A Comprehensive Review of Big Data Mining Security: Cryptographic-Hybrid Intrusion Detection and Decentralized Rights Management

Pasupuleti Nagendra Babu[1], S. Ramakrishna[2]
[1]*Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool, India.*
[2]*Professor, Department of Computer Science, S. V. University, Tirupati, India.*

***Abstract*—The exponential proliferation of digital data generated by web applications, social networking platforms, and the Internet of Things (IoT) has ushered in the era of Big Data. While Big Data Analytics (BDA) and data mining empower organizations to extract actionable insights for competitive advantage, they simultaneously introduce severe security vulnerabilities and privacy leakage risks. Traditional perimeter-based security mechanisms, such as firewalls and standard encryption, fail to scale across distributed, high-velocity, and heterogeneous datasets. This review paper analyzes a comprehensive research thesis addressing these challenges. First, it examines the foundational architecture of Big Data (characterized by the 5Vs) and evaluates existing Privacy-Preserving Data Mining (PPDM) techniques, including anonymization, perturbation, and cryptographic models.**

Second, it reviews two novel architectural proposals presented in the thesis:

(1) the Secured Crypto-Hybrid Algorithm (SCHA), which integrates K-Means clustering and Local Outlier Factor (LOF) anomaly detection with AES-256 encryption to secure network traffic; and

(2) a decentralized security paradigm for Big Data platforms, utilizing OpenSDRM (Open Digital Rights Management) for user-centric social media control and Software-Defined Networking (SDN) for active intrusion prevention. Empirical evaluations using NSL-KDD and Kyoto 2006+ datasets demonstrate that the proposed hybrid frameworks significantly reduce false positive rates while enhancing detection accuracy and data confidentiality.

***Index Terms*—Big Data Analytics, Privacy-Preserving Data Mining (PPDM), Secured Crypto-Hybrid Algorithm (SCHA), Intrusion Detection System (IDS), AES-256, OpenSDRM, Software-Defined Networking (SDN).**

I. INTRODUCTION

The 21st century is defined by an unprecedented information explosion driven by rapid advancements in information technology, globalization, mobile computing, and social media. Organizations across healthcare, finance, retail, and telecommunications routinely process datasets reaching exabytes and zettabytes. This massive influx of information termed Big Data cannot be captured, stored, or analyzed using conventional database management tools.

The thesis under review establishes that Big Data is governed by five core dimensions, known as the 5Vs:

1. Volume: The sheer scale of generated data, requiring distributed storage and parallel processing frameworks like Hadoop and MapReduce.
2. Velocity: The accelerating speed at which data streams are generated and processed, shifting paradigms from batch processing to real-time analytics.
3. Variety: The structural heterogeneity of data, encompassing structured (relational databases, CSV), semi-structured (XML, JSON), and unstructured formats (video, audio, social media text, sensor logs).
4. Value: The extraction of meaningful, actionable business intelligence from raw datasets.
5. Veracity: The trustworthiness, accuracy, and consistency of data, mitigating the impact of noise and anomalies.

To extract value from these complex ecosystems, organizations rely on Big Data Analytics (BDA) and Data Mining the computational process of discovering

non-trivial, previously unknown patterns using machine learning, statistical analysis, and pattern recognition. However, aggregating vast amounts of personal, financial, and behavioral data creates a critical vulnerability. Cyber-attacks ranging from Denial of Service (DoS) and probing to malware and ransomware threaten the confidentiality, integrity, and availability of network resources. Furthermore, collaborative data mining across untrusted party's risks exposing sensitive individual identities.

The primary objective of the analyzed thesis is to bridge the gap between high-efficiency data mining and robust information security. It achieves this by categorizing existing threats, evaluating Privacy-Preserving Data Mining (PPDM) methodologies, and introducing integrated cryptographic and network-level security architectures.

II. RELATED WORK

A substantial portion of the thesis is dedicated to surveying the evolution of data mining algorithms, Big Data security challenges, and Privacy-Preserving Data Mining (PPDM) frameworks.

2.1 Classification of PPDM Techniques: Following the taxonomy established by Tran and Hu [10] and Agarwal and Yu, the thesis categorizes input PPDM methodologies into four primary domains:

Non-Perturbation (Anonymization) Methods: These techniques sanitize identifiable attributes before data release. Sweeney [183] introduced k-anonymity, which uses generalization and suppression so that any given record is indistinguishable from at least k-1 other records. To overcome homogeneity and background knowledge attacks inherent to k-anonymity, subsequent researchers developed l-diversity (ensuring sensitive attributes have at least l well-represented values) and t-closeness (Li and Venkatasubramanian [156]), which maintains the distribution of sensitive attributes close to the overall table distribution. Cao and Karras [125] further extended this to β-likeness.

Perturbation and Randomization Methods: Evfimievski [184] and Dwork [163] pioneered methods that introduce controlled statistical noise to mask individual records while preserving aggregate distributions. Perturbation is executed via additive noise or multiplicative transformations (such as Random Projection proposed by Liu et al.). Differential privacy (Dwork) utilizes mechanisms like Laplace noise to maximize query accuracy while minimizing record identification risks, though Hu et al. [71] noted this can induce a 15% to 30% loss in data utility.

Cryptographic and Secure Multi-Party Computation (SMC) Methods: Introduced by Lindell and Pinkas, cryptographic approaches allow mutually untrusted parties to jointly compute functions over private inputs without revealing raw data. Advanced implementations include Fully Homomorphic Encryption (FHE) and Attribute-Based Encryption (ABE) such as Key-Policy ABE (Goyal et al.) and Ciphertext-Policy ABE (Bethencourt et al. [159]) which allow computations and fine-grained access control directly over encrypted cloud data.

Hybrid Methods: Researchers such as Kaur [27] and Lohiya and Ragha [122] demonstrated that combining suppression, randomization, and generalization minimizes information loss while maintaining high privacy guarantees.

2.2 Big Data Security and Intrusion Detection Systems (IDS): The literature survey highlights that traditional security norms designed for relational DBMS fail when applied to unstructured, distributed environments like Hadoop and cloud computing.

Cloud and Hadoop Vulnerabilities: Inukollu et al. [96] and Zhang et al. [124] identified severe security gaps in Hadoop Distributed File Systems (HDFS) and MapReduce, particularly regarding inter-node communication, data-at-rest protection, and compliance auditing. To secure cloud databases, Arasu et al. [111] developed Cipherbase using tamper-proof FPGA hardware, while Popa et al. (referenced via MIT's CryptDB) enabled SQL queries over encrypted datasets via a proxy architecture.

Machine Learning in Intrusion Detection: Network Intrusion Detection Systems (IDS) rely heavily on data mining to detect misuse (signature-based) and anomalies. Tavallaee et al. [142] analyzed the KDD'99 dataset, revealing high record redundancy (78% in training, 75% in testing) that biases classifiers against infrequent attacks like User-to-Root (U2R). Furthermore, Choi [147] emphasized that no single algorithm achieves a low false alarm rate across all attack vectors, necessitating hybrid, multi-algorithmic approaches.

III. METHODOLOGY

To understand and mitigate security threats within data mining workflows, the thesis establishes a structured methodological foundation covering data mining paradigms, threat taxonomies, and evaluation metrics.

3.1 Core Data Mining Techniques: The thesis categorizes data mining methodologies into supervised and unsupervised learning frameworks:
Classification (Supervised): Models map input variables to predefined categorical labels. Key algorithms evaluated include Decision Trees (C4.5, Recursive Partitioning, Random Forest), Naïve Bayes (probabilistic classifiers based on Bayes' theorem assuming feature independence), Support Vector Machines (SVM) (constructing optimal hyperplanes with maximum margins), K-Nearest Neighbors (KNN) (majority-vote classification in multidimensional space), and Artificial Neural Networks (ANN).
Clustering (Unsupervised): Partitioning unlabeled datasets into groups of similar objects. The thesis contrasts Hierarchical Clustering (agglomerative bottom-up and divisive top-down dendrograms) with Partitioning Clustering, specifically focusing on K-Means, which iteratively assigns data points to the nearest cluster centroid to minimize squared error.
Association Rule Mining & Regression: Algorithms like Apriori, Eclat, and FP-Growth discover co-occurrence relationships (measured via support, confidence, lift, and conviction), while linear and logistic regression predict continuous or binary outcomes.

3.2 Taxonomy of Data Mining and Network Attacks
The research defines six primary cyber-attack vectors targeting data mining infrastructures:

1. Denial of Service (DoS) & Distributed DoS (DDoS): Flooding network resources with excessive traffic from single or multiple domains to exhaust capacity and crash servers.
2. Probing & Port Scanning: Scanning network topologies to identify active hosts, open ports, and system vulnerabilities.
3. Spoofing: Falsifying IP addresses to masquerade as a trusted internal client.
4. User to Root (U2R): Exploiting vulnerabilities to escalate normal user privileges to root/administrator access.
5. Malware & Spyware: Malicious executable scripts designed to damage systems or covertly exfiltrate sensitive user credentials.
6. Ransomware: Malicious software that encrypts target systems and demands ransom for decryption keys, identified in the thesis as one of the most severe modern threats.

3.3 Experimental Datasets and Evaluation Metrics: To benchmark security models, the methodology utilizes two standard network intrusion datasets and the WEKA data mining suite:

- NSL-KDD: A refined version of the KDD CUP 99 dataset that eliminates redundant records, comprising 43 distinct features.
- Kyoto 2006+: Real-world network traffic data collected from honeypots and servers at Kyoto University, containing 24 statistical and security features.

Model performance is rigorously evaluated using a confusion matrix True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to calculate:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

IV. PROPOSED WORK SECURED CRYPTO-HYBRID ALGORITHM (SCHA)

In this, the dual challenge of detecting sophisticated network intrusions while protecting the underlying datasets from unauthorized access. The author proposes the Secured Crypto-Hybrid Algorithm (SCHA), an integrated framework combining unsupervised anomaly detection with symmetric-key cryptography.

4.1 Architectural Phases of SCHA: The SCHA framework operates across four sequential phases:

Phase	Core Processes	Tools & Techniques
1. Dataset Mining	Ingestion of benchmark network traffic logs.	NSL-KDD (43 features) & Kyoto 2006+ (24 features).

2. Preprocessing	Data cleaning, missing value imputation, normalization, and feature reduction.	Byte-sequence extraction, hexadecimal dumps, and n-gram frequency generation via WEKA.
3. Training & Testing	Hybrid unsupervised anomaly and outlier detection.	K-Means Clustering + Local Outlier Factor (LOF).
4. Encryption & Decryption	Securing processed datasets against unauthorized tampering or leakage.	AES-256 (Advanced Encryption Standard) block cipher.

4.2 Algorithmic Mechanics:

1. Feature Extraction via N-Grams: Raw network features (such as Source/Destination IP addresses) are converted into byte arrays and hexadecimal dumps. Using WEKA, n-grams are generated, retaining high-frequency features while discarding low-frequency noise to optimize computational efficiency.
2. Hybrid Anomaly Detection (K-Means + LOF):

- First, K-Means clustering partitions the network traffic into k clusters by iteratively updating cluster centroids (mc) until convergence.
- Second, the Local Outlier Factor (LOF) algorithm evaluates the local density deviation of each data point relative to its k-nearest neighbors. By calculating the local approachable density (\(\text{lad}_{k}\)) and comparing it against neighboring densities, LOF assigns an outlier score. Data points exhibiting significantly lower density than their neighbors (LOF > 1) are flagged as anomalous intrusions (e.g., DoS, Malware, Spyware).

3. Cryptographic Protection (AES-256): Once classification and outlier detection are complete, the resulting datasets are encrypted block-by-block using a 256-bit AES cipher with an initialization vector (Iv) and XOR chaining operations. This ensures that stored analytical outputs remain confidential and immune to brute-force decryption.

4.3 Empirical Results and Performance Analysis

Simulation results presented in the thesis demonstrate that SCHA substantially outperforms standalone classifiers (such as Random Forest and J48) and conventional baseline models:

- False Positive Reduction: Compared to baseline detection methods which suffered from high false positive rates in Spoofing (30%), Malware (25%), and Probing (25%) the finalized SCHA model reduced false positives to 1%–4% across all attack categories.
- High Detection Accuracy: SCHA achieved overall detection accuracies of 99% for Normal traffic, 96% for DoS, 94% for DDoS, 93% for Malware, and 90% for both Probe and Ransomware attacks.

V. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The reviewed thesis provides a rigorous, multi-layered investigation into the intersection of Big Data Analytics, Privacy-Preserving Data Mining, and network cybersecurity. While data mining unlocks transformative value across healthcare, finance, and commerce, the velocity, volume, and variety of modern datasets render traditional, isolated security tools obsolete.

The research demonstrates that securing Big Data ecosystems requires an integrated, hybrid approach:

1. At the analytical and storage layer, combining unsupervised machine learning (K-Means and LOF) with robust symmetric encryption (AES-256) in the SCHA framework effectively identifies complex network intrusions with up to 99% accuracy while safeguarding sensitive datasets.
2. At the application and network layers, decoupling user data from centralized platforms via OpenSDRM restores privacy control to end-users, while programmable SDN-based IDS/IPS architectures enable real-time, automated threat containment.

5.2 Future Scope: Building upon these findings, the thesis outlines several critical avenues for future research:

- Post-Quantum Cryptography: Integrating quantum-resistant encryption algorithms into Big Data frameworks to protect long-term sensitive data against future quantum computing threats.

- Federated Learning & Edge Computing: Shifting privacy-preserving analytics toward decentralized edge devices and IoT nodes using lightweight cryptographic protocols and federated learning.
- Blockchain for Data Provenance: Incorporating distributed ledger technologies to establish immutable audit trails and enhance data integrity across multi-party cloud environments.
- Real-Time Adaptive AI: Advancing deep learning models capable of autonomous, real-time anomaly detection over non-stationary streaming data while adhering to evolving global regulations (such as GDPR and India's DPDP Act).

REFERENCES

[1]. Yasa, R. N., I. K. S. Buana, Girinoto, H. Setiawan, and R. B. Hadiprakoso, "Modified RNP privacy protection data mining method as big data security," in *Proc. International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, 2021.

[2]. Ekwunife, N., "National security intelligence through social network data mining," in *Proc. IEEE International Conference on Big Data (Big Data)*, 2020.

[3]. Hassani, H., C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing*, vol. 4, no. 1, Art. no. 1, 2020.

[4]. Faker, O., and E. Dogdu, "Intrusion detection using big data and deep learning techniques," in *Proc. ACM Southeast Conference (ACMSE)*, 2019, pp. 18–20.

[5]. Ur Rehman, M. H., I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial Internet of Things," *Future Generation Computer Systems*, vol. 99, pp. 247–259, 2019.

[6]. Tran, H., and J. Hu, "Privacy-preserving big data analytics: A comprehensive survey," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 207–218, 2019.

[7]. Chamikara, M. A., P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "An efficient and scalable privacy-preserving algorithm for big data and data streams," *Computers & Security*, vol. 87, Art. no. 101570, 2019.

[8]. Oussous, A., F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University – Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.

[9]. Abouelmehdi, K., A. Beni-Hessane, and H. Khaloufi, "Big healthcare data: Preserving security and privacy," *Journal of Big Data*, vol. 5, no. 1, Art. no. 1, 2018.

[10]. Günther, W. A., M. H. R. Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, 2017.

[11]. Sivarajah, U., M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263–286, 2017.

[12]. Mehmood, A., I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.

[13]. Hu, X., M. Yuan, J. Yao, Y. Deng, L. Chen, Q. Yang, et al., "Differential privacy in Telco big data platform," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1692–1703, 2015.

[14]. Wu, X., X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[15]. Inukollu, V. N., S. Arsi, and S. R. Ravuri, "Security issues associated with big data in cloud computing," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 6, no. 3, pp. 45–56, 2014.

[16]. Arasu, A., S. Blanas, K. Eguro, R. Kaushik, D. Kossmann, R. Ramamurthy, and R. Venkatesan, "Secure database-as-a-service with Cipherbase," in *Proc. ACM SIGMOD International Conference on Management of Data*, 2013, pp. 1033–1036.

[17]. Roy, I., S. T. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and privacy for MapReduce," in *Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2010, pp. 297–312.

[18]. Tavallaee, M., E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symposium on*

Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6.

[19]. Li, N., T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd International Conference on Data Engineering (ICDE)*, 2007, pp. 106–115.

[20]. Sweeney, L., "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.