP methods to interpret ambiguity or complexities in the content of log entries. Lastly, LogSentinel detects brute force attacks through identifying repeat failures in login attempts. The input into LogSentinel is log data that is formatted as comma separated value (CSV), and the output from LogSentinel is formatted as structured Excel reports that allow users to easily interpret the results of their analysis. LogSentinel provides several benefits compared to traditional log analysis methods. These include improved accuracy of classification, better flexibility in handling multiple types of log formats and reduced labor associated with manual analysis. Furthermore, the authors demonstrate the effectiveness of LogSentinel's performance in applying its functionality to cybersecurity monitoring and automated system management applications.

KEYWORDS: Log Classification, Machine Learning, Natural Language Processing, Cybersecurity, Anomaly Detection, Log Analysis, Hybrid Model

I. INTRODUCTION

Log data plays a crucial role in monitoring, debugging, and securing modern digital systems from any kind of threat. As IT (Information Technology) infrastructure is expanding rapidly, applications and computing assets such as servers and network devices are also generating a substantial amount of log data. This growth makes it challenging, protracted, and susceptible to errors to process logs manually. Therefore, automated log analysis systems are now necessary for good system management and keeping an eye on the security of digital systems [3].

The conventional approaches to log analysis are mostly based on rule-based systems, such as predefined patterns and regular expressions. Though the above-mentioned methods are effective when it comes to analyzing structured logs, they may not be suitable for dealing with unstructured or dynamically created logs that come from contemporary systems. The inadequacies associated with the proposed approach can affect the process of log classification or anomaly detection. In order to overcome the problem, machine learning and deep learning methods are increasingly being applied for log classification and anomaly detection [10].

Recently, the advancement of large language models (LLMs) has significantly improved the techniques for classification of log data. By leveraging LLMs, log data can be interpreted with contextuality, thus making them more useful for dealing with complex and unstructured data [2][4]. Nevertheless, each approach has its own disadvantages. Systems that use regex are less flexible, machine-learning systems depend significantly on the quality and diversity of training data, and large language model-based methods might be computationally expensive and time-consuming [5].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This paper proposes LogSentinel, a new hybrid approach to classify log data that combines different methodologies. The system uses three-stage classification process: first is rule-based classification, second is machine learning models, and the last one is natural language processing. In our proposed approach, the Python regular expression (re) library is used to instantly categorize logs based on pre-defined rules, then the machine learning model is utilized for semi-structured log files that couldn't be classified by regex. Finally, natural language processing (NLP) techniques are applied to analyze complex and unstructured log messages.

This system takes the input in CSV format and generates output in an Excel file with a structured report for easy interpretation. This combination of various techniques and models in one solution allows LogSentinel to provide better classification quality, improved flexibility, and less labour-intensive processing of log files.

II. RELATED WORK

Currently, studies have been more focused on building tools that are able to spot suspicious or unusual patterns automatically. Many experts claim that log classification frameworks mostly require machine learning, deep learning, and natural language processing technical methodologies to operate effectively. Due to its inherent simplicity and interpretability, rule-based systems are accessible to lay users. For this reason, such approaches have traditionally been the predominant choice in the field. However, these methods are not effective when dealing with unstructured, noisy, or large-scale log data, limiting their applicability in modern systems [3]. Significant methodological evolutions have been observed in log-based anomaly detection as machine learning techniques have improved. Scientific investigations claim that Long Short-Term Memory networks, Transformer-based networks, and embedding-based approaches show strong effectiveness when these methods search for anomalies in logs [8][9]. Such technologies find complicated designs in the historical data of the logs because these technologies find anomalies with higher precision than previous working systems. The performance of these models is highly dependent on the quality and organization of input data. Well-structured and representative datasets significantly improve the accuracy and effectiveness of anomaly detection systems. The tool is only as effective as the data used to provide the result of the tool. Therefore, having all available data can help increase accuracy and effectiveness of results. If the data is not structured or organized, finding the necessary information for your purpose becomes difficult. The tools will not be very helpful. These tools need information to work well, and that information needs to be organized in a way that is easy to understand. Favorable circumstances for log analysis are produced by the arrival of large language models. Many experts believe that when a model like LogLLM is used, large language models demonstrate that they comprehend unstructured or noisy logs. Anomaly detection is performed by these models with a small requirement for preliminary arrangement of information while the models work [2][4].

Contextual understanding is used by these research methods so that the precision of classifying data becomes higher. Experts claim that a different technique was proposed which brings machine learning and LLM methods together [5]. Log parsing and preprocessing activities are recognized widely as being significant because these steps make the working position of analyzing logs more productive. It has been observed that the log parsing process significantly influences the effectiveness of an anomaly detection model [7]. Different techniques like rule-augmented filtering and log data purification are employed so that the quality of information improves while unwanted data pieces stay low [10]. This activity of preprocessing is necessary because the correctness of results in future analysis depends on this stage. Sophisticated systems like RAG and embedding approaches are also present which have made anomaly detection better than before [12][13].

Furthermore, researchers have proposed some unified frameworks to integrate various strategies into a single system. By doing so, these systems achieve greater extensibility and ensure that the solutions remain versatile enough for diverse real-world applications [8]. Researchers leverage graph-based unsupervised learning to detect anomalies within unstructured log data. By finding these deviations from already established behavioural patterns independent of past incidental data, this hybrid methodology effectively isolates irregularities in systems where predefined schemas are lacking. Research works have been finished recently to look into how log analysis can serve the safety of computers and the study of crimes [10]. Despite these big changes, several challenges remain in the world of log analysis. Existing tools usually pick only one way of working, which means these tools can only manage logs that match that specific style. Ability to grow, the cost of computer work, and the power to change are things that must be fixed so that log analysis works well. The integrated framework emerged as a response to the functional deficiencies found in current

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

systems. Rules for sorting, smart computer instructions, and natural language processing techniques are all used by the hybrid approach.

III. PROPOSED METHODOLOGY

A. *System Overview:*

The proposed LogSentinel System represents a hybrid approach to classify logs and detect anomalies. A primary function is to classify log messages while identifying irregularity, i.e., attempting unauthorized access to an account or repeatedly failing to log in to an account (Brute Force Attacks). Rather than using only one method to accomplish this task, the proposed system will combine pre-defined rules (regex) and natural language processing (NLP) with machine learning algorithms that have been trained on examples of data. By combining these methods into a single application, the proposed system should be capable of managing a variety of log formats; including well-formatted logs, poorly formatted logs, and unformatted logs. Each type of input may require a different contribution from each of the components used by the proposed system.

The system accepts log data as input from the user in CSV (Comma Separated Value) format and processes the data through several phases including clean-up of log data, classification of data, anomaly detection of data, and generation of reports. Blending techniques enhance how accurately, smoothly and rapidly logs can be interpreted instead of using one approach. While each phase has its own functionality, each phase builds upon outputs generated by previous phases in an unobtrusive manner.

B. *Methodology Description:*

The Proposed methodology consists of the following steps:

Step 1: Input and Preprocessing

This system takes the log data in CSV file format as input. Before processing the data for classification purposes, it is first pre-processed to eliminate any noise and to normalize the format of the log.

Step 2: Rule-Based Classification (Regex)

At the start of the process, a preliminary regex processing was applied on the log entries for classification using a basic pattern matching rule. That step allowed the faster identification of the well-known and formatted log types. The use of this rule-based method allows the system to filter out the standardized logs as soon as possible thus reducing the initial computing delay at the time of classification.

Step 3: Machine Learning-Based Classification

The next step involves log entries that are not categorized via regex. Here, logs are processed by a machine learning model based on BERT embeddings for accurate classification of semi-structured log data. This model categorizes the semi-structured logs based on the features learned from them.

Step 4: NLP/LLM-Based Processing

In the case of remaining complex and unclear log messages, advanced natural language (NLP) processing methods are used by utilizing a large language model (LLM). In this phase, contextually aware interpretation of log messages is achieved.

Step 5: Brute-Force Attack Detection

The system also incorporates a security measure that detects brute force attack. This algorithm monitors how often there have been multiple failed log-ins originating from the same IP address within a limited time frame. Multiple attempts to sign-in from a single IP can indicate the Brute-Force Attack possibility.

Step 6: Output Generation

After undergoing the classification and detection process, there comes the output of the process in an Excel report format. This includes the categorized logs, detected threats, and overall summary of all logs, hence making it easy to interpret and analyze the outputs.

© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404034

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. *System Workflow:*
The overall workflow of the system follows a sequential pipeline:

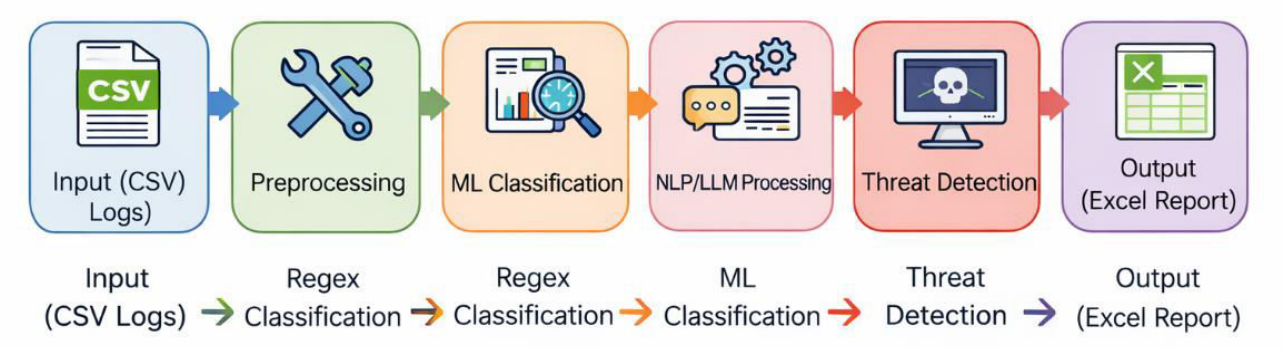


Fig. 1. System Workflow of LogSentinel

This hybrid workflow ensures that each log message is processed using the most suitable technique, improving overall system performance.

D. *Advantages of the Proposed Method:*
- Combines multiple techniques for improved accuracy
- Efficient handling of structured and unstructured logs
- Reduces manual effort in log analysis
- Detects security threats such as brute-force attacks
- Generates structured and easy-to-analyze reports

IV. RESULTS

The proposed system, LogSentinel, was implemented to evaluate its performance in automated log classification and anomaly detection. The system underwent empirical evaluation using sample log data in CSV format, comprised of both normal and anomalous log entries.

The classification process was carried out using different methods together. Initially, rule-based techniques which use regular expressions were used to organize log messages with a clear structure. This approach of pattern-matching provided fast and efficient classification for known patterns. Because some log messages could not be categorized by the regular expressions, a machine learning model which uses BERT embedding was used to find new patterns. For the remaining complex and unstructured logs, NLP-based processing was applied to improve the correctness of the results.

In addition to the above-mentioned functions, an additional function for detecting brute force attacks was implemented. A function that monitors repeatedly failed login attempts with a certain time frame by the same source allows the system to monitor and detect suspect actions. This is another example of how this system can be used to monitor threats in addition to categorizing logs.

The results were generated in the form of structured Excel reports, which included categorized logs and detected anomalies. Compared to traditional rule-based methods, the proposed hybrid approach provides improved flexibility and accuracy in handling diverse log formats. The integration of machine learning and NLP techniques enhances the system's ability to process unstructured data effectively.

Furthermore, this multi-step approach will reduce reliance on a single method and improve the overall efficiency of the system. Thus, these findings show that the developed system performs efficiently, can scale appropriately to handle large amounts of data from various sources, and is suitable for use in real-world implementations of automated log management and cyber security monitoring systems.

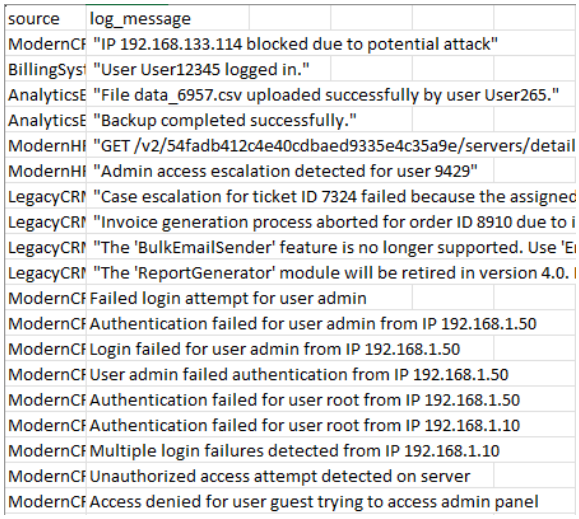
© 2026 IJIRCCE | Volume 14, Issue 4, April 2026 DOI: 10.15680/IJIRCCE.2026.1404034

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

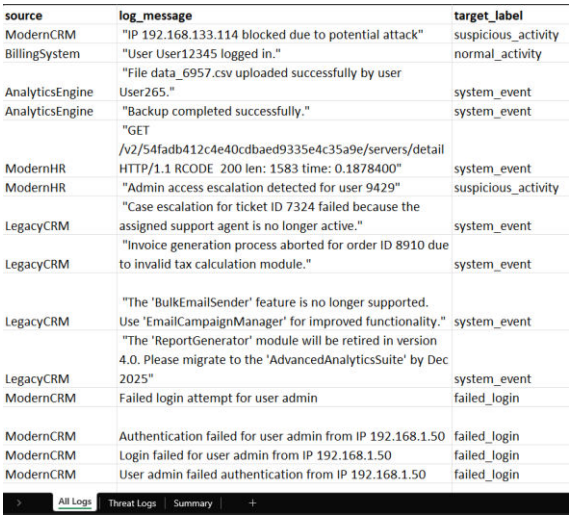
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Overall, the proposed system demonstrates reliable performance in log classification and anomaly detection, making it suitable for real-world applications in cybersecurity monitoring and system management.



source	log_message
ModernCF	"IP 192.168.133.114 blocked due to potential attack"
BillingSyst	"User User12345 logged in."
AnalyticsE	"File data_6957.csv uploaded successfully by user User265."
AnalyticsE	"Backup completed successfully."
ModernHI	"GET /v2/54fadb412c4e40cdbaed9335e4c35a9e/servers/detail
ModernHI	"Admin access escalation detected for user 9429"
LegacyCRI	"Case escalation for ticket ID 7324 failed because the assigned
LegacyCRI	"Invoice generation process aborted for order ID 8910 due to i
LegacyCRI	"The 'BulkEmailSender' feature is no longer supported. Use 'EI
LegacyCRI	"The 'ReportGenerator' module will be retired in version 4.0. I
ModernCF	Failed login attempt for user admin
ModernCF	Authentication failed for user admin from IP 192.168.1.50
ModernCF	Login failed for user admin from IP 192.168.1.50
ModernCF	User admin failed authentication from IP 192.168.1.50
ModernCF	Authentication failed for user root from IP 192.168.1.50
ModernCF	Authentication failed for user root from IP 192.168.1.10
ModernCF	Multiple login failures detected from IP 192.168.1.10
ModernCF	Unauthorized access attempt detected on server
ModernCF	Access denied for user guest trying to access admin panel

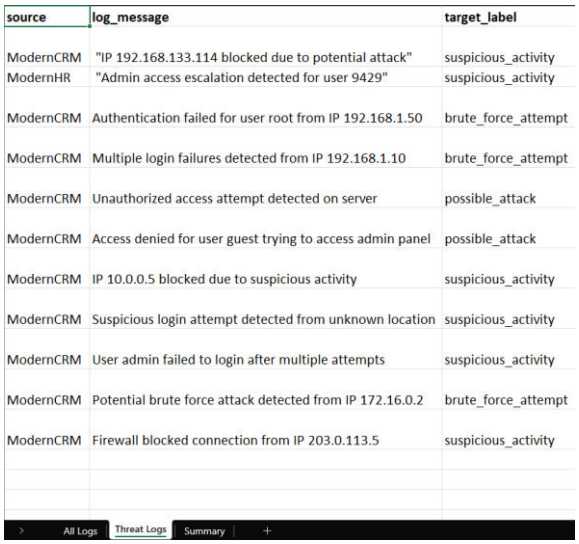
Fig. 2. Sample Input Log Data (CSV Format)



source	log_message	target_label
ModernCRM	"IP 192.168.133.114 blocked due to potential attack"	suspicious_activity
BillingSystem	"User User12345 logged in."	normal_activity
AnalyticsEngine	"File data_6957.csv uploaded successfully by user User265."	system_event
AnalyticsEngine	"Backup completed successfully."	system_event
ModernHR	"GET /v2/54fadb412c4e40cdbaed9335e4c35a9e/servers/detail HTTP/1.1 RCODE 200 len: 1583 time: 0.1878400"	system_event
ModernHR	"Admin access escalation detected for user 9429"	suspicious_activity
LegacyCRM	"Case escalation for ticket ID 7324 failed because the assigned support agent is no longer active."	system_event
LegacyCRM	"Invoice generation process aborted for order ID 8910 due to invalid tax calculation module."	system_event
LegacyCRM	"The 'BulkEmailSender' feature is no longer supported. Use 'EmailCampaignManager' for improved functionality."	system_event
LegacyCRM	"The 'ReportGenerator' module will be retired in version 4.0. Please migrate to the 'AdvancedAnalyticsSuite' by Dec 2025"	system_event
ModernCRM	Failed login attempt for user admin	failed_login
ModernCRM	Authentication failed for user admin from IP 192.168.1.50	failed_login
ModernCRM	Login failed for user admin from IP 192.168.1.50	failed_login
ModernCRM	User admin failed authentication from IP 192.168.1.50	failed_login

All Logs | Threat Logs | Summary | +

Fig. 3. Output Excel Report Showing All Classified Logs



source	log_message	target_label
ModernCRM	"IP 192.168.133.114 blocked due to potential attack"	suspicious_activity
ModernHR	"Admin access escalation detected for user 9429"	suspicious_activity
ModernCRM	Authentication failed for user root from IP 192.168.1.50	brute_force_attempt
ModernCRM	Multiple login failures detected from IP 192.168.1.10	brute_force_attempt
ModernCRM	Unauthorized access attempt detected on server	possible_attack
ModernCRM	Access denied for user guest trying to access admin panel	possible_attack
ModernCRM	IP 10.0.0.5 blocked due to suspicious activity	suspicious_activity
ModernCRM	Suspicious login attempt detected from unknown location	suspicious_activity
ModernCRM	User admin failed to login after multiple attempts	suspicious_activity
ModernCRM	Potential brute force attack detected from IP 172.16.0.2	brute_force_attempt
ModernCRM	Firewall blocked connection from IP 203.0.113.5	suspicious_activity

All Logs | Threat Logs | Summary | +

Fig. 4. Output Excel Report Showing Threat Logs

Label	Count
system_event	7
suspicious_activity	6
failed_login	5
brute_force_attempt	3
possible_attack	2
normal_activity	1

All Logs | Threat Logs | Summary | +

Fig. 5. Output Excel Report Summary of All logs

V. CONCLUSION AND FUTURE WORK

LogSentinel is a hybrid tool used for both identifying anomalies in system log data through log classification and automatic log-data processing. Unlike other tools that rely on only one strategy, LogSentinel combines traditional rule-based strategies using pre-defined rules (regular expressions) with machine-learned patterns and natural-language-processing based text-interpreting techniques to analyze the variety of log-entry types including formatted-line logs, partially-structured line logs, and unformatted-text logs. By integrating these multiple strategies in layers, classic-analysis gaps are minimized and speed in the identification and labeling of log events significantly increases.

Experimental results indicate that the hybrid approach efficiently identifies and labels log entries to detect potential security threats associated with brute-force attacks. The application of regular expression-based rules allows the system

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

to rapidly identify existing patterns; meanwhile, the incorporation of machine-learning and natural-language-processing techniques enables the system to process ambiguous and complex log entry data. In addition, structured Excel reports generated from the system simplify result interpretation and minimize user effort needed to perform manual analysis of the system's output.

While the system effectively performs its intended function, there are some limitations. For example, the performance of machine-learning models is dependent upon high-quality training data; furthermore, employing advanced language-processing techniques within the system can lead to increased computational overhead. As well, LogSentinel operates using static data sets, which means it cannot process logs in real time.

Future enhancements to the system will include developing real-time-log analysis capabilities, enhancing model optimization to increase performance and scalability. Incorporation into an enterprise environment through integration with a Security Information and Event Management (SIEM) system can also expand the use cases for this type of system. Developing and applying more advanced deep-learning architectures in conjunction with increasing the diversity of the dataset being analyzed can assist in the development of more robust and reliable anomaly-detection processes.

REFERENCES

1. Anfeng Peng, Ajesh K. Chathoth, Stephen Lee, "Log Anomaly Detection with Large Language Models via Knowledge-Enriched Fusion," arXiv preprint arXiv:2512.11997, 2025.
2. Wei Guan, Jian Cao, Shiyou Qian, Jianqi Gao, Chun Ouyang, "LogLLM: Log-based Anomaly Detection Using Large Language Models," arXiv preprint arXiv:2411.08561, 2024.
3. Max Landauer, Sebastian Onder, Florian Skopik, Markus Wurzenberger, "Deep Learning for Anomaly Detection in Log Data: A Survey," Machine Learning with Applications, vol. 12, 100470, 2023.
4. Viktor Beck, Max Landauer, Markus Wurzenberger, Florian Skopik, Andreas Rauber, "System Log Parsing with Large Language Models: A Review," arXiv preprint arXiv:2504.04877, 2025.
5. Fatemeh Hadadi, Qinghua Xu, Domenico Bianculli, Lionel C. Briand, "LLM Meets ML: Data-efficient Anomaly Detection on Unstable Logs," ACM Transactions on Software Engineering and Methodology, 2025.
6. Yinang Gao, Tongyi Luo, Kai Huang, et al., "LogLAA: An Adaptive Integrated Log Anomaly Analysis Framework," Cybersecurity, vol. 9, 141, 2026.
7. Zanis Ali Khan, Donghwan Shin, Domenico Bianculli, Lionel C. Briand, "The Impact of Log Parsing on Deep Learning-based Anomaly Detection," Empirical Software Engineering, vol. 29, 139, 2024.
8. Leeladhar Chourasiya, Sushma Khatri, U.K. Lilhore, et al., "Advanced System Log Analyzer for Anomaly Detection and Cyber Forensic Investigations using LSTM and Transformer Networks," Journal of Cloud Computing, vol. 14, 60, 2025.
9. Musaad Alzahrani, "Investigating the Impact of Log-Sequence Embeddings on Anomaly Detection: A Systematic Study," Information, vol. 17, no. 3, 228, 2026.
10. Shenglin Zhang, Ziang Chen, Zijing Que, et al., "LogPurge: Log Data Purification for Anomaly Detection via Rule-Enhanced Filtering," arXiv preprint arXiv:2511.14062, 2025.
11. Junjie Huang, Zhihan Jiang, Zhuangbin Chen, Michael R. Lyu, "ULog: Unsupervised Log Parsing with LLMs through Log Contrastive Units," arXiv preprint arXiv:2406.07174, 2024.
12. Jonathan Pan, Swee Liang Wong, Yidi Yuan, "RAGLog: Log Anomaly Detection using Retrieval-Augmented Generation," arXiv preprint arXiv:2311.05261, 2023.
13. Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, Hongyi Liu, Ying Li, "XRAGLog: A Resource-Efficient Context-Aware Log Anomaly Detection Method Using Retrieval-Augmented Generation," 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details