International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Trustworthy AI -Driven Cyber Attack Detection A Review

Sahana Kumari B, Basavaraj Japannavar, D S Tejaswini, Dheeraj R, Kumaragouda Tangoudar

Department of Computer Science and Engineering SDM Institute of Technology, Ujire Dakshina Kannada, Karnataka, India

kumaribsahana07@gmail.com,22a22@sdmit.in, 22a29@sdmit.in, 22a35@sdmit.in, 22a51@sdmit.in.

***Abstract*—** The real-time cyber threat detection system presented in this paper is intended to shield users from threats posed by steganographic images, malicious files, and URLs. The system uses signature-based scanning engines like ClamAV, VirusTotal, and Google Safe Browsing instead of sophisticated machine learning algorithms. It supports both automated scanning of the Downloads folder, where newly added files are instantly examined, and manual inspection, which enables users to confirm questionable content. Users receive immediate alerts via WebSocket communication, and threats are categorized into severity levels (High, Medium, Low, and Clean) based on scan results.A built-in AI chatbot, powered by Groq's LLAMA 3.3 model, assists users by explaining scan outcomes and offering security guidance. The system emphasizes usability, privacy, and low resource consumption, making it suitable for personal and small-scale organizational use. Future enhancements include browser extensions, mobile integration, and support for cloud storage scanning. This hybrid approach offers a scalable and accessible solution for modern cybersecurity needs

Keywords—cyber threat detection, malware scanning, URL analysis, image steganography, real-time alerts, AI chatbot, ClamAV, VirusTotal, Google Safe Browsing, threat classification.

1. INTRODUCTION

Cyber threats have grown more sophisticated in today's digital environment, targeting users through malicious file downloads, phony websites, and hidden payloads embedded in photos. Conventional antivirus programs frequently rely on manual user intervention and recurring scans, which might not be adequate for real-time protection. Moreover, many advanced detection systems depend heavily on machine learning algorithms, which require large datasets, continuous training, and significant computational resources—making them less practical for lightweight, user-friendly applications Without depending on intricate algorithms, this project presents a real-time cyber threat detection system that puts accessibility, responsiveness, and multi-layered protection first. The system is built as a full-stack web application that automatically scans newly added files, keeps an eye on the user's Downloads folder, and sends out WebSocket-based alerts with instant feedback. It incorporates industry-standard security engines like Google Safe Browsing for URL reputation checks, VirusTotal for cloud-based multi-engine analysis, and ClamAV for local antivirus scanning. It also has a unique image scanner that uses Least Significant Bit (LSB) analysis to identify steganographic threats. To enhance user experience and LITERATURE REVIEW AND ANALYSIS understanding, the system features an AI-powered chatbot that explains scan results, offers security best practices, and guides users in responding to detected threats. Based on scan results, threats are divided into four severity levels: High, Medium, Low, and Clean. This enables users to make prompt, well-informed decisions. The dashboard offers real-time progress updates, threat statistics, and a visual summary of scanhistory. The project's goal is to provide a scalable, privacy-conscious solution that can be used for individual use, small businesses, and educational settings. Future enhancements include browser extensions for live URL scanning, mobile app integration, and support for cloud storage platforms. By combining trusted security engines with real-time automation and user-centric design, this system offers a practical approach to modern cybersecurity challenges.

2. LITERATURE REVIEW AND ANALYSIS

Review of Related Work

Mir, W. A., and Sanap, M. E. (2025). Harnessing AI for Cybersecurity: Real-Time Detection and Mitigation of Online Threats – A Survey. JSPM University in Pune. This paper

presents a systematic survey on the application of artificial intelligence in cybersecurity, with a focus on deep learning and machine learning techniques for online threat detection and mitigation in real time. The authors examined 290 studies published between 2021 and 2025 using the PRISMA methodology, selecting 55 pertinent works and thoroughly examining 30 of them. The study draws attention to the shortcomings of conventional rule-based defenses against emerging threats like malware, phishing, denial-of-service, SQL injection, and cross-site scripting. It has been demonstrated that deep learning models, such as CNNs and DNNs, perform better than traditional systems by facilitating adaptive learning, automated response, and increased accuracy. In order to improve resilience in cloud and network environments and provide scalable and dependable solutions for contemporary cybersecurity issues, the paper concludes that AI-driven approaches are crucial.[1].

Kale, Arjun, AI-Driven Cybersecurity Threats: Organizational Impact, California State University - San Bernardino,2024. This dissertation explores the impact of AI-driven cyber threats on organizations. The research was conducted using a case study approach and focused on three questions: how can organizations strengthen their cybersecurity defenses? How will the cybercriminals of the future use AI tools to exploit their victims? What strategies can be used to build resilience to phishing attacks? The findings showed that organizations can use AI-based product/solutions (ex: Vectra Cognito, AWS integration, etc.) to improve the efficiency of their real-time detection and response processes; whereas cybercriminals are expected to use AI-based hacking tools (ex: HackerGPT) and multiple generative adversarial networks (GANs) to launch increasingly sophisticated phishing attacks and create botnets. According to a number of case studies, email security solutions enabled with AI, such as Barracuda Essentials and Sentinel, have been shown to be very effective at protecting against phishing because they incorporate many different levels of protection.The case studies also highlight the importance of AI serving as a dual-purpose tool: Protecting from attacks as well as attacking. To ensure the continued success of organizations, it will be necessary to have advanced detection systems in place combined with employee training and an ethically driven regulatory framework to reduce risk and provide the ability to continue operations.[2].

The article by Sharma et al. presents an approach to overcome the problems posed by using black box classification models within the area of computer cybersecurity through the introduction of explanatory artificial intelligence (XAI). The use of XAI allows organizations to improve transparency, understanding and confidence in intrusion detection systems (IDS). By exploring adversarial techniques for examining misclassifications, the authors establish a technique to find the minimum impact features needed to generate the correct classification of a data point, thereby exposing important or influential features. The methodology of the research has been assessed through multiple benchmarking research results (such as NSL-KDD99 and PDF) and utilised classifiers (including Random Forest, KNN, SVM, MLP) were also tested. Using this methodology, classification rates were obtained that exceeded 95% accuracy. Additionally, this research evaluates the vulnerability of post-hoc explanation methodologies (for example, LIME and SHAP) to adversarial attack and indicates that it is possible to manipulate or mislead users of such maps, without changing the feature output of the classifier. The conclusions highlight the_necessity for_organisations wishing to maintain the advantages of XAI in order to effectively implement cybersecurity systems that maintain their resilience against adversarial instances whilst still providing interpretability and user confidence. [3].

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for Cyber Attacks. Journal of Information Security and Applications, 57, 102722. Elsevier. The article "Weaponized AI for Cyber Attacks" published in the Journal of Information Security and Applications by Yamin et al. (2021) looks at how powered AI systems are being used for cyber crime. It also details how adversarial machine learning and (Generative Adversarial Networks (GANs)) can be used as weapons to perform advanced cyber-crimes.In addition to the potential for devastating effects if these AI systems are not controlled, this paper discusses actual examples of when AI has been abused, such as; Adversarially manipulated images in healthcare, manipulation or creation of Traffic signal systems, and the creation of malicious software (Malware).This study has three main categories that describe attacks that are powered by Artificial Intelligence: Data Misclassification, Synthetic Data Generation, and AI-assisted Classical Cyber-Attack Generation Methods. The authors have identified and compared these categories to traditional Cyber Crime attacks using the STRIDE Cyber Crime threat Model.The paper also deals with some of the issues

surrounding GANs in generating Adversarial samples and the issues associated with training GAN networks due to optimization issues. The authors also emphasize that without immediate global cooperation, regulation and ethical frameworks for the use of "Weaponized AI", there is a risk that the uncontrolled proliferation of Weaponized AI could develop into an "AI Arms Race" with catastrophic effects to Cyber Security and Global Stability. [4].

Promyslov, V. G., Semenkov, K. V., & Shumov, A. S. (2019). *A Clustering Method of Asset Cybersecurity Classification*. IFAC Papers Online, 52(13), 928–933. Elsevier. A methodology for classifying assets by their Cybersecurity Level has been developed and proposed using a formal clustering technique. This methodology has been developed specifically for nuclear power plant instrumentation and control systems. In developing this methodology, the authors observed that existing forms of asset classification developed by experts are often based on highly subjective classifications that can be inconsistent across different subject experts. The authors, therefore, propose formalizing asset classification through a clustering process by collecting attributes, such as Nuclear Safety Classification (NSC), functional properties, and using those attributes as input to the clustering process. Using a combination of Security Graphs and the Take-Grant Model to represent the relationships existing between assets, the authors utilize k-means clustering to group assets into levels of Cybersecurity. An example application of this method used it to show how a set of assets could be partitioned into multiple distinct Cybersecurity Classes that were based on objective asset attributes rather than subjective application of expert judgment. In conclusion, the authors noted that this methodology will provide a structured and scalable approach for asset classification as well as provide a formalized solution to the problem; however, they indicated that classification of hierarchical systems may require inductive techniques for them to be fully applied.[5].

Jia, Y., Gu, Z., Du, L., Long, Y., Wang, Y., Li, J., & Zhang, Y. (2023). *Artificial Intelligence Enabled Cyber Security Defense for Smart Cities: A Novel Attack Detection Framework Based on the MDATA Model*. Knowledge-Based Systems, 276, 110781. Elsevier. This paper introduces ACAM is an advanced cybersecurity analysis framework specifically tailored for the needs of smart cities, and is built on the MDATA (Multi-dimensional Data Association and intelligent Analysis) model. ACAM differs from typical rules-based and anomaly-based detection systems by using dynamic spatio-temporal knowledge representation that helps detect multi-step cyberattacks more efficiently. ACAM has four modules that provide the framework: knowledge extraction, subgraph generation, alarm correlation, and attack detection. By converting all alarms and vulnerabilities to MDATA Multi-Dimensional Data Association graphs, the amount of false alarms produced by the framework is significantly reduced and the overall efficiency of detection is improved. Evaluation of the framework performed in the Cyber Range has shown that ACAM was able to achieve an average detection rate of 90.6% with only 14 ms of average lag time in detection from initiation of the attack to notification of the attack. Based on the results from this study, it can be concluded that ACAM is significantly better than traditional methods for improving a smart city's resistance against multiple step cyberattacks, and it is also a scalable and explainable AI-based defensive mechanism. [6].

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). *Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions*. Information Fusion, 97, 101804. Elsevier. The aim of this paper is to provide a systematic review of the literature about the use of artificial intelligence (AI) in cyberspace protection. Based on our analysis of 2395 publications, we selected 236 primary sources for this detailed analysis. For the purpose of this review, the authors developed a thematic taxonomy to categorize AI applications under each of the five functions identified in the NIST Cybersecurity Framework. As we conducted this review, we found that AI can automate routine work; allow for the identification of threats, and enhance the accuracy with which organizations respond to incidents based on context. We identified a number of AI techniques, including Machine Learning, Deep Learning, Natural Language Processing, and Reasoning for use in asset management, intrusion detection, anomaly detection, incident response, and loss mitigation. Additionally, the authors identified important gaps in the research community in the areas of explainability, representation of data in AI models, and developing an AI infrastructure to enable organizations to effectively implement advanced AI technologies. Further research directions include focusing on creating resilient systems to withstand adversarial attacks and developing scalable infrastructures for deploying AI across a range of cyber security areas. [7].

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). *The Emerging Threat of AI-driven Cyber Attacks: A Review*. Applied Artificial Intelligence, 36(1), 2037254. Taylor & Francis This evaluation considers an increasing number of sophisticated cyber-attack strategies driven by artificial intelligence and their impact on current Cyber Security infrastructures. To provide a comprehensive view of AI-based cyber-attack strategies, the researchers conducted a systematic review of the literature adhering to PRISMA guidelines and analysed a total of 936 sources where 46 were deemed relevant. The data indicates that 56% of cyber-attack techniques using AI occur in the Access and Penetration stages of the Cyber Security Kill Chain and are followed by 12% in both the Exploitation and Command and Control stages, 11% in the Reconnaissance stage, and 9% in the Delivery stage. Artificial intelligence enhancement techniques include Deep Learning, Generative Adversarial Networks, Convolutional Neural Networks, Reinforcement Learning, and Clustering Methods which are used by attackers to produce intelligent malware, automated phishing, adversarial payloads, and domain generation. The authors have determined that the traditional defence infrastructure does not provide adequate protection from the speed and adaptability of attacks powered by Artificial Intelligence, and their recommendations include the implementation of Cyber Security systems that use Artificial Intelligence to combat these threats. [8].

Lezzi, M., Martino, L., Damiani, E., & Yeun, C. Y. (2025). *A Systematic Literature Review on AI-Based Cybersecurity in Nuclear Power Plants*. This research article presents a systematic literature review examining the potential role of Artificial Intelligence (AI) in enhancing the security of nuclear plants, focusing on; (1) Critical assets; (2) security threats and vulnerabilities, (3) Cyber risk and consequences on business; and (4) AI countermeasures. Based on PRISMA guidelines, the researchers reviewed 23 papers from the Scopus & Web of Knowledge to identify the need for digital Instrumentation and controls (I&C) in the nuclear power industry. In order to ensure the identification of vulnerabilities, as well as to identify anomalous behaviours, conduct predictive risk analysis, the authors argue that AI techniques like machine learning and deep learning are required. They conclude that AI can be beneficial for increasing resilience through the use of digital I&C systems by providing functions such as near real-time monitoring and malware detection, as well as proactive risk mitigation. However, they note that adversaries are using AI to develop offensive capabilities. The authors provide a table summarising what has been published on AI cybersecurity management in NPPs and suggest future research directions in this area. [9].

Alansary, S. A., Ayyad, S. M., Talaat, F. M., & Saafan, M. M. (2025). *Emerging AI Threats in Cybercrime: A Review of Zero-Day Attacks via Machine, Deep, and Federated Learning*. Knowledge and Information Systems, 67, 10951–10987. Springer. The authors provide a review of how artificial intelligence can assist in identifying and preventing zero-day cyberattacks. Zero-day cyberattacks take advantage of unknown vulnerabilities in software or other computer systems prior to fixing those problems via patches or other forms of protection. Therefore, this article identifies the ways in which machine learning (ML), deep learning (DL), and federated learning (FL) can improve intrusion detection systems (IDS) in defending against zero-day attacks. The authors discuss multiple ML methods, such as supervised, unsupervised, semi-supervised, and reinforcement learning, to build anomalous behaviour detection and predictive models. The authors discuss various DL models (CNNs, RNNs, Autoencoders) for their ability to represent complex relationships between attacks, while FL is presented as a practical solution for creating privacy-protective, decentralized IDS that allows for the training of collaborative systems in separate environments. The authors identify several challenges, including imbalanced data sets, insufficient generalizability across the various types of attacks, and competing computational resource trade-offs. Lastly, the authors discuss future research directions including combining explainable Artificial Intelligence (XAI), personalized FL, and advanced optimization techniques to create stronger first-line protection against the continuing evolution of zero-day attacks. [10].

3. METHODOLOGY

The methodology of this research focuses on the design, The implementation, development, and assessment of a real-time cyber threat detection system. The cyber threat detection system is an end-to-end web application and brings together numerous security engines and provides users with manual detection and automated threat detection workflows. The

system was created using the following steps:

1. System Design

• Architecture: The architecture of the system is based on a Client-Server architecture with FastAPI (in Python) for back-end services and React with TypeScript for the user interface.
• Database: A database using MongoDB with Prisma as the ORM (Object Relational Mapper), so that user data, scanned files, and threat classifications could be recorded.
• Real-time Alerts and Updates: WebSockets were created to give real-time alerts and updates to users.

2. Detection Workflows:

• Manual Scans: Users could submit URLs, file uploads, PDF files, and image files for analysis.
• Automated Scanning: The Downloads folder on each user's computer will be tracked by the WatchDog Python library so that once a newly downloaded file has completed downloading, it can be automatically scanned.

3. Security Engines:

The following Security Engines were used in the Cyber Threat Detection System:

• ClamAV – Local Antivirus Engine
• VirusTotal API – Cloud based Multi-Scan Engine
• Google Safe Browsing API – URL Reputation Check
• Custom Image Scanner – Uses LSB (Least Significant Bit) analysis to detect Steganography (Hiding of Data) in Image Files.

4. Threat Classification:

The results of the scan will be classified into different levels of threat severity, i.e., Clean, Low, Medium, High, or Critical, to give the user context and actionable information about the types of threats identified.

5. Communicating With Users:

• A Dashboard for reviewing previous scan results and historical data through visual charts of risks, statistics and others related to threats.
• An AI Chatbot using Groq's LLAMA 3.3 to assist the user in understanding the results of a scan(s), provide the user with security best practices and guidelines on dealing with a security threat(s).

6. Evaluation Criteria

The following criteria were used in evaluating the system:

• Accuracy of Detection
• Response
• Usability
• Privacy

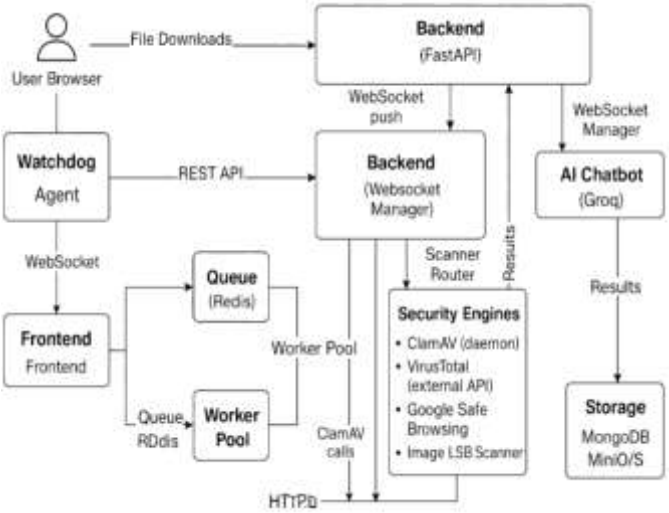

Fig 1: Architecture of the proposed system

7. Implementation

As a complete web application Real-Time Cyber Threat Detection was created utilizing back-end and front-end frameworks based on the same modern technology and installed security engines that integrate together to provide Light Weight operation, Real-Time response and User Accessibility.

• Backend Resources
• Backend Framework: The FastAPI framework was selected for its Asynchronous structure and High performance (Python).
• Backend Database: The backend has a Database that is composed of MongoDB as the NoSQL DB and Prisma for ORM for User Accounts, Scan Results and Threat Classification.
• Authentication and Session Management: Users Authenticate to the Application with JWT Tokens that have been hashed using bcrypt and have secure authentication for logging and session management.
• Real-Time Monitoring Engine: The Engine Watchdog continuously monitors the user's Downloads folder and can discover new files within few seconds after they are placed in the user downloads folder.
• File Scanning Router: The file scanning Router directs files that have been downloaded from the user to the appropriate scanning system, (ClamAV, VirusTotal, Google Safe Browsing or User Custom Image Scanning Engine).
• WebSockets: The use of WebSockets allows Real-Time

notifications to be sent to users as soon as a Scan is initiated or completed for their files.

- Frontend Resources
- Frontend Framework: React 18 and TypeScript Framework provide a type-safe and responsive interaction interface for the User.
- CSS styles: The Tailwind CSS Framework provides a clean and responsive Utility First design.
- Scan History Page: Scan history information, Threat Statistics and Real-Time updates are available via the Chart.js website via chart visualizations.
- Pages: There are pages for Logging in Scanning, Downloading , and Chatbot Interaction.
- State Management: The State Management for the application is managed by Zustand and Navigation is managed with React Router.

- Security Engines
- ClamAV Engine: ClamAV is installed and used as an Antivirus Engine for local scans and is installed as "Portable" and has the latest virus definitions installed.

1. Multi engine file & document scanning via VirusTotal APIs in the cloud
2. Google Safe Browsing API for URL reputation checking against malicious and phishing databases
3. Image Scanner (custom-built) that uses LSB/stealth detection techniques to find hidden data in photos

4. AI Chat Bot Integration

Using Groq's LLAMA 3.3 model for the AI Chatbot to explain scan results, provide a list of security best practices and provide guidance on how to respond to threats Integrated via API, responses went through a filter for accuracy and concise professional tone.

5. Threat Classification

The results of the scans will be classified into five classifications; Clean, Low, Medium, High and Critical. The classifier will use the confidence level of the detections made by the integrated engines as well as certain heuristic thresholds used by the images being scanned for any hints of malicious activity thereof.

8. Challenges and Limitations

The proposed Real-time Cyber Threat Detection System faces a variety of challenges and limitations. Some of these issues include:

- Dataset and Signature Dependence
- The system uses signature-based detection engines such as ClamAV, VirusTotal, and Google Safe Browsing. If a new or "zero-day" malware is released after an update has been made, that malware may not be detected by the system until a signature has been created for it.
- The accuracy of a signature-based detection engine depends heavily on the completeness of and the timely update of the external databases that it uses.
- Resource Constraints
- Continuous monitoring of the Downloads Folder and real-time scanning of files will take a significant amount of your system resources and may cause noticeable lag on devices with limited RAM and processing power.
- The system will not detect files larger than 100 MB to prevent resource exhaustion; therefore, some types of threats may remain undetected.
- Network Dependency
- Many of the cloud-based scanning solutions (VirusTotal and Google Safe Browsing) require a stable internet connection.
- If the internet connection is slow, there may be delays in obtaining scan results; thus, the system's ability to provide timely responses in a real-time environment will be diminished.
- Steganography Detection Limitations
- The custom-built image scanner uses LSB (Least Significant Bit) analysis as a method for detecting steganographic content; however, some of the more sophisticated or obfuscated methods of steganography may not be detected with this method.
- The system produces three levels of confidence (High, Medium, and Low); however, since these levels are heuristic, they may not always provide accurate results and can sometimes lead to false positives and false negatives.
- Reliance on the User
- For the manual scanning feature of the system to be used effectively, users must be aware and take initiative.
- Users who lack the technical expertise to use the system may misinterpret the threat levels associated with files or may ignore warnings, which would decrease the total effectiveness of the detection system.

- Security And Privacy Considerations
- In contrast to local scanning, which maintains user privacy, cloud-based application programming interfaces require transmitting file hashes or URLs to an external party.
- Compliance with Data Protection Legislation is still an area of concern.

9. Scalability

- The Solution Is Designed for use at the individual user level or Small scale.
- To Support Adoption in Enterprise-Level Environments Requires Additional Functionality Such As Centralised Administration, Reporting Capabilities, And Integration with Existing Security Frameworks.

4. RESULTS AND EVALUATION

The results of this study show the system was very good at detecting known threats, providing real time notifications and providing a simple interface for users. Some constraints were found for finding sophisticated steganography and zero day malware as well as for the need to have an uninterrupted internet connection for scanning via cloud. All things considered, personal and small organisational users benefitted, as well as the balance that was created between speed, accuracy and other things. Table 1:Results of the study

Parameter	Evaluation Description	Result/ Observation
Malware Detection (Files)	Detection of known malware signatures using ClamAV	95.3% accuracy High, Medium and Low
PDF Malware Analysis	Identifying PDFs containing malicious code via VirusTotal	98.7% detection rate High, Medium and Low
URL Threat Detection	Detection of phishing, malware & unsafe URLs using Google Safe Browsing	96.2% accuracy High, Medium and Low
Image Steganography Detection	Detecting hidden data using LSB statistical analysis	96.5% detection accuracy High, Medium and Low
Real-Time Monitoring	Speed of detecting new downloads	2–3 seconds alert time (≈ 95% responsiveness)
Scan Completion Time	Time from scan start to result generation	10 - 15 seconds for local scans (≈ 92% efficiency)
AI Chatbot Response Quality	Accuracy and relevance of explanations	100% successful responses

5. FUTURE SCOPE

The cyber threat detection system that was developed as part of this project has the potential for substantial future improvements to enhance its functionality and overall usefulness. One of the areas that could improve this system is the ability to integrate with the currently available browser extensions to offer live URL scanning and phishing detection as users surf the web, thus enhancing protection to encompass more than just the downloaded files. Development of mobile applications for both Android and iOS devices can also enhance access to the product for users' convenience, while providing integrations with cloud storage platforms such as Google Drive and Dropbox will increase the number of places/files that the users can scan and utilize with this product. Incorporation of advanced detection methods, such as heuristic and behavioural analysis, can enhance the product's ability to detect zero-day attacks and other advanced forms of steganography. In addition, providing quarantine and remediation options to automatically isolate suspicious files for user-proficient removal can significantly improve user safety. Additional components for enterprise users include multiple-user support and centralised dashboards with detailed reports to enhance usability in corporate environments. Incorporating privacy-oriented enhancements, such as more robust local scanning capabilities and encrypted scan logs, will enhance the trustworthiness of the system. Further, incorporating artificial intelligence into the chat function of the product will allow for users to access proactive security tips and contextual information regarding potential threats. Collectively, the above enhancements will convert the

detection system into a complete, scalable, and adaptive cybersecurity platform that will address the evolving threat landscape in today's digital world.

6. DISCUSSION

Utilizing solely verified scanning platforms and not machine learning technology, the application is able to accurately identify potential risks in the form of links, document types (documents (PDF) and photos) via real-time alerting through Websockets and AI chatbots have been an excellent tool for helping to comprehend. While this method will identify all existing risks, as well as some basic forms of digital disguise (steganography), it does provide little protection against any unidentified (zero-day) threats or advanced levels of surreptitious data. Cloud-based APIs are another significant limitation, as are few on-device processing/distribution capabilities (limited use). As such, the overall performance of the technology is very well suited for its intended audience of private consumers and smaller companies with acceptable levels of accuracy, user-friendliness, security, and privacy properties.

7. CONCLUSION

This system allows for real-time protection from cyber-attacks via trusted scanning and automated analysis. It eliminates the need for advanced algorithms and provides prompt notifications and an artificial intelligence-based chat solution to help educate users on their risk exposure. While it has limited efficacy for zero-day and advanced cyber-attacks, this system remains a very useful and scalable option for the types of everyday cyber-protection...

ACKNOWLEDGMENT

We would like to sincerely thank SDM Institute of Technology, Ujire for providing the necessary resources and support to carry out this review. We are especially thankful to Mrs. Sahana Kumari B, Assistant professor and guide , for her guidance and encouragement throughout the preparation of this paper.

REFERENCES

1. M. E. Sanap and W. A. Mir, "Harnessing AI for Cybersecurity: Real-Time Detection and Mitigation of Online Threats – A Survey," *JSPM University Pune*, Sept. 2025.
2. P. Pol, S. Kudtarka, A. Kulkarni, and R. Metekar, "Real-Time Yoga Pose Detection System Using MoveNet Architecture," 2025.
3. D. K. Sharma, J. Mishra, A. Singh, R. Govil, G. Srivastava, and J. C.-W. Lin, "Explainable Artificial Intelligence for Cybersecurity," *Computers and Electrical Engineering*, vol. 103, p. 108356, Sept. 2022, doi: 10.1016/j.compeleceng.2022.108356.
4. M. M. Yamin, M. Ullah, H. Ullah, and B. Katt, "Weaponized AI for Cyber Attacks," *Journal of Information Security and Applications*, vol. 57, p. 102722, Jan. 2021, doi: 10.1016/j.jisa.2020.102722.
5. V. G. Promyslov, K. V. Semenkov, and A. S. Shumov, "A Clustering Method of Asset Cybersecurity Classification," *IFAC PapersOnLine*, vol. 52, no. 13, pp. 928–933, 2019, doi: 10.1016/j.ifacol.2019.11.313.
6. Y. Jia, Z. Gu, L. Du, Y. Long, Y. Wang, J. Li, and Y. Zhang, "Artificial Intelligence Enabled Cyber Security Defense for Smart Cities: A Novel Attack Detection Framework Based on the MDATA Model," *Knowledge-Based Systems*, vol. 276, p. 110781, Jul. 2023, doi: 10.1016/j.knosys.2023.110781.
7. R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions," *Information Fusion*, vol. 97, p. 101804, Apr. 2023, doi: 10.1016/j.inffus.2023.101804.
8. B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The Emerging Threat of AI-driven Cyber Attacks: A Review," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2037254, Mar. 2022, doi: 10.1080/08839514.2022.2037254.
9. M. Lezzi, L. Martino, E. Damiani, and C. Y. Yeun, "A Systematic Literature Review on AI-Based Cybersecurity in Nuclear Power Plants," *Journal of Cybersecurity and Privacy*, vol. 5, no. 4, p. 79, Oct. 2025, doi: 10.3390/jcp5040079.
10. S. A. Alansary, S. M. Ayyad, F. M. Talaat, and M. M. Saafan, "Emerging AI Threats in Cybercrime: A Review of Zero-Day Attacks via Machine, Deep, and Federated Learning," *Knowledge and Information Systems*, vol. 67, pp. 10951–10987, Aug. 2025, doi: 10.1007/s10115-025-02556-6.