

MACHINE LEARNING-BASED INTRUSION DETECTION FOR BIG DATA SECURITY IN CLOUD HEALTHCARE SYSTEMS

Bhagyashree P
Computer Science and Engineering
Reva University
Bangalore, India
bhagyashree.pandian@gmail.com

Subhatra Ravikumar
Computer Science and Engineering
Reva University
Bangalore, India
subhadil09@gmail.com

Swethashree RN
Computer Science and Engineering
Reva University
Bangalore, India
swethashreern@gmail.com

Ayesha Sana
Computer Science and Engineering
Reva University
Bangalore, India
ayashaafiyah7@gmail.com

Dr. Raje Siddiraju Upendra
School of Electronics and Communication Engineering
Reva University
Bangalore, India
upendra.rs@reva.edu.in

Abstract: This paper presents a big data-driven security framework for protecting cloud-based healthcare data using machine learning-based intrusion detection. With the rapid adoption of cloud computing in healthcare, large volumes of sensitive patient data are stored and processed in distributed environments, increasing the risk of cyber threats and unauthorized access. To address this challenge, the proposed system integrates big data analytics with machine learning techniques to detect and classify network intrusions in real time. The NSL-KDD benchmark dataset is used for model training and evaluation. Three machine learning algorithms—Random Forest, Support Vector Machine, and Logistic Regression—are implemented and compared using accuracy, precision, and recall metrics. Experimental results show that the Random Forest model achieved superior performance with detection accuracy exceeding 99 percent across multiple attack categories, including DoS, Probe, R2L, and U2R attacks. The proposed framework enhances the security, scalability, and reliability of cloud-based healthcare systems. This study demonstrates the effectiveness of combining big data analytics and machine learning for intelligent and proactive cybersecurity in healthcare cloud environments.

Keywords: Big Data Analytics, Cloud Security, Healthcare Data, Intrusion Detection System, Machine Learning, Random Forest

I. INTRODUCTION

Healthcare systems increasingly rely on cloud computing for storing and managing large volumes of medical data. Cloud platforms provide scalability, flexibility, and cost efficiency; however, they also introduce serious security and privacy challenges. Unauthorized access, data breaches, and cyberattacks pose significant risks to sensitive patient information. Traditional security mechanisms are often insufficient for detecting evolving cyber threats in large-scale cloud environments. The integration of big data analytics and machine learning enables intelligent monitoring and real-time threat detection. Machine learning-based Intrusion Detection Systems (IDS) can analyze large network traffic data and identify abnormal patterns. Identify applicable funding agency here. If none, delete this effectively. This paper proposes a big data-driven machine learning framework for securing healthcare data stored in cloud environments. Main Contributions:

- Development of a big data-enabled cloud security framework for healthcare

- Comparative analysis of RF, SVM, and LR for intrusion detection
- Evaluation using NSL-KDD dataset across multiple attack types
- Identification of the most effective model for healthcare cloud security

II. LITERATURE REVIEW

As we are in an era of technologies like cloud environment and data analytics, where we are profiting a lot, at the same time we are facing some issues in data protection. We have few old methods for attack identification but they have their own drawback. To resolve these drawbacks we have started taking help of Machine Learning. Machine Learning started identifying attacks by using already know information/data. In that Random Forest has become more powerful for large data, gradient Boosting for positive rate, SVM for feature scope, Logistics regression for calculations made attack identification easy. In healthcare domain, data protection is highly essential, then Machine Learning was a hope. Deep learning approaches, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) needs high processing, and by using Principal Component Analysis (PCA) and correlation-based methods reduced high processing. But all these evolution were not domain specific such as health care. To answer this, we are proposing approach where we improve handling large data , protection of data and stability of health care data in cloud environment.

III. EXISTING SYSTEM

Current healthcare data handling system mainly depends on cloud computing which have many rewards like handling large data, reasonable cost, retrieving distantly, loading electronic health record (EHR), concurrent patient observation. However, this system uses old methods for attack identification which has many drawbacks, also preserving and keeping database up-to date needs in-person work. Due to extensive network large data and data processing results in slow attack identification and also give negative results. Also current system has less coordination between data processing and data protection , so the

act of ML may be less effective while handling large data, protection of data, and stability of healthcare data in cloud environment

IV. PROPOSED SYSTEM

In order to overcome the shortcomings of current cloud- based healthcare security solutions, this study suggests a big data-driven machine learning framework for safe health- care data management in cloud environments and intelligent intrusion detection. By combining big data analytics with cutting-edge machine learning algorithms, the suggested sys- tem seeks to deliver scalable, accurate, and real-time cyber threat detection. There are four main layers in the suggested framework: 1. Layer of Data Acquisition 2. Layer of Big Data Processing 3. Detection Layer for Machine Learning 4. Layer of security Response and Monitoring The system is made to examine massive network traffic and spot malicious activity that compromises cloud infrastructure for healthcare. The NSL-KDD benchmark dataset serves as a sample dataset for network traffic analysis in this project. Four main assault categories and typical behavior are represented by the dataset's forty-one input attributes and one class label. DoS (Denial of Service) Probe attacks R2L (Remote-to-Local) U2R, or user- to-root. The dataset is frequently used to assess intrusion de- tection methods and replicates actual network traffic patterns. Preprocessing is essential to enhancing model performance since cloud settings produce massive amounts of diverse data. The big data preprocessing procedures listed below are carried out: Missing or inconsistent values are eliminated, data is normalized and standardized, categorical properties are converted using one-hot encoding, features are transformed for better model learning, and datasets are balanced to lessen bias. These actions match the system with big data analyt- ics principles by enhancing data quality and facilitating the effective management of large-scale datasets. The machine learning detection module, which uses several algorithms to categorize network activity as benign or malevolent, is the main part of the suggested system. He models listed below are used and contrasted. Random Forest is an approach to ensemble learning properly manages high dimensional data resilient to over fitting and noise. SVM, or support vector machine is the second approach which is high dimensional categorization that works and captures nonlinear patterns using kernel functions. LR, or Logistic Regression computationally effective offers excellent interpretability ideal for locations with a lot of data to find the best algorithm for cloud security in healthcare, a comparison analysis is conducted. Standard classification measures are used to assess the model's performance: Precision Accuracy and Sensitivity. These metrics offer a thorough assessment of the system's capacity to minimize false alarms and accurately identify intrusions. The solution sends security administrators real-time notifications as soon as an intrusion is discovered. The suggested framework backs: Early identification of threats, ongoing observation of cloud network operations, Determining the nature and intensity of the attack, Assistance with proactive security administration. This makes it possible for healthcare institutions to react promptly and stop possible data breaches. Large-scale real- time healthcare cloud settings are supported by the suggested framework, even if the current implementation leverages NSL- KDD for experimental assessment. Big data platforms like Spark for distributed data processing or Apache Hadoop can be used to expand the system. Systems for cloud storage that enable scalable data management, analysis of real-time streaming for ongoing observation. Because of this, the suggested approach is scalable and appropriate for big data applications in healthcare in the future. This project's principal contribu- tions consist of: Creating an intrusion detection framework for healthcare cloud systems that is supported by big data, integrating several machine learning models for comparison analysis, and detecting various attack

types (DoS, Probe, R2L, and U2R), Finding the best model for extensive security monitoring and a scalable architecture appropriate for actual healthcare cloud settings

The performance comparison table presents the evaluation of three machine learning algorithms—Random Forest, Support Vector Machine (SVM), and Logistic Regression—used for intrusion detection in the proposed system. The models are assessed using three key metrics: accuracy, precision, and recall.

Accuracy represents the overall correctness of the model in classifying both normal and malicious network traffic. Precision indicates the proportion of correctly identified intrusion instances among all detected intrusions, while recall measures the model's ability to correctly identify actual attack instances. From the results, it is observed that the Random Forest algorithm achieves the highest performance across all evaluation metrics, with an accuracy of 99.2

The SVM model also performs well, achieving an accuracy of 97.8

Overall, the results indicate that Random Forest is the most suitable model for intrusion detection in cloud-based healthcare environments due to its high accuracy and reliability.

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)
Random Forest	99.2	98.9	99.1
SVM	97.8	97.5	97.6
Logistic Regression	95.6	95.2	95.4

The proposed system follows a layered architecture for detecting intrusions in cloud-based healthcare environments using big data analytics and machine learning techniques.

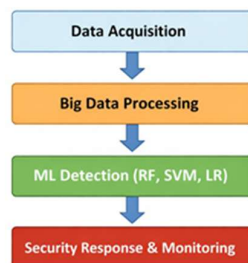


Fig. 1. Proposed Big Data-Driven Intrusion Detection Framework for Health-care Cloud

The framework consists of four major layers: Data Acquisition, Big Data Processing, Machine Learning Detection, and Security Response and Monitoring.

Data Acquisition Layer

This layer is responsible for collecting network traffic and healthcare-related data from cloud environments. In this work, the NSL-KDD dataset is used as the primary data source. The collected data includes both normal and

malicious traffic patterns such as DoS, Probe, R2L, and U2R attacks. Big Data Processing Layer

In this layer, the collected data is preprocessed and transformed to make it suitable for machine learning models. The preprocessing steps include data cleaning, normalization, handling missing values, and encoding categorical features. This layer ensures efficient handling of large-scale data and improves the overall quality of input data.

Machine Learning Detection Layer

This is the core layer of the system where intrusion detection is performed. Multiple machine learning algorithms—Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR)—are applied to classify network traffic as normal or malicious. A comparative analysis is carried out to determine the most effective model, with Random Forest showing superior performance.

Security Response and Monitoring Layer

This layer is responsible for generating alerts and monitoring system activity in real time. When an intrusion is detected, the system triggers alerts for administrators and identifies the type of attack. This enables quick response and helps prevent data breaches in healthcare cloud systems. “Temperature/K”.

ACKNOWLEDGMENT

The authors would like to thank Reva University for providing the facilities and encouraging research atmosphere needed to complete this work. We would like to express our sincere gratitude to Dr. Raje Siddiraju Upendra, Associate Professor and Head-Research, for his invaluable advice, unwavering support, and perceptive recommendations during this study.

We also thank everyone who helped us finish this study successfully, whether directly or indirectly.

REFERENCES

- [1]. A, Jayasudha, and D, Pugazhenth, “Classification of Textures using Patch Based Energy Features of Selected Wavelet Coefficients”, *Journal of Applied Sciences*, Vol. 14, No. 7, 2014, pp. 709-713.
- [2]. Z. Yang, X. Liu, T. Li, D. Wu, and J. Wang, “A systematic literature review of anomaly-based network intrusion detection,” in *Proc. CS*, 2022, pp. 1–6.
- [3]. V. Kathiresan, S. Karthik, P. Divya, and D. P. Rajan, “A comparative study of machine learning techniques for intrusion detection,” in *Proc. Int. Conf. Computer Communication and Informatics (ICCCI)*, 2022, pp. 1–6.
- [5]. A. Kiran et al., “Intrusion detection system using machine learning,” in *Proc. ICCI*, 2023, pp. 1–4.
- [6]. J. A. Abraham and V. R. Bindu, “Intrusion detection and prevention using machine learning and deep learning approaches: A review,” in *Proc. ICAECA*, 2021.
- [7]. U. Dixit, S. Bhatia, and P. Bhatia, “Comparison of machine learning algorithms for intrusion detection systems,” in *Proc. COM-IT-CON*, 2022.

- [8]. G. Yedukondalu et al., "Machine learning framework for network intrusion detection," in Proc. ICIRCA, 2021.
- [9]. V. W. Samawi, S. A. Yousif, and N. M. G. Al-Saidi, "Automatic machine learning for intrusion detection using Auto-WEKA," in Proc. ICSGRC, 2022.
- [10]. Waskle, S. Parashar, and U. Singh, "Intrusion detection using PCA with Random Forest approach," in Proc. ICESC, 2020.
- [11]. Bhati and M. Khari, "An ensemble model for network intrusion detection using AdaBoost and Random Forest," Springer, 2022.
- [12]. B. S. Bhati and C. S. Rai, "Analysis of support vector machine- based intrusion detection techniques," Arabian Journal for Science and Engineering, vol. 45, pp. 2371–2383, 2020.
- [13]. S. Krishnaveni et al., "Anomaly-based intrusion detection using support vector machine," Springer, 2020.
- [14]. Selvaraj and S. Sundaravaradhan, "Challenges and opportunities in IoT healthcare systems: A systematic review," SN Applied Sciences, vol. 2, no. 1, 2020.
- [15]. V. S. Naresh et al., "Internet of Things in healthcare: Architecture, applications, challenges and solutions," CSE, vol. 35, no. 6, 2020.
- [16]. T. Nithya et al., "A comprehensive survey of machine learning: Advancements, applications and challenges," in Proc. ICAISS, 2023.
- [17]. K. Gai, M. Qiu, and H. Zhao, "Privacy-preserving big data analytics for cloud-enabled healthcare systems," IEEE Access, vol. 8, pp. 15935– 15945, 2020.
- [18]. M. Chen, Y. Ma, J. Song, C. F. Lai, and B. Hu, "Smart healthcare: Big data analytics and cloud computing," IEEE Network, vol. 31, no. 1, pp. 14–20, 2017.
- [19]. A. Javaid et al., "A deep learning approach for network intrusion detection system," Procedia Computer Science, vol. 155, pp. 394–401, 2019.
- [20]. Maheswari et al., "Machine learning performance evaluation using feature selection," Springer, 2023.