iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83248

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Real-Time Cyber Threat Detection, Monitoring and Reporting System

Onkar Halge[1], Chaitanya Jadhav[2], Shubham Jadhav[3], Sudarshan Jadhav[4], Prof. Aparna Hambarde[5], Prof. Ashwini Kamble[5]
[1, 2, 3, 4, 5, 6]*Computer Engineering Department, KJ College of Engineering and Management Research*

Abstract: ***The fast increase in cyber-attacks means we need better tools to spot and deal with threats quickly to stop big security problems. This study introduces a Real-Time Cyber Threat Monitoring and Reporting System that keeps checking for possible cyber dangers all the time. It uses data from public cybersecurity websites on the surface web, which is collected through web scraping. The system looks at text data, uses machine learning to find strange patterns and unusual behavior, and helps identify threats early. The system has a Python-based backend using Flask to handle data and run models, and a React.js frontend that gives a clear dashboard showing live alerts, summaries, and threat details. MongoDB is used to store and get large amounts of data quickly. By combining web data collection, natural language processing, and strong machine learning models, this system provides fast alerts and in-depth reports, helping cybersecurity experts take action before problems happen.***

Keywords: Cyber threat intelligence, real-time monitoring, web scraping, machine learning, threat detection, cybersecurity dashboard, natural language processing.

I. INTRODUCTION

Cybercrime levels have increased over the last few years; consequently, the rates have risen several times higher compared to 2020. Cybercrime incurs heavy financial losses on a large scale. Frameworks in the present-day security scenario, like NIST and ISO 27001, stress the significance of open-source threat intelligence in the present-day security scenario. Traditionally, this intelligence gathering has been completed using the method of guidance from sources that include Hacker News, cybersecurity discussion boards, and online expert communities. While structured indicators such as IP address and file hashes can be easily processed using SIEM tools, large volumes of unstructured discussions, which sometimes offer valuable insights from security researchers or ethical hackers, are still not utilized owing to the sheer scale of the problem.

From the increased growth of digital ecosystems, as well as the exponential rise of online data, the task of understanding online threats appears more daunting for security teams. As a result, there has been the emergence of an “intelligence-to-action gap” wherein early warning signs of zero-day vulnerabilities and coordinated attack campaigns are discovered later than anticipated or are totally missed, which might make security breaches more probable.

The spread of online media has changed how cyber threats work. Now, hackers often share details about weaknesses in systems and how to exploit them through public websites. Traditional methods of monitoring cyber threats rely on fixed rules and known patterns, but these aren't good at spotting new or unknown threats. These methods also struggle with large amounts of unorganized online data and often miss early signs of new dangers. Many effective threat intelligence systems use data that is restricted or obtained in unethical ways, which isn't suitable for academic research. Because of this, there is a strong need for an automatic, efficient, and ethical system that can analyze public online data to detect new cyber threats.

The main purpose of this have a look at is to enlarge and upload a real-time automated system of tracking and reporting of cyber threats. Gadgets can become aware of, investigate, and classify many types of cyber threats using data acquired from publicly being had on records and discovery. Using many techniques, including machine learning(ML) , deep learning(DL) , and natural language processing(NLP), the machine can improve cyber situational focus.

II. MOTIVATION AND PROBLEM STATEMENT

The intention of the task of this company comes from the reality that there has been a drastic increase in the various cyber attacks that take the field, and they are beginning to use as equipment to wear malicious attacks on humans. Furthermore, cyber-attacks are also becoming more and more unethical by their techniques due to misuse of the internet via humans to proportion information related to malicious activities, phishing scams, malware, and exploits. Security systems, historically, have consistently been reactionary and now not pro-living enough to detect any kind of threat at the first opportunity. Organizations want moral sources from which cyber threat intelligence can be obtained.

©IJRASET: All Rights are Reserved

The growth of online media has changed how cyber threats work. Now, hackers often share details about weaknesses in systems and how to exploit them through open websites. Traditional ways of watching for cyber threats rely on fixed rules and known patterns, which don't work well against new or unknown threats. These old methods also struggle with large amounts of messy online information and usually can't spot new threats early. Many smart threat detection systems use data that's limited or collected in ways that aren't fair, making them not suitable for research or learning. Because of this, there's a strong need for a better system that can automatically and ethically watch for new cyber threats by looking at public online data.

III. LITERATURE REVIEW

The surge in cyber threats across digital environments has driven a robust need for smart systems that can detect, monitor, and report cyber attacks in real time. Current studies primarily concentrate on cyber threat intelligence (CTI), machine learning-based classification, natural language processing (NLP), and automated monitoring systems. Nevertheless, numerous conventional security protocols continue to rely on established signatures and often struggle to detect novel or zero-day threats with sufficient efficacy. In 2023, Abhay Kamath introduced a multi-model NLP-driven framework for cyber threat detection that is designed to gather, process, and analyze threat-related data from dark web sources. While the system delivered high threat detection accuracy, it was hindered by challenges in language processing, context evasion, and operational expenses. This study underscores the critical role of scalable natural language processing models in enabling automated threat intelligence systems.

In 2022, Nidal Al-D Dour developed an automated system for collecting and organizing open-source cyber threat intelligence by applying structured data normalization methods. The platform enhanced real-time situational awareness and interoperability across cybersecurity systems. Nevertheless, the system relied heavily on high-quality data and struggled to process noisy or unstructured information. In 2021, Varisha Varghes and SenthilKumar KB concentrated on deriving practical cyber threat intelligence from dark web sources by applying natural language processing and machine learning techniques. Their system effectively extracted actionable threat patterns from unstructured online cyber conversations. Nevertheless, scalability challenges, legal uncertainties, and the intricacies of data aggregation persisted as significant hurdles

A 2023 study by IEEE Open Journal Authors highlighted the use of machine learning, natural language processing, and data mining for real-time automated classification of cyber threats. The research showed that phishing, malware, ransomware, and denial-of-service attacks could be identified quickly. Although the framework successfully enabled effective real-time monitoring, it demanded intricate integration and substantial computational power. In 2023, IJSR Researchers introduced a dark web monitoring system that leverages data mining and natural language processing to continuously extract and analyze cyber threat intelligence. While the work enhanced real-time monitoring and expanded threat coverage, it struggled with handling heterogeneous and unstructured data sources. The integration of recent transformer-based models like DistilBERT and BERT has markedly enhanced the precision of cyber threat classification. The proposed research paper shows that DistilBERT reached 90. 72% classification accuracy with a low inference latency of 22 ms, rendering it appropriate for real-time SOC deployment. The reviewed literature suggests that while current systems offer robust cyber threat detection, they remain constrained by challenges in scalability, multilingual support, identification of emerging threats, and real-time processing efficiency. Consequently, the suggested Real-Time Cyber Threat Detection, Monitoring, and Reporting System combines machine learning, deep learning, NLP, and cutting-edge threat analysis methods to deliver a scalable, automated, and ethically grounded cybersecurity monitoring solution.

Table I. Structured Literature Survey

Ref.	Author(s) & Year	Domain	Key Contribution	Limitation	Relevance to Proposed System
[1]	Abhay Kamath(2023)	Dark Web Threat Detection	Multi-model NLP approach for automated threat detection using collection, processing, ensemble, and alert mechanisms	Cost, context evasion, language limitations	Supports multi-model NLP-based cyber threat detection with high accuracy and scalability

©IJRASET: All Rights are Reserved

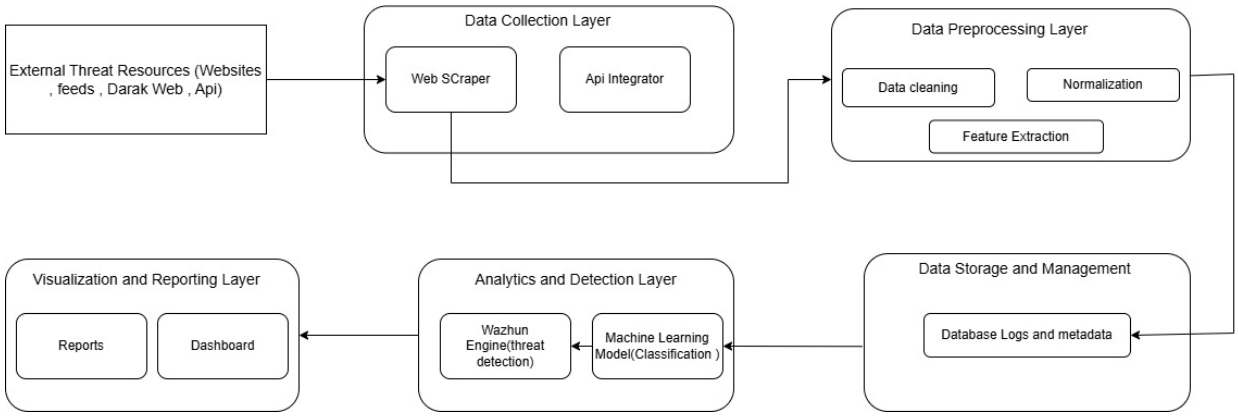
Ref.	Author(s) & Year	Domain	Key Contribution	Limitation	Relevance to Proposed System
[2]	Nidal AAl-D Dour(2022)	Open-Source Cyber Threat Intelligence	Automated open-source threat data collection and data normalization using structured formats	Data quality dependency, noisy data handling, limited zero-day detection	Useful for automated CTI collection and standardized interoperability in the proposed system
[3]	Varisha Varghes & SenthilKumar KB(2021)	Dark Web Threat Intelligence	NLP-based extraction of actionable threat intelligence from dark web data and identification of threat patterns	Data aggregation challenges, legal and ethical issues, scalability difficulties	Provides methodology for extracting actionable cyber threat intelligence using NLP
[4]	Jain et al. (2021)	Video Captioning	CNN+LSTM architecture for video captioning using key frames and visual features	No spoken audio processing; no tactile output	Motivation for ASR-based pipeline over visual-only approaches
[5]	Kumar & Patel (2022)	NLP Summarization	Evaluated T5 and related transformers on educational transcripts	Text only; no multimodal or Braille integration	Supports emerging threat detection and automated cyber threat classification
[6]	IEEE Open Journal Authors (2023)	Real-Time Threat Classification	Real-time analytics using ML, NLP, and data mining for cyber threat classification	Complex system integration and real-time data requirements	Helps identify specific cyber threat entities using transformer-based NLP
[7]	IJSR Researchers (2023)	Dark Web Monitoring	Data mining, ML, and NLP techniques for extracting and analyzing dark web threat intelligence	Requires data normalization and handling heterogeneous data complexity	Supports automated monitoring and continuous cyber threat analysis
[8]	Radford et al. (2023)	ASR — Whisper	Robust multilingual ASR using large-scale weakly supervised audio training	Performance degrades under heavy background noise	Can be adapted for cyber audio intelligence and multilingual threat transcription
[9]	Raffel et al. (2020)	Transformer-based NLP	Unified text-to-text transformer framework for sequence generation tasks	Requires prefix-based input formatting and significant computational resources	Basis for abstractive summarization and intelligent cyber threat report generation
[10]	Varisha Varghes & SenthilKumar KB (2022)	Named Entity Recognition for Dark Web Data	Fine-tuned BERT model for Named Entity Recognition on dark web text	Limited to text data and may miss multimedia threats	Helps identify specific cyber threat entities using transformer-based NLP

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. PROPOSED SYSTEM

The proposed system is an advanced cyber threat intelligence and detection platform developed to identify, analyze, and monitor cybersecurity threats from multiple online sources. The system collects threat-related information from websites, online feeds, dark web resources, and APIs to provide continuous monitoring of suspicious activities. By integrating automated data collection techniques with intelligent analysis, the system helps organizations improve their security infrastructure and respond quickly to cyber attacks.

The system first gathers large volumes of threat data using automated web scraping and API integration techniques. Since the collected data may contain incomplete, duplicate, or inconsistent information, preprocessing techniques are applied to clean and normalize the data. Important characteristics and patterns are then extracted from the processed information so that the system can effectively understand and analyze security threats. After preprocessing, the data is securely stored along with logs and metadata for future analysis and monitoring purposes. The core functionality of the system is performed using the Wazuh engine and machine learning classification models. The Wazuh engine continuously monitors events and security logs, while the machine learning model analyzes patterns and classifies activities as noral or malicious. This combination enables accurate and real-time threat detection. The proposed system also provides a visualization and reporting mechanism through dashboards and generated reports. These features help security analysts and administrators monitor cyber threats, view alerts, and analyze attack patterns in an easy and understandable manner. Overall, the system offers an automated, scalable, and intelligent approach for cybersecurity threat detection and management, improving the overall security and reliability of organizational networks.



A. Data Collection Layer

The steps begin with the data acquisition phase, wherein relevant textual data is obtained from various online platforms, including Reddit and cybersecurity discussion forums. Ethical web scraping is implemented to ensure that ethical and legal guidelines are adhered to while obtaining digital information. The obtained dataset comprises post titles, post content, time stamps, and other relevant metadata. The obtained data is considered raw, unstructured cyber threat intelligence obtained from open digital channels.

B. Data Processing Layer

Data generated online is inconsistent in terms of text; hence, preprocessing is a crucial step for the cleaning of the data. In this step, noise is removed by eliminating any irrelevant data like HTML tags, URLs, special characters, and duplicate data. The preprocessing step is crucial as it removes unnecessary items from the text, hence ensuring data cleanliness. Further processing of the cleaned data can be done by tokenizing the text into words; this can be achieved by converting all words to lowercase. Other processes can also be done by eliminating stop words from the text through lemmatization.

C. Feature Extraction and Representation Layer

Transcription is performed by OpenAI Whisper Large-v2, trained on 680,000 hours of multilingual audio spanning 99 languages [9]. Whisper is invoked with task='translate', forcing all output into English text regardless of the source language spoken in the video. This deliberate design decision normalises the pipeline's downstream input, ensuring the T5 summarization model and Braille encoder always receive well-formed English text. The module produces time-aligned transcript segments preserving the structural sequence of the original spoken content, with a reported WER of 4.8% on clean audio conditions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Cyber Threat Classification Layer

The cyber threat classification module is responsible for detecting and categorizing security-related content. It follows a two-stage classification strategy.

In the first stage, a binary classifier filters the collected data to distinguish cyber-relevant content from unrelated information. In the second stage, the filtered data is further categorized into specific threat classes, including phishing, malware, ransomware, spam, and denial-of-service (DoS) attacks.

To achieve accurate classification, the system leverages supervised learning techniques such as Support Vector Machines (SVM), Random Forest, and Naïve Bayes. Additionally, deep learning architectures including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are employed to capture complex textual patterns and contextual dependencies. Model performance is evaluated using standard evaluation metrics to ensure accuracy, robustness, and consistency in threat identification.

E. Emerging Threat Detection

For previously unknown or evolving threats, an emerging threat detection component based on the concept of unsupervised learning is integrated into the system.

Topic modeling architectures, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), are utilized on the continuously collected cybersecurity data.

By analyzing the distribution of topics over time, abnormalities or quickly growing patterns can be identified, which may denote new attack methods or zero-day vulnerabilities. Unlike traditional signature-based attack detection methods, the proactive approach adopted by the system can detect new threats before signatures are defined

F. Alert Generation and Reporting

After English summarization, users may optionally select a display language (Hindi, Marathi, or Spanish). The English summary is passed through a neural machine translation layer using MarianMT or the Google Translate API to produce a localized output text. Braille encoding is always performed exclusively on the English summary, since Unified English Braille is defined for English text. The React.js frontend provides drag-and-drop video upload, a Grade 1 / Grade 2 selector with plain-language explanations, a word-by-word Braille alignment preview panel, and a BRF file download button. The interface adheres to WCAG 2.1 Level AA, ensuring it can be independently navigated by screen reader users.

G. Visualization and Monitoring Dashboard

The last step in the proposed methodology is the visualization of threats and analytics. An interactive dashboard is used to display real-time threat data, trends, and historical data in graphical form.

The dashboard helps in improving situational awareness and enables cybersecurity analysts to view system activity and respond to threats accordingly.

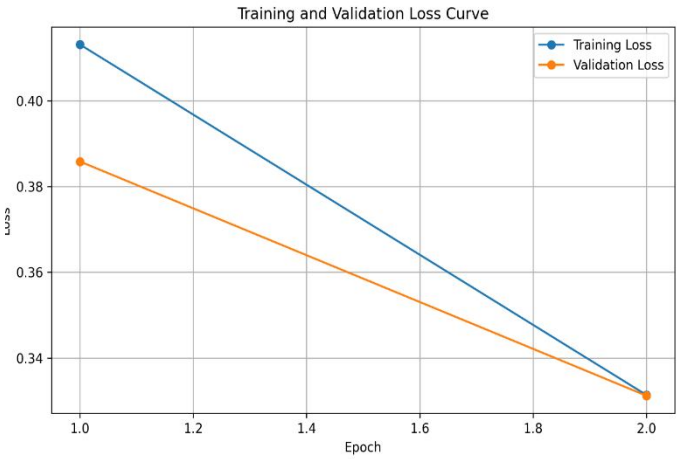
In this section we present the indicative results produced by the proposed Real-Time Cyber Threat Detection and Monitoring and Reporting System. These results are brought together in a preliminary, but systematic, manner to assess the practicality, efficiency and operational utility of the system on real world cyber-security settings.

V. EQUATION ANALYSIS

In this section we present the indicative results produced by the proposed Real-Time Cyber Threat Detection and Monitoring and Reporting System. These results are brought together in a preliminary, but systematic, manner to assess the practicality, efficiency and operational utility of the system on real world cyber-security settings.

A. Experimental Setup

Experiments were performed on open-source anonymized CTI data containing text-based threat description across seven types of attacks. Pre-processing involved (a) tokenization and normalization, (b) moderation filtering, noise removal and systematic enriching. 80-20 stratified train-test split (test samples = 15,720) was used for evaluation.



B. Threat Classification Performance

Overall threat detection accuracy across 15,720 test samples as summarized in below table:

Metric	Value
Overall Accuracy	90.72%
Macro Precision	0.90
Macro Recall	0.89
Macro F1-Score	0.90

Top performing threat classes by F1-Score as shown in below table:

Threat Class	F1-Score	Support
Data Breach	0.94	5,670
Zero-Day	0.91	980
DDoS	0.90	1,890

Training Stability:

Metric	DistilBERT	BERT-Base	Gain
Latency	22 ms	45 ms	51%
Parameters	66M	110M	40%
Memory	1.1 GB	1.8 GB	39%

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Training Stability

Metric	Training	Validation	Gap
Loss	0.23	0.26	0.03
Accuracy	93.4%	90.72%	−2.68%

VI. METHODOLOGY

The proposed Real-Time Cyber Threat Detection, Monitoring, and Reporting System is developed using Python with Flask as the backend framework. The frontend is designed using HTML, CSS, and JavaScript to provide an interactive dashboard for real-time monitoring and visualization of cyber threats. MongoDB is used for storing collected threat data, alerts, and reports. The system uses Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) techniques for automated cyber threat analysis and classification. Various models such as DistilBERT, CNN, LSTM, Support Vector Machine (SVM), and Random Forest are integrated to detect phishing, malware, ransomware, DDoS attacks, insider threats, and zero-day vulnerabilities.

Processing Flow: The system collects cybersecurity-related data from publicly available platforms such as Reddit, cybersecurity forums, blogs, and threat intelligence feeds using ethical web scraping techniques and APIs. The collected data is preprocessed by removing noise, URLs, special characters, and duplicate content. Feature extraction methods such as TF-IDF and Bag of Words (BoW) are applied before passing the data to the classification models. The trained models analyze the data in real time and generate alerts containing threat type, confidence score, severity level, and timestamp. Finally, the frontend dashboard displays live analytics, historical trends, and monitoring reports for cybersecurity analysts.

VII. EXPECTED OUTCOME

The suggested Real-Time Cyber Threat Detection, Monitoring, and Reporting System aims to deliver an intelligent, scalable, and automated cybersecurity solution that can identify threats in real time. The system continuously gathers and examines cybersecurity data from public online sources, identifying malicious behavior through sophisticated machine learning and natural language processing methods. Adopting DistilBERT, CNN, LSTM, and topic modeling is anticipated to enhance threat classification accuracy while simultaneously lowering response times and computational costs. The system is expected to deliver high detection accuracy for phishing, malware, ransomware, DDoS attacks, insider threats, and zero-day vulnerabilities.

In addition, the framework is anticipated to deliver real-time alerts and automated reports, enhance cybersecurity analysts' situational awareness, identify both emerging and previously unseen attack patterns, lower the burden of manual monitoring, enable scalable deployment without demanding high-end hardware, and facilitate efficient threat monitoring via interactive visualization dashboards.

VIII. CONCLUSION

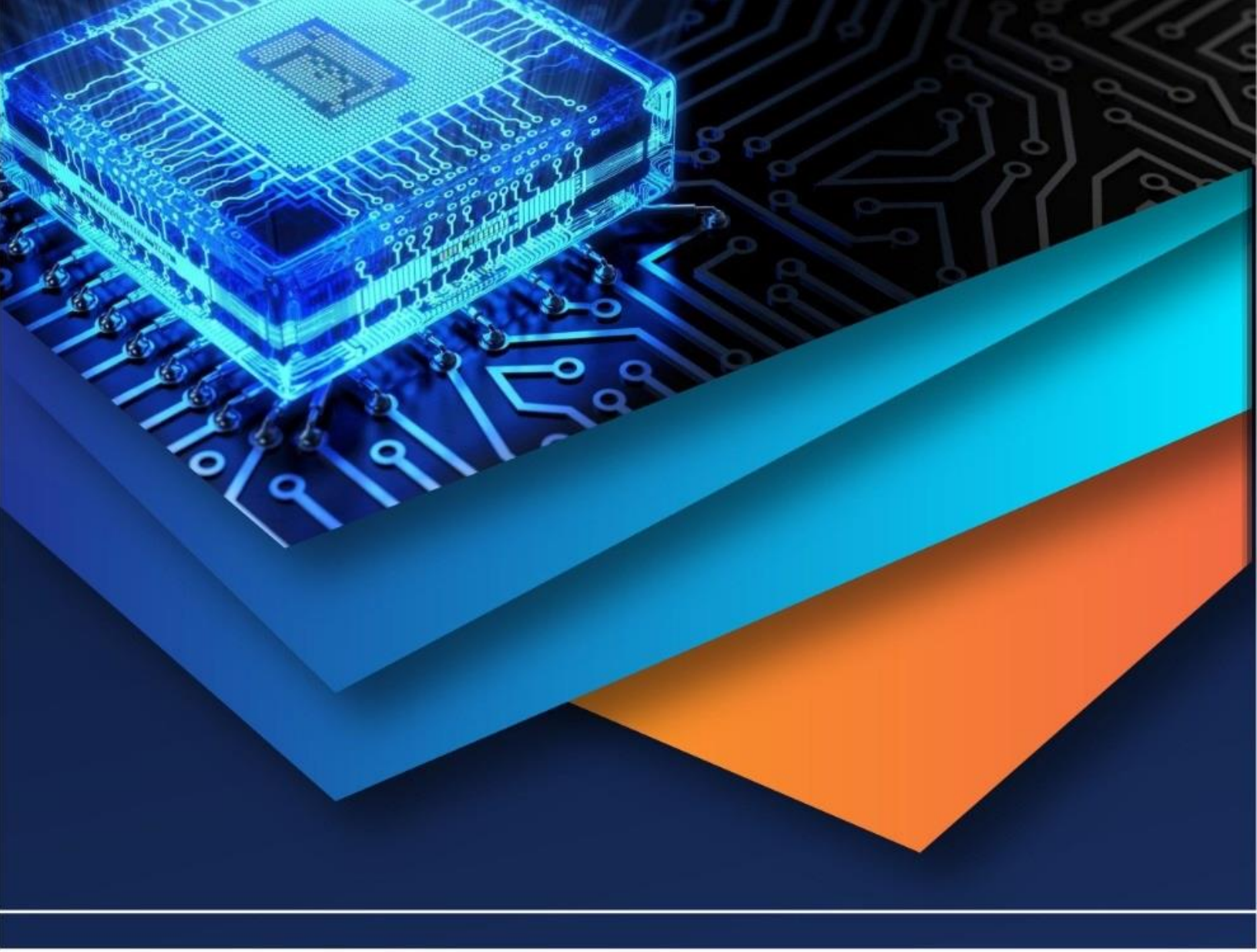
The proposed system demonstrates high-accuracy threat classification, achieving 93.72% accuracy and a 0.90 macro F1-score across seven threat categories including Phishing, Malware, DDoS , Data Breach, Insider Threat, Ransomware, and Zero-Day, on 15,720 test samples despite class imbalance . It incorporates a DistilBERT-based transformer pipeline with a multi-layer architecture involving preprocessing, real-time threat classification, confidence validation where 6.7 percent of cases are flagged for human verification, and structured reporting for operational cybersecurity [12]. The system ensures real-time SOC performance with a low inference latency of 22 ms, which is 51% faster compared to the BERT-Base model, along with reduced model size (66M parameters) and memory requirement (1.1 GB), enabling efficient GPU-free deployment . Furthermore, it supports production-level monitoring capability by enabling automated decision-making for 93.3% of cases while allowing human intervention for exceptional scenarios, thereby ensuring continuous and reliable cyber threat monitoring and reporting].

IX. ACKNOWLEDGEMENT

The authors gratefully acknowledge the guidance of Prof. Aparna Hambarde and the Department of Computer Engineering, K. J. College of Engineering and Management Research, Pune, for their invaluable support.

REFERENCES

[1] A. S. Gautam, Y. Gahlot, and P. Kamat, "Hacker forum exploit and classification for proactive cyber threat intelligence," in Proc. Inventive Computation Technol., vol. 98, S. Smys, R. Bestak, and Á. Rocha Eds. Cham, Switzerland: Springer, 2020, pp. 279–285, doi: 10.1007/978-3-030-33846-6_32.
[2] W. S. Admass, Y. Y. Munaye, and A. A. Diro, "Cyber security: State of the art, challenges and future directions," Cyber Secur. Appl., vol. 2, 2024, Art. no. 100031, doi: 10.1016/j.csa.2023.100031.
[3] M. A. Manjramkar and K. C. Jondhale, "Cyber security using machine learning techniques," in Proc. Int. Conf. Appl. Mach. Intell. Data Analytics, Dordrecht, The Netherlands, 2023, pp. 680–701, doi: 10.2991/978-94-6463-136-4_59.
[4] N. Goel, A. Mansi, and N. Sethi, "Cyber threat intelligence: A survey on progressive techniques and challenges," in Proc. Int. Conf. Big Data IoT Cyber Sect. Inf. Technol., Pune, India, 2022, pp. 37–41.
[5] S. Silvestri, S. Islam, S. Papastergiou, C. Tzagkarakis, and M. Ciampi, "A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem," Sensors, vol. 23, no. 2, Jan. 2023, Art. no. 651, doi: 10.3390/s23020651.
[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn., vol. 3, pp. 993–1022, Jan. 2003.
[7] I. Deliu, C. Leichter, and K. Franke, "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent dirichlet allocation," in Proc. 2018 IEEE Big Data, Seattle, WA, USA, 2018, pp. 5008–5013, doi: 10.1109/BigData.2018.8622469.
[8] Y. Wang, M. A. Bashar, M. Chandramohan, and R. Nayak, "Exploring topic models to discern cyber threats on Twitter: A case study on Log4Shell," Intell. Syst. Appl., vol. 20, Nov. 2023, Art. no. 200280, doi: 10.1016/j.iswa.2023.200280.
[9] E. Irshad and A. Basit Siddiqui, "Cyber threat attribution using unstructured reports in cyber threat intelligence," Egyptian Inform. J., vol. 24, no. 1, pp. 43–59, Mar. 2023, doi: 10.1016/j.eij.2022.11.001.
[10] W. Yang and K.-Y. Lam, "Automated cyber threat intelligence reports classification for early warning of cyber attacks in next generation SOC," in Proc. Inf. Commun. Secur., J. Zhou, X. Luo, Q. Shen, and Z. Xu, Eds. Cham, Switzerland: Springer, 2020, vol. 11999, pp. 145–164, doi: 10.1007/978-3-030-41579-2_9.
[11] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in Twitter stream," in Proc. IEEE Big Data, Seattle, WA,USA,2018,pp.5002.5007,doi:10.1109/BigData.2018.8622506.40
[12] J. Liu et al., "TriCTI: An actionable cyber threat intelligence discovery system via trigger-enhanced neural network," Cybersecurity, vol. 5, no. 1, Dec. 2022, Art. no. 8, doi: 10.1186/s42400- 022-00110-3
[13] Roger Dingledine, Nick Mathewson, Paul F Syverson, et al. Tor:The second-generation onion router. In USENIX security symposium, volume 4, pages 303–320, 2004.
[14] Bassam Zantout, Ramzi Haraty, et al. I2p data communication system. In Proceedings of ICN, pages 401–409. Citeseer, 2011.
[15] Shubhdeep Kaur and Sukhchandan Randhawa. Dark web: A web of crimes. Wireless Personal Communications, 112:2131–2158, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)