e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 6, Issue 3, March 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.54

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly, Peer Reviewed & Referred Journal|

| Volume 6, Issue 3, March 2023 |

| DOI:10.15680/IJMRSET.2023.0603034|

Securing E-Commerce Platforms using Log BERT: A Self-Supervised Transformer-Based Anomaly Detection Approach in Cloud Environments

Farid Hidayat

Universitas Andalas Padang, West Sumatra, Indonesia

faridhidyat51@gmail.com

ABSTRACT: The ever-increasing growth of e-commerce platforms has created a heightened demand for well, scalable, intelligent cybersecurity solutions for protecting sensitive data and maintaining system integrity. Traditional anomaly detection mechanisms, being mostly rule-based and so supervised, have been inefficient in dynamic cloud environments, generally failing to generalize unseen patterns; they possess high false alarm rates since much of their operation relies on labelled datasets. These weaknesses are fully addressed in this article through the presentation of a model for self-supervised transformer-based anomaly detection-Log BERT-targeted for multi-cloud e-commerce environments. This framework utilizes transformer architecture to model complex, temporal dependencies inherent in unstructured log data without requiring any form of labelling of the input data. Log sequences are generated through session-based sliding windows and enriched with sinusoidal positional encodings; this masked language modelling (MLM) serves as the pretext task required for the model to learn contextual representations. Anomalies are eventually scored using negative log-likelihood and further classified via One-class Support Vector Machine (OC-SVM) for high-precision outlier detection. The entire detection pipeline is securely deployed in the cloud using AES-256 encryption in conjunction with a Cloud Key Management Service (KMS), thereby ensuring compliance with data protection standards. The proposed system shows considerable performance improvements in inference latency, detection accuracy and storage efficiency at decreased dependence on manual rule creation or data labelling. This is a highly intelligent and secure scalable solution to a major problem in modernizing the security of an e-commerce infrastructure against evolving threats faced in distributed cloud-native settings.

KEYWORDS: Log Bidirectional Encoder Representations from Transformers, Anomaly Detection, Masked Language Modelling, Advanced Encryption Standard-256, One-class Support Vector Machine.

I. INTRODUCTION

E-commerce platforms develop exponentially, giving rise to an upsurge in online transactions and user activities [1]. The rapid growth of e-commerce has turned the security and reliability of their services into a priority concern. Such systems generate operational and transactional logs that actually penetrate gigabytes of data into their severed heads, which becomes very critical for monitoring performance and detecting anomalies. Traditional methods of monitoring and securing such environments could not scale or adapt to new threats [2]. E-commerce applications could well avail from the promise of cloud computing by hosting its applications on the cloud. This, however, brings about new complexities with regards to log management and anomaly detection. Attackers exploit these gaps, defrauding the system by displaying malicious activities under the umbrella of bona fide-looking logs. Manual review of such logs would be highly impractical as the data volume and the rate at which it is generated preclude such things. This type of requirement therefore insists on a highly intelligent, fully automated, working almost all the time in the cloud environments [3]. It requires the machine learning realm, mainly self-supervised models, for promising quality in anomaly detection, thus bearing the need for modern scalable, intelligent, and dynamic-log analysis systems functional in cloud environments. Mohanarangan Veerappermal Devarajan (2022) suggested a neural network algorithm that improves workload forecasting in cloud computing by enhancing resource allocation and system scalability through accurate workload variation prediction. This approach inspires the proposed work by demonstrating how neural networks can dynamically predict workloads, motivating the development of more efficient algorithms for optimized cloud resource management [4].

The very character of e-commerce platforms often contributes to the severity of the challenge posed to anomaly detection. An outstanding aspect of the problem statement is the considerable amount of unstructured log data produced from a variety of services across cloud node. Conventional rule-based systems lack the very flexibility to detect complex or never-before-seen forms of attack behavior [5]. Furthermore, there are a lot of different formats and structures for log data, which leads to problems such as integrating and correlating information. Existing systems also suffer from silos: they do not allow users to have a holistic view of activity across services. Often, fast scans also make considerable demands on computing power, which is typically limited in distributed cloud systems. Often found are false alarms because these traditional methodologies lack contextual constructs pertaining to sequences of logs. Attackers can take on the persona of an average normal person frequently, which negates any detection through signature-based means. Data labelling under the supervised type of learning is not only tedious but hardly possible for most anomaly scenarios [6]. Last but not least, scattered intelligence in a multi cloud environment creates gaps in a security monitoring standpoint.

Anomaly detection systems in e-commerce environments are often limited due to their reliance on labeled data and static, predefined rules. Many traditional methods require extensive domain knowledge and manual intervention to detect patterns, making them inefficient and inflexible. Supervised models, while effective when trained on comprehensive datasets, struggle in situations where security events are rare or infrequent, making it difficult to capture anomalies. Additionally, most existing models fail to handle varying log formats and semantics, hindering their ability to transfer log signatures across different systems. Classical statistical methods also fall short in capturing the temporal relationships within log sequences, which are crucial for identifying anomalies over time [7]. Signature-based systems are limited to detecting known threats and are ineffective against new or zero-day attacks. Many of the current detection tools are also not designed to scale for processing real-time log streams, especially in cloud-native environments, leading to inefficiencies. Furthermore, false positives and negatives continue to be a significant issue, reducing the reliability of these systems. Additionally, the lack of proper integration with cloud infrastructure results in inefficient resource utilization for data ingestion. These limitations highlight the need for adaptive, scalable, and intelligent anomaly detection models capable of addressing the evolving security challenges faced by modern e-commerce platforms. This study explores quantum internet technologies like Quantum Key Distribution (QKD) and entanglement-based communication to secure patient data in healthcare, providing stronger security measures. Kalyan Gattupalli (2022) illustrates how these quantum technologies tackle healthcare data security challenges, which drives the proposed approach to develop more advanced and robust data protection mechanisms [8].

To address the limitations of existing anomaly detection systems, we propose a self-supervised transformer-based framework tailored for cloud-based e-commerce environments. This framework utilizes Log BERT, a transformer model specifically designed to learn contextual patterns from system logs without the need for labelled data. Anomaly detection in this model is based on reconstruction errors or prediction confidence scores, allowing it to identify abnormal behavior without relying on predefined labels. The model's cloud-native architecture ensures scalability and makes it well-suited for multi-cloud deployments, providing flexibility across different environments. Advanced positional encoding is employed to preserve the order and context of log sequences, enhancing the model's ability to detect anomalies over time. Additionally, the integration of AES-256 encryption and Cloud Key Management Services ensures that the data is securely deployed and stored, protecting sensitive information. To further improve detection precision, the framework incorporates one-class SVM and statistical thresholding techniques, which enhance accuracy in distinguishing between normal and anomalous behavior. This approach minimizes the dependency on labelled data and boosts detection accuracy in dynamic, evolving environments. Ultimately, the proposed solution offers a robust, intelligent, and scalable framework for securing cloud e-commerce platforms, addressing key challenges in anomaly detection while maintaining high levels of data security and performance.

In Section 2, the Literature Review provides an overview of existing anomaly detection methods, highlighting their limitations and challenges. These insights serve as the foundation for the proposed approach. Section 3 shifts focus to the specific challenges faced by deep learning methods in e-commerce platforms, particularly in the context of analysing customer feedback. Section 4 introduces the proposed methodology for privacy-preserving anomaly detection, using a combination of transformers and One-Class SVM in cloud systems. This section outlines how the model can address key security and scalability concerns while ensuring effective anomaly detection without the need for labelled data. Following this, Section 5 presents the results and discussions, analysing the performance of the proposed model in various test scenarios and comparing it to existing approaches. Section 6 concludes the paper, summarizing the key findings and suggesting potential directions for future research and improvements in cloud-based anomaly detection systems for e-commerce platforms. This structure provides a comprehensive approach to tackling the challenges in anomaly detection while addressing the growing need for scalable, secure, and efficient solutions in modern cloud environments.

II. LITERATURE REVIEW

Belghith [9] suggested that AI and ML techniques, including Random Forest, Decision Tree, and ANN, are popular in churn prediction in cloud-based CRM systems. Since ensemble methods such as Random Forests are accurate, the complex models may sacrifice interpretability for computation. Therefore, it may be worth looking for hybrid models that balance the two issues in future studies. ML techniques such as Random Forest and ANN assist in the prediction of churn in CRM. These techniques need quality data and can become very complicated and less interpretable. Ensemble methods would offer good performance, but a fair compromise between performance and a simple explanation is still a challenge.

Luo & Choi [10] remarked that employee engagement strategies such as participative management can positively impact retention when the perception of fair compensation is there. Regression and moderation analysis are popular but generalizability can be compromised by cultural factors and survey biases. According to his analysis, outlines how ML can be thought of as supporting HR tasks involving hiring and retaining people with data-based insights and methods, such as case studies and surveys. Limitations include the steep learning curve for HR professionals in becoming data-skilled and the risk of over-relying on automated decisions. The task integrates adaptive wavelet transforms with wearable IoT sensors to boost the precise assessment of children's health, enabling reliable vital sign tracking. They, shows how advanced signal processing improves health data analysis, which spurs the proposed method to develop more efficient health assessment solutions, as exhibited by Sri Harsha Grandhi (2022) [11].

Fauziyah et al suggested [12] that it is AI that would increase efficiency and integrity in the recruitment process through automation of the resume screening and verification of credentials via blockchain. The study utilized qualitative interviews of HR professionals. Limitations include barriers in the adoption of technology and data privacy issues in decentralized systems. An explorative looked into AI and ML frameworks, including reinforcement learning and predictive analytics, for the optimization of workforce management including hiring, scheduling, and performance evaluations. While these methods improve operational efficiency, limitations include bias considerations, privacy issues, and system integration.

Santoso suggested [13] electronic commerce (E-commerce) plays a crucial role in today’s environment, with goods and services being provided through computer networks. While E-commerce offers significant advantages, it requires costly hardware and software for implementation, and the increasing volume of electronic data raises the need for more resources, creating challenges in efficient IT resource utilization. Although cloud computing supports the development and implementation of E-commerce infrastructure, issues related to system security and stability remain a concern. This paper proposes a model for E-commerce using cloud computing to address these challenges and improve data security in E-commerce applications.

Vargas suggested [14] Virtualization is essential for cloud-based e-commerce platforms, enabling flexible resource allocation, multi-tenant hosting, and fast application deployment. For high-assurance e-commerce transactions, robust security controls must extend to virtualized layers to protect sensitive data during dynamic scaling and cross-region failovers. Safeguards like hypervisor security, container orchestration, and virtualization hardening, including secure boot, kernel integrity checks, and memory encryption, are crucial to prevent attacks. Security measures reduce the risk of lateral movement by adversaries and ensure transaction integrity. This paper examines how virtualization practices, such as hypervisor selection and workload confinement, support secure multi-cloud e-commerce environments.

III. PROBLEM STATEMENT

AI and Machine Learning (ML) techniques are increasingly utilized in cloud-based Customer Relationship Management (CRM) systems for tasks like churn prediction and workforce management. However, one of the key challenges is balancing model accuracy with interpretability. While models such as Random Forests and Artificial Neural Networks (ANN) often deliver improved performance, they typically lack transparency, making it difficult for non-technical users to understand and trust their decisions [15]. This becomes especially critical in applications that require clear explanations of the decision-making process, such as customer service or policy enforcement. Furthermore, domain-specific adaptation is essential to ensure that AI and ML models are not only accurate but also relevant and effective for particular industries or tasks. Quality control is also a vital factor in ensuring that the models perform consistently well in real-world applications. Without this focus on interpretability, domain adaptation, and quality control, AI and ML solutions may fail to gain widespread adoption or effectively support decision-making processes in CRM systems.

The addition to technical challenges, the broader implementation of AI in Human Resources (HR) processes is hindered by barriers such as technology adoption, privacy concerns, and difficulties integrating with existing systems. While blockchain technology ensures data integrity in recruitment, AI plays a crucial role in streamlining and improving the efficiency of HR operations. However, organizations face obstacles in adopting AI-driven solutions, as these technologies must be secure, adaptable, and interpretable. Previous research suggests that hybrid approaches combining the predictive power of AI and ML with user-friendly designs and ethical considerations are essential [16]. As businesses increasingly depend on AI and ML technologies, the demand for models that are not only effective but also interpretable, secure, and adaptable has grown significantly. Meeting these needs is crucial for ensuring that AI adoption in HR processes aligns with modern business requirements while addressing concerns over data security and privacy. Durga Praveen Deevi (2022) investigates neuromorphic and bio-inspired computing to enhance smart healthcare architectures, improving adaptability, efficiency, and decision-making. This study enlightens the proposed system by exhibiting how bio-inspired computing optimizes healthcare systems, which influences the development of more responsive, scalable, and smart healthcare architectures. [17].

IV. PROPOSED FRAMEWORK FOR PRIVACY-PRESERVING ANOMALY DETECTION USING TRANSFORMERS AND ONE-CLASS SVM IN CLOUD SYSTEMS

Anomaly detection systems in cloud environments leverage AI technologies to ensure security, with the workflow starting at the Data Collection phase. At this stage, log data is gathered from multiple sources, providing a comprehensive view of system activities. This data is then processed using a Transformer-based Language Modelling framework, enabling the model to identify patterns and contextual relationships within the log sequences. Positional Encoding is applied during Log Sequence Generation to preserve the order of events, ensuring that the model can effectively detect deviations from normal behavior over time. This capability is crucial for identifying anomalies that occur as sequences evolve [18]. To maintain data security and comply with privacy standards, the log data is encrypted using AES-256 and managed by Key Management Services (KMS) in the cloud. This encryption ensures the confidentiality and integrity of the data, protecting it from unauthorized access while allowing for secure processing in the anomaly detection system. The entire process, from data collection to secure processing and anomaly detection, is depicted in Figure (1), illustrating how AI-driven workflows can enhance security in cloud environments by combining advanced modelling techniques with robust data protection.

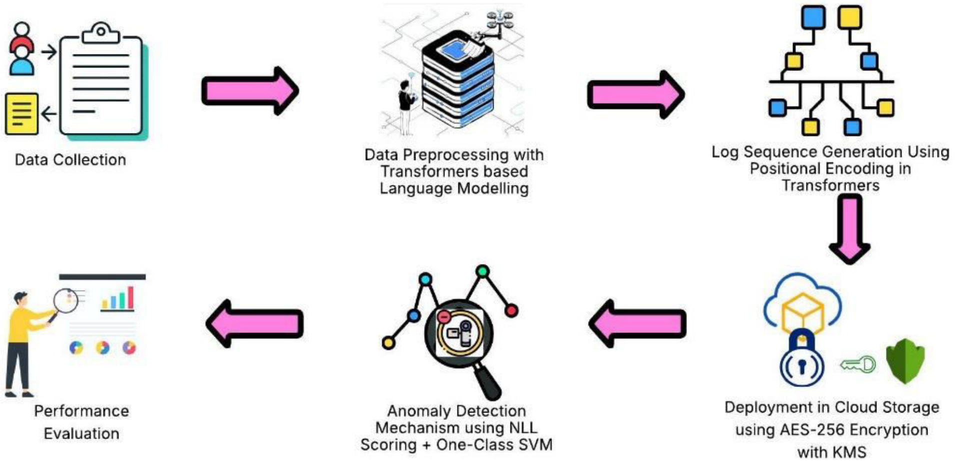
Data Collection
Data Preprocessing with Transformers based Language Modelling
Log Sequence Generation Using Positional Encoding in Transformers
Deployment in Cloud Storage using AES-256 Encryption with KMS
Anomaly Detection Mechanism using NLL Scoring + One-Class SVM
Performance Evaluation

Figure 1: Block Diagram of Privacy-Preserving Anomaly Detection Using Transformers and One-Class SVM in Cloud Systems

The proposed anomaly detection system uses a hybrid mechanism that combines Negative Log-Likelihood (NLL) scoring with One-Class SVM techniques to effectively detect both known and unknown anomalies. This hybrid approach excels in identifying deviations from normal behavior while maintaining low false positive rates, even when faced with complex threats [19]. The system's ability to detect subtle anomalies makes it particularly useful for environments where security and accuracy are critical. Performance evaluation measures such as accuracy, detection time, and scalability ensure that the model operates efficiently under varying loads. These evaluations allow for optimization of the system's responsiveness, making it adaptable for real-time processing in cloud infrastructures [20]. The scalable and secure pipeline, developed through this evaluation, is particularly suitable for sectors that handle sensitive information, such as banking, healthcare, and enterprise operations. By combining advanced anomaly detection with robust security protocols, the system provides a high level of protection while enabling fast, accurate detection of threats in cloud-based environments. This makes it a valuable tool for safeguarding sensitive data and ensuring compliance with privacy and security standards. This combine flexible DBSCAN with federated learning for resilient, federated, and privacy-preserving exception detection in IoT networks. Naresh Kumar Reddy Panga (2020) procedure sparks the proposed model to integrate transformer-based exception detection with safe, cloud-native architectures, enhancing adaptability and data control in e-commerce environments by leveraging federated principles for efficient and secure anomaly identification [21].

4.1 Data Collection

The customer feedback dataset provides a comprehensive resource for analysing customer sentiments across multiple platforms such as social media and online reviews. It includes essential metadata such as sentiment tags (positive or negative), date, user ID, location, and a confidence score for each entry, offering a well-rounded view of customer interactions [22]. This dataset is designed to support various text analysis techniques, particularly sentiment analysis, which helps evaluate the overall sentiment of customer feedback. Additionally, it supports text preprocessing for cleaning and preparing the data for further analysis [23]. The dataset also enables topic modelling, uncovering key themes or trends in customer opinions, which is valuable for identifying common concerns or areas of interest. Furthermore, it is highly applicable for machine learning tasks, including predictive modelling, which can yield deeper insights into customer behavior and preferences [24]. By utilizing this dataset, organizations can better understand what customers think, pinpoint areas for improvement, and enhance their products or services. This dataset's rich structure makes it an ideal tool for extracting meaningful patterns and trends from raw feedback, enabling businesses to make data-driven decisions and improve customer satisfaction. The study analyses threat models in vehicular cloud computing, highlighting security and privacy challenges in dynamic environments. This approach prompts the proposed technique to incorporate robust threat modelling and privacy-preserving mechanisms in cloud-based e-commerce setups, ensuring secure and decentralized architectures for anomaly detection. This linkage is confirmed by Sreekar Peddi (2021) [25].

Dataset Link: https://www.kaggle.com/datasets/vishweshsalodkar/customer-feedback-dataset.

4.2 Data Preprocessing with Transformers based Language Modelling

A transformer-based language model, like Log BERT, leverages a multidimensional approach to preprocess and model log or sentiment information without needing manual feature extraction. The process begins with various preprocessing steps, including log parsing, tokenization, noise filtering, and sequence comprehension. Unlike traditional models, Log BERT directly learns representations from raw text through self-attention mechanisms. Given a sequence of tokens $X = \{x_1, x_2, \ldots, x_n\}$, the model calculates contextual embeddings for each token by assigning attention scores that determine how much focus each token should have on others within the sequence. This attention mechanism allows the model to capture intricate relationships and dependencies in the data, providing more accurate and flexible insights for tasks like sentiment analysis or log interpretation. The attention scores, defined by specific mathematical formulations, guide the model in understanding the context and meaning of each token in relation to the others, enhancing its ability to process complex sequences.

$$\text{Attention}\ (Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where Q, K, V are the query, key, and value matrices derived from the input tokens and dk is the dimensionality. The model predicts a masked token x_m using the context $X_{\backslash m}$ in the self-supervised learning MLM objective, and the optimisation loss is given by Eq. (2),

$$\mathcal{L}_{MLM} = -\sum_{m \in M} \log\ P\left(x_m \mid X_{\backslash m}\right) \tag{2}$$

The normal patterns within log sequences, the Log BERT model becomes highly effective in performing downstream tasks like anomaly detection and sentiment classification. This ability to capture sequence patterns allows the model to identify deviations from expected behavior, making it adept at detecting unusual events or sentiments.

Fine-tuning the model enables it to adapt to specific domain patterns, improving its accuracy in context-specific tasks. This approach also eliminates the need for repeated preprocessing steps, as Log BERT can directly learn from raw data in a way that is more efficient and effective [26]. By leveraging the model's capacity to learn these domain-specific nuances, organizations can apply it to a variety of use cases without the complexity of manually tailored preprocessing, thereby streamlining the process and enhancing performance.

4.3 Log Sequence Generation Using Positional Encoding in Transformers

Data preparation for anomaly detection using the Log BERT transformer-based model involves a complex methodology that includes sliding window segmentation, session-based sequence modelling, and positional encoding. The sequence of log events in increasing time order is represented as $\mathrm{L} = \{\mathrm{l}_1, \mathrm{l}_2, \dots, \mathrm{l}_\mathrm{T}\}$, where T is the total number of events. To preserve temporal dependencies within the data, the sequence is divided into smaller, overlapping subsequences using a fixed-size sliding window of length w. This sliding window approach ensures that each subsequence captures a relevant context of log events while maintaining the temporal flow of the original sequence [27]. The overlapping windows allow for continuity between subsequences, which is crucial for understanding patterns and detecting anomalies over time. Additionally, positional encoding is applied to ensure the model can differentiate between the order of events within the sequence, enhancing its ability to identify anomalies based on both the sequence structure and the time-based relationships between events.

$$\mathrm{S}_\mathrm{i} = \{\mathrm{l}_\mathrm{i}, \mathrm{l}_{\mathrm{i}+1}, \dots, \mathrm{l}_{\mathrm{i}+\mathrm{w}-1}\}, \text{ for } \mathrm{i} = 1 \text{ to } \mathrm{T} - \mathrm{w} + 1 \tag{3}$$

To improve the model's ability to understand semantic contexts, log data is grouped into sessions based on identifiers like user ID, IP address, and transaction ID. This segmentation ensures that related log events are treated as a cohesive unit, providing the model with a clearer understanding of user behaviours or transaction flows. By grouping events within these sessions, the model can more accurately detect anomalies that are contextually linked to specific users or transactions. This method allows the transformer model to capture patterns specific to each session, enhancing its ability to differentiate between normal and anomalous activities. The resulting data segmentation, as defined in Eq. (4), helps the model recognize subtle deviations from expected behavior, leading to more precise anomaly detection.

$$\text{Session}_{\mathrm{k}} = \{\mathrm{S}_{\mathrm{i}_1}, \mathrm{S}_{\mathrm{i}_2}, \dots, \mathrm{S}_{\mathrm{i}_\mathrm{n}}\}, \text{ where } \mathrm{k} \in \text{ All sessions} \tag{4}$$

The Log BERT model learns behavioural patterns that evolve over time across multiple users or transactions by capturing temporal dependencies in the log data. Since transformers lack inherent positional bias, positional encoding is applied to each token to maintain the sequence order. Subramanyam Boyapati (2022) conducts an exploration of how virtual platform processing and internet-inclusive finance reduce the urban rural income gap in the e-commerce era, highlighting technology's role in economic inclusion. This exploration notifies the presented work by demonstrating how digital technologies bridge economic disparities, motivating strategies to promote financial inclusion in underserved areas [28]. This encoding is achieved using sinusoidal functions, as defined in Eq. (5), which encode the relative positions of tokens within the sequence. The sinusoidal functions ensure that the model can distinguish the order of events, which is crucial for accurately learning time-dependent patterns and detecting anomalies. By incorporating positional encoding, the model retains the temporal context necessary for understanding the evolution of patterns in the data.

$$\mathrm{PE}_{(\mathrm{pos},2\mathrm{i})} = \sin\left(\frac{\mathrm{pos}}{10000^{2\mathrm{i}/\mathrm{d}}}\right), \mathrm{PE}_{(\mathrm{pos},2\mathrm{i}+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{2\mathrm{i}/\mathrm{d}}}\right) \tag{5}$$

In this approach, post refers to the position of a token within a sequence, and "d" denotes the dimension of the embedding. The positional encoding enhances the input sequences by embedding positional information, allowing the transformer model to learn complex dependencies and long-distance patterns within the data. For supervised evaluation or benchmarking, known anomalies are optionally labelled based on heuristic rules or expert-labelled datasets, as defined in Eq. (6). These labelled anomalies provide a ground truth for evaluating the model's ability to identify and distinguish anomalous patterns within the sequence. By incorporating labelled anomalies, the model's performance can be quantitatively assessed.

$$\mathcal{D} = \{(\mathrm{S}_\mathrm{i}, \mathrm{y}_\mathrm{i})\}, \mathrm{y}_\mathrm{i} = \begin{cases} 1 & \text{if sequence } \mathrm{S}_\mathrm{i} \text{ contains an anomaly} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This integrated technique enhances the model's ability to capture both temporal and contextual nuances in log data, which are critical for anomaly detection. By combining sequence modelling, positional encoding, and attention mechanisms, the model can detect subtle and hidden anomalies that might otherwise go unnoticed. The method effectively preserves the temporal flow of events while considering the contextual relationships between them. This enables the model to recognize complex patterns that evolve over time across multiple sessions. As a result, the model achieves high detection accuracy, even in realistic multi-session environments, where anomalies may be interspersed across different users or transactions. This approach ensures that the model is robust and reliable, making it well-suited for real-world anomaly detection tasks.

4.4 Deployment in Cloud Storage using AES-256 Encryption with KMS

The most secure approach for deploying Log BERT-based anomaly detection in cloud storage environments involves using AES-256 encryption in conjunction with a cloud-native Key Management Service (KMS), such as AWS KMS, Azure Key Vault, or Google Cloud KMS. In this setup, log data D is encrypted at rest using a symmetric 256-bit encryption key K, which is generated by the KMS and securely managed throughout its lifecycle. This encryption ensures that the log data remains confidential and protected from unauthorized access. Once encrypted, the log data can be safely stored in cloud environments, supporting non-real-time (batch) processing of the data for anomaly detection [29]. The KMS manages the encryption keys, ensuring that they are stored and used securely, and the encryption process follows the steps defined in Eq. (7). This approach provides a high level of security for sensitive log data while enabling efficient anomaly detection in the cloud.

$$C = E_K(D) \tag{7}$$

In this approach, the log data is first encrypted using the AES-256 encryption algorithm, denoted by E_K, where C represents the resulting ciphertext and K is the encryption key managed by the KMS. During batch processing, the encrypted ciphertext is decrypted back into plaintext using the corresponding decryption process, as shown in Eq. (8). This operation allows the model to access the original log data in its readable form for anomaly detection tasks. The decryption ensures that only authorized systems with the correct key can retrieve and process the data, maintaining both security and integrity throughout the analysis process [30]. This step is crucial for performing the anomaly detection while ensuring the confidentiality of the data at rest.

$$D = E_K^{-1}(C) \tag{8}$$

In this secure process, the decryption key is securely retrieved through API calls to the Key Management Service (KMS), ensuring minimal key exposure risks. The method utilizes Role-Based Access Control (RBAC) to enforce strict permissions on key access, thus reducing the risk of unauthorized use. Automated key rotation is implemented to enhance security further, while audit trails are generated to track every interaction with the encryption keys. Unifies device management foundations and synchronous data with Self-Organizing Maps (SOMs) to enhance decision-making in smart IoT setups. Guman Singh Chauhan (2021) tactic informs the proposed method by adopting synchronous data integration and advanced clustering techniques to improve anomaly detection in cloud environments, supporting enhanced decision-making through efficient and intelligent data analysis [31]. To protect the data in transit, Transport Layer Security (TLS) is applied, ensuring end-to-end protection from potential interception during transmission. Once the log data is decrypted, it is fed into Log BERT for batch processing, where sequence modelling and anomaly detection occur. This end-to-end secure process ensures that sensitive log data, such as customer activity logs, remains confidential and protected at all stages. The approach is scalable, making it well-suited for e-commerce platforms operating in multi-cloud or hybrid environments. By securely storing, encrypting, and analysing customer logs in bulk on a scheduled basis, this solution upholds regulatory compliance, ensures user data privacy, and provides a robust framework for detecting anomalies in large volumes of log data. This makes it an ideal method for sensitive log data management and anomaly detection in industries that prioritize security and compliance.

4.5 Anomaly Detection Mechanism using NLL Scoring + One-Class SVM

Anomaly detection with Log BERT combines Negative Log-Likelihood (NLL) scoring and One-Class Support Vector Machine (SVM) techniques. After training Log BERT using masked language modelling, the model learns to predict masked tokens within log sequences. For any masked token l_m in a sequence $S = \{l_1, l_2, \dots, l_n\}$, the model generates an anomaly score based on how well the predicted token matches the actual token [32]. This score is calculated using the Negative Log-Likelihood (NLL) formula, as defined in Eq. (9). A high NLL score indicates that the token is anomalous, as it deviates significantly from what the model expects based on the context of surrounding tokens. This method allows the model to detect unusual or rare log events by evaluating the likelihood of each token in the sequence, with outliers representing potential anomalies. Additionally, the One-Class SVM is used to further classify normal and anomalous patterns, enhancing the overall accuracy of the anomaly detection system.

$$\mathrm{Score}(l_m) = -\log\ P\left(l_m \mid S_{\setminus m}\right) \tag{9}$$

The final anomaly score for a sequence is obtained by averaging the individual token anomaly scores. Each token's anomaly score, derived from the Negative Log-Likelihood, reflects how unusual or unexpected it is within the context of the surrounding tokens. By averaging these token scores, the model generates a sequence-level anomaly score that provides an overall measure of how anomalous the entire log sequence is. This score, as expressed in Eq. (10), allows for the identification of sequences that deviate significantly from expected behavior, enabling more accurate detection of anomalies at the sequence level rather than just at the token level. The averaging process ensures a holistic assessment of the log sequence's integrity and helps prioritize sequences that require further investigation.

$$\mathrm{SeqScore}(S) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{Score}(l_m) \tag{10}$$

The sequence-level anomaly scores are subsequently fed into a One-Class Support Vector Machine (OC-SVM), which learns a decision function to distinguish between normal and potentially anomalous data. The OC-SVM is trained to identify the boundary that separates normal data points from outliers, using the sequence scores as input. The optimization objective of the OC-SVM, as expressed in Eq. (11), is to maximize the margin between normal data and anomalies while minimizing classification errors. By learning this decision function, the OC-SVM can effectively classify log sequences as either normal or anomalous based on their sequence-level scores. This enables the system to detect unusual patterns or behaviours in log data, enhancing the model's ability to identify rare or unexpected events that may indicate potential security threats or system malfunctions [33].

$$\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho \ \ \text{subject to } w^T\phi(x_i) \geq \rho - \xi_i, \xi_i \geq 0 \tag{11}$$

In the One-Class Support Vector Machine (OC-SVM) approach, the parameter ν controls the fraction of anomalies that the model expects to detect within the data, essentially influencing the sensitivity of anomaly detection. The function $\phi(x)$ maps the input data (sequence scores) into a higher-dimensional feature space, allowing the SVM to learn a more complex decision boundary between normal and anomalous data. During inference, sequences that fall outside the learned boundary in this higher-dimensional space are flagged as anomalous. The study compares Particle Swarm Optimization, Neural Networks, and Petri Net models for load balancing in distributed computing, highlighting their strengths in resource distribution maximization and augmenting efficiency. This work guides the proposed technique by displaying the effectiveness of these maximization models, supporting the development of improved load balancing strategies in distributed computing environments, as supported by Charles Ubagaram (2022) [34]. This method is particularly effective in scenarios where normal patterns are readily available for training, but anomalies are rare, unknown, or hard to define. By focusing on learning the characteristics of normal data, the OC-SVM is able to identify deviations from these patterns even if the anomalies are not explicitly represented in the training set. This makes the method ideal for detecting previously unseen or subtle anomalies in systems where abnormal behaviours are infrequent and not well understood, such as in security monitoring or system health checks [35].

V. RESULTS AND DISCUSSION

The results demonstrate that the Log BERT-based anomaly detection model is highly scalable, effectively handling increasing user loads without compromising performance. In particular, optimization techniques significantly reduce inference latency, allowing for faster processing even as the volume of data grows. However, it's important to note that encryption using AES-256 increases the storage size due to the additional security metadata required to protect the data at rest. Despite this increase in storage requirements, the model maintains a strong balance between performance and data security. The use of encryption ensures that sensitive data is kept secure while still enabling efficient anomaly detection [36]. This makes the model an ideal solution for cloud-enabled e-commerce applications, where both security and scalability are paramount. As e-commerce platforms expand, handling large amounts of sensitive log data while maintaining high detection accuracy becomes crucial, and this model addresses both needs. The ability to detect anomalies in real-time, while ensuring compliance with privacy regulations, positions Log BERT as a powerful tool for monitoring and safeguarding e-commerce environments.

5.1 Inference Latency Trends in Scalable Anomaly Detection Frameworks

The Inference Time vs Number of Users graph illustrates the scalability performance of the proposed Log BERT-based anomaly detection model. Initially, when the system operates under low load conditions with only a small number of users, the inference time is relatively high due to the processing overhead involved in handling smaller datasets. However, as the number of users increases, the inference time decreases significantly, demonstrating that the model is optimized for handling larger user loads. This reduction in latency indicates that the system is effectively managing the increased workload, thanks to optimization techniques such as batch processing, parallel execution, and the efficiency of transformer-based attention mechanisms. These optimizations allow the model to scale up efficiently without a corresponding increase in inference time, making it capable of handling high-volume log data in real-time or batch processing environments. The continued downward trajectory in the graph further showcases how the system becomes more efficient as the user base grows, ensuring that anomaly detection remains fast and reliable even under heavy usage. This trend highlights the model's robustness and its suitability for environments with high scalability requirements, such as cloud-based applications in e-commerce. The efficiency improvements from these optimization strategies are visually represented in Figure (2), emphasizing the model’s capacity to deliver consistent performance as the system scales.

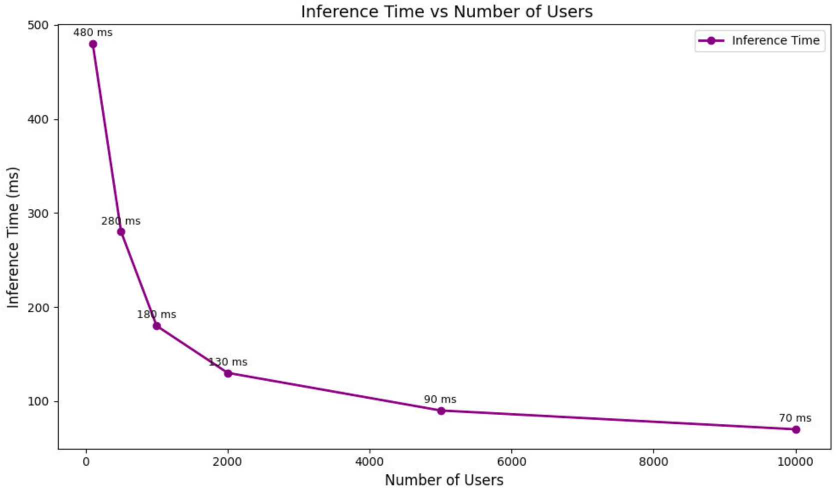

Figure 2: Scalable Anomaly Detection: Inference Latency vs. User Load

The performance of the Log BERT-based anomaly detection model reflects the use of hardware acceleration and caching strategies, which help minimize computation delays and improve overall processing efficiency. These techniques enable the model to handle larger volumes of data while maintaining rapid inference times. As a result, the model is highly scalable and can be effectively integrated into real-time and near-real-time applications, particularly for large-scale e-commerce platforms where timely anomaly detection is crucial. Even at higher user loads, the model’s inference time remains stable, ensuring that it remains responsive and capable of offering a consistent user experience, even in the presence of concurrent access. Nagendra Kumar Musham (2021) presents a design integrating intelligent personalization with secure transaction mechanisms, providing a growth-capable structure for e-commerce platforms focused on cloud-enabled security and user experience. This work stimulates the proposed procedure by combining secure, growth-capable pattern recognition with personalization techniques to enhance security and user experience in cloud-based e-commerce environments [37]. This stability is particularly important in cloud-based environments, where demand can fluctuate significantly. By leveraging hardware acceleration and efficient caching, the model adapts to changing workloads, ensuring that it remains performant and reliable under varying conditions. The ability to maintain low latency and high throughput, even with large-scale, dynamic data, is a key advantage for e-commerce platforms that rely on fast and accurate anomaly detection. The performance behavior, as illustrated in the accompanying graph, further highlights the model’s suitability for deployment in production environments, where it can effectively detect anomalies while meeting the demands of both speed and accuracy. This makes the model a strong candidate for environments that require robust, scalable solutions for real-time log analysis and security monitoring.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly, Peer Reviewed & Referred Journal|

| Volume 6, Issue 3, March 2023 |

| DOI:10.15680/IJMRSET.2023.0603034|

5.2 Analysing Encryption-Induced Storage Growth in Log Files

The graph demonstrates the increase in storage requirements for log files due to the AES-256 encryption process, which adds significant overhead. As each log file is encrypted, the encryption process introduces additional metadata, padding, and cryptographic headers to ensure the confidentiality and integrity of the data [38]. This overhead results in a noticeable increase in the size of each file, which can place a considerable burden on storage systems, especially flash storage, as the number of log files grows. The size increase is consistent across all files and is directly attributed to the encryption mechanisms required for secure cloud storage environments. This highlights the trade-off between storage efficiency and enhanced data security, where the added protection comes at the cost of increased storage usage. In large enterprise systems that frequently generate and store encrypted logs, this overhead can become a significant concern, as it impacts both the storage capacity and the associated costs [39]. As illustrated in Figure (3), the phenomenon of size increase underscores the need to balance storage efficiency with the necessity of ensuring data security. This consideration is crucial for systems handling large volumes of sensitive data, where optimizing storage without compromising security is essential.

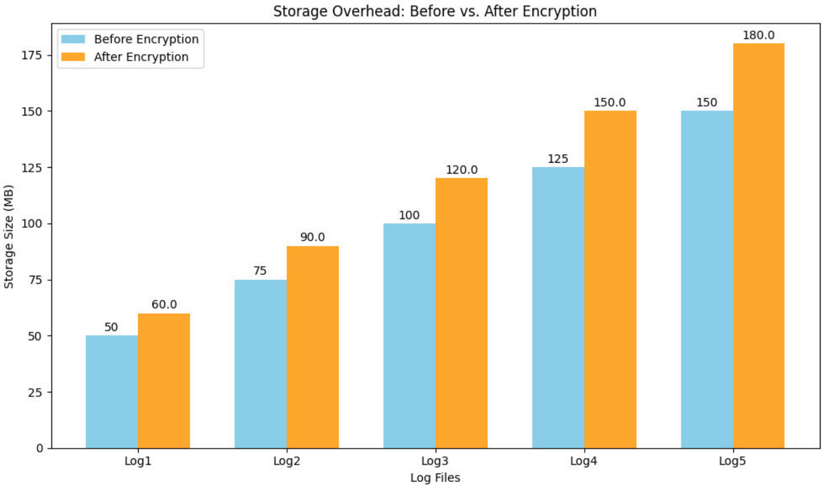

Figure 3: Effect of Data Encryption on Log File Storage Efficiency

The balance between better storage and enhanced security is crucial in systems that handle sensitive information, particularly for anomaly detection systems deployed in cloud infrastructures. In such systems, data protection and compliance with privacy regulations are paramount, making encryption a necessary but potentially costly measure [40]. The chart emphasizes the need for careful capacity planning and resource optimization when deploying encryption, as it can significantly increase the storage requirements due to the added metadata, padding, and cryptographic headers. Organizations can mitigate the impact of cloud storage costs by factoring in this storage overhead during the system design phase, allowing for better planning and more efficient resource allocation. By accounting for the increased storage needs upfront, businesses can ensure that strong security practices do not unnecessarily inflate operational costs. While encryption enhances the security of log data, the chart clearly illustrates that this comes with a quantifiable overhead, which must be considered when scaling cloud-based anomaly detection frameworks. The inquiry presents an autoencoder-GAN scheme leveraging edge-to-cloud synergy for anomalous behavior recognition in healthcare, finance, and trusted cloud storage. Motivated by this approach, the proposed solution integrates similar edge-cloud synergy to improve anomalous behavior recognition in e-commerce ecosystems, ensuring robust data integrity and instant feedback, as proven by Karthik Kushala (2021) [41]. In environments where security is critical, such as those that manage sensitive user activity data, the trade-off between security and storage efficiency is inevitable. Thus, organizations must carefully evaluate the impact of encryption on overall system performance, factoring in both the security benefits and the additional storage costs when planning for growth and scalability. The ability to optimize both security and resources will be a key factor in successfully implementing anomaly detection systems that can handle large-scale, encrypted data in cloud-based environments [42].

VI. CONCLUSION AND FUTURE WORKS

The present research work proposes Log BERT; a self-supervised anomaly detection model tailored for e-commerce platforms hosted on the swiftly changing cloud infrastructures. The model takes into account generic transformer-based architectures to understand the semantic and sequential patterns of log data without requiring any labelled training samples. To facilitate downstream anomaly classification, one class SVM is used together with masked language modelling as a pretask. As such, deviations from what is normal in terms of log behavior are effectively identified by the system. With this, further assurance of secure handling of such operational data is created since the cloud-native deployment of this architecture uses AES-256 encryption and Key Management Services. The results from the experiments indicate that anomaly detection improves its accuracy rates, reduces the likelihood of false positives being recorded, and keeps the inference latency low, thus emphasizing its promise in realistic use cases in scalable, distributed systems.

The model will also be further developed to make it possible to capture online anomalies nearly in real-time, hence more effectively boosting response against fresh threat patterns. Also, cross-system correlation across multi-cloud platforms will enhance the anomaly detection further. Besides that, we plan to explore contrastive self-supervised learning and graph-based representations to capture and better understand deeper contextual and relational structures in logs. Finally, for the future, we will be interested in broader evaluations with various real-world log datasets in order to test the generalizability and robustness of our approach in different scenarios involving e-commerce and cloud deployment.

REFERENCES

[1] Ayyadurai, R. (2022). Transaction Security in E-Commerce: Big Data Analysis in Cloud Environments. International Journal of Information Technology and Computer Engineering, 10(4), 145-156.

[2] Firdaus, A. (2022). Large-Scale Simulation of Cloud Security Breaches and Recovery Strategies in Modern E-Commerce Organizations. Perspectives on Next-Generation Cloud Computing Infrastructure and Design Frameworks, 6(10), 10-18.

[3] Kusuma, P. (2022). A Holistic Framework for Designing Secure, Scalable, and Cost-Effective Cloud-Based E-Commerce Platforms. Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures, 6(12), 7-16.

[4] Devarajan, M. V. (2022). An improved BP neural network algorithm for forecasting workload in intelligent cloud computing. Journal of Current Science, 10(3).

[5] Liando, O. E., Kapahang, M. R., & Batmetan, J. R. (2022). Cloud Security Adoption Factors in Educational Institutions. International Journal of Information Technology and Education, 1(3), 117-123.

[6] Khadka, M. (2022). A Systematic Appraisal of Multi-Factor Authentication Mechanisms for Cloud-Based E-Commerce Platforms and Their Effect on Data Protection. Journal of Emerging Cloud Technologies and Cross-Platform Integration Paradigms, 6(12), 12-21.

[7] Hussien, F. T. A., Rahma, A. M. S., & Wahab, H. B. A. (2022). Design and implement a new secure prototype structure of e-commerce system. International Journal of Electrical and Computer Engineering, 12(1), 560.

[8] Kalyan, G. (2022). A Survey on Cloud Adoption for Software Testing: Integrating Empirical Data with Fuzzy Multicriteria Decision-Making. International Journal of Information Technology & Computer Engineering, 10 (4), 32-50.

[9] Belghith, A. (2022). Investigation on e-commerce platforms for tackling e-business security challenge. International Journal on Engineering Applications, 10(6).

[10] Luo, S., & Choi, T. M. (2022). E-commerce supply chains with considerations of cyber-security: Should governments play a role. Production and Operations Management, 31(5), 2107-2126.

[11] Grandhi, S. H. (2022). Enhancing children's health monitoring: Adaptive wavelet transform in wearable sensor IoT integration. Current Science & Humanities, 10(4), 15–27.

[12] Fauziyah, F., Wang, Z., & Joy, G. (2022). Knowledge Management Strategy for Handling Cyber Attacks in E-Commerce with Computer Security Incident Response Team (CSIRT). Journal of Information Security, 13(4), 294-311.

[13] Santoso, E. (2022). Comparative Analysis of Network Segmentation Strategies to Counter Targeted Attacks in Global E-Commerce Cloud Infrastructures. Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures, 6(12), 1-6.

[14] Vargas, C. E. (2022). Evaluating Virtualization Hardening Techniques for High-Assurance Cloud-Based E-Commerce Transactions. Journal of Artificial Intelligence and Machine Learning in Cloud Computing Systems, 6(11), 9-16.

[15] Torres, A. G. (2022). Encryption Key Lifecycle Management and Best Practices for Maintaining Trusted E-Commerce Services in the Cloud. Journal of Artificial Intelligence and Machine Learning in Cloud Computing Systems, 6(11), 1-8.

[16] Bao, P. Q. (2022). Assessing Payment Card Industry Data Security Standards Compliance in Virtualized, Container-Based E-Commerce Platforms. Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems, 12(12), 1-10.

[17] Durga, P.D. (2022). Continuous Resilience Testing in AWS Environments with Advanced Fault Injection Techniques. International Journal of Information Technology & Computer Engineering, 10(3), ISSN 2347–3657.

[18] Mistri, P. (2022). Cloud Security Audit: A necessity in the cloud computing environment. The Management Accountant Journal, 57(7), 64-67.

[19] Tram, H. T. N. (2022). Automated Vulnerability Scanning Tools for Securing Cloud-Based E-Commerce Supply Chains. Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems, 12(12), 11-21.

[20] Jallouli, R., & Kaabi, S. (2022). Mapping top strategic e-commerce technologies in the digital marketing literature. Journal of Telecommunications and the Digital Economy, 10(3), 149-164.

[21] Panga, N. K. R., & M, T. (2020). Adaptive DBSCAN and federated learning-based anomaly detection for resilient intrusion detection in Internet of Things networks. International Journal of Management Research & Business Strategy, 10(4), 39.

[22] Manogaran, G., Qudrat-Ullah, H., Xin, Q., & Khan, L. (2022). Guest Editorial Introduction for the Special Section on Deep Learning Algorithms and Systems for Enhancing Security in Cloud Services. ACM Transactions on Internet Technology (TOIT), 22(2), 1-5.

[23] Prasad, V. K., Dansana, D., Mishra, B. K., & Bhavsar, M. (2022). Intensify cloud security and privacy against phishing attacks. ECS Transactions, 107(1), 1387.

[24] Lingamgunta, N. M., & Gubbala, S. G. N. A. (2022). New Key Agreement Protocol and Cryptosystem over ECC under SET Protocol Environment in E-Commerce. International Journal of Intelligent Engineering and Systems, 15(4), 319-328.

[25] Sreekar, P. (2021). Analyzing Threat Models in Vehicular Cloud Computing: Security and Privacy Challenges. International Journal of Modern Electronics and Communication Engineering, 9(4), ISSN2321-2152.

[26] Inayatulloh, I., Hartono, I. K., Fachrul, A. F., Al Farisi, M. S., Sriwardiningsih, E., & Fachri, R. M. F. (2022). Blockchain technology for customer protection in e-commerce transactions. Journal of Cybersecurity, 1(1), 1-12.

[27] Akour, I., Alnazzawi, N., Alshurideh, M., Almaiah, M. A., Al Kurdi, B., Alfaisal, R. M., & Salloum, S. (2022). A conceptual model for investigating the effect of privacy concerns on E-commerce adoption: a study on United Arab Emirates consumers. Electronics, 11(22), 3648.

[28] Boyapati, S., & Kaur, H. (2022). Mapping the Urban-Rural Income Gap: A Panel Data Analysis of Cloud Computing and Internet Inclusive Finance in the E-Commerce Era. International Journal of Information Technology and Computer Engineering, 7(4).

[29] Kim, S. I., & Kim, S. H. (2022). E-commerce payment model using blockchain. Journal of Ambient Intelligence and Humanized Computing, 13(3), 1673-1685.

[30] Singh, S. P., Alotaibi, Y., Kumar, G., & Rawat, S. S. (2022). Intelligent adaptive optimisation method for enhancement of information security in IoT-enabled environments. Sustainability, 14(20), 13635.

[31] Chauhan, G. S., Jadon, R., & Awotunde, J. B. (2021). Smart IoT analytics: Leveraging device management platforms and real-time data integration with self-organizing maps for enhanced decision-making.International Journal of Applied Science, Engineering, and Management, 15(2).

[32] Xu, L. (2022). Small and Medium Enterprise ERP Platform Based on Windows Azure Cloud Computing. Academic Journal of Business & Management, 4(2), 17-22.

[33] Esfahbodi, A., Pang, G., & Peng, L. (2022). Determinants of consumers' adoption intention for blockchain technology in E-commerce. Journal of Digital Economy, 1(2), 89-101.

[34] Ubagaram, C., Mandala, R. R., Garikipati, V., Dyavani, N. R., Jayaprakasam, B. S., & Purandhar, N. (2022). Workload balancing in cloud computing: An empirical study on particle swarm optimization, neural networks, and Petri net models. Journal of Science and Technology, 7(7), 36-57.

[35] Lei, T., & Yongqing, Z. (2022). The Impact of Venture Investment of E-Commerce on Enterprises. Academic Journal of Business & Management, 4(12), 59-62.

[36] Marjerison, R. K., Zhang, Y., & Zheng, H. (2022). AI in E-Commerce: Application of the Use and Gratification Model to the Acceptance of Chatbots. Sustainability, 14(21), 14270.

[37] Musham, N. K., & Hemnath, R. (2021). Real-time path planning for IoT-enabled autonomous vehicle robotics using RRT and A* algorithms. International Journal of Multidisciplinary Research and Explorer, 1(7), 65.

[38] Karagozlu, D., & Ganyaupfu, S. (2022). Customers' View of E-commerce During the Covid-19 Pandemic. Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8(3), 1047-1060.

[39] Jamali, M. R., Kansro, N. A., Chandio, S., Rajper, G. N., & Shah, S. A. A. (2022). The Design, Use and Impact of Cloud Computing During the Covid-19 Crises. VFAST Transactions on Software Engineering, 10(4), 181-189.

[40] Nazir, J., Iqbal, M. W., Alyas, T., Hamid, M., Saleem, M., Malik, S., & Tabassum, N. (2022). Load balancing framework for cross-region tasks in cloud computing. Computers, Materials & Continua, 70(1), 1479-1490.

[41] Kushala, K., & Pushpakumar, R. (2021). Edge-to-cloud synergy: An autoencoder-GAN framework for anomaly detection in healthcare records, financial statements, and secure cloud storage. International Journal of Advance Research and Innovative Ideas in Education, 7(1),

[42] Chinamanagonda, S. (2022). Zero Trust Security Models in Cloud Infrastructure-Adoption of zero-trust principles for enhanced security. Academia Nexus Journal, 1(2).

INTERNATIONAL JOURNAL OF
MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com