The Scientific Temper (2024) Vol. 15 (2): 2181-2189
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.30
https://scientifictemper.com/

RESEARCH ARTICLE

Improvement of data analysis and protection using novel privacy-preserving methods for big data application

Mohamed Azharudheen A.*, Vijayalakshmi V.

Abstract

Due to the increasing volume of data, the importance of data analysis systems has become more critical. An intrusion detection system is a type of software that monitors and analyzes the data collected by a network or system. Due to the increasing volume of data collected in the medical field, it has become harder for traditional methods to detect unauthorized access and manipulation of the data. To advance the efficiency of big data analysis, various techniques are used in IDS. This paper proposes a method that combines the deep learning network and proposed optimization algorithm. The goal of this paper is to develop a classification model that takes into account the hidden layer nodes of the DBN and then implement a PSO algorithm to improve its structure. The results of the simulations show that the Spark-DBN-PSO algorithm achieves a 99.04% accuracy rate, which is higher than the accuracy of other deep neural network (DNN) and artificial neural network (ANN) algorithms. The results of the research demonstrate that the proposed methodology performs superior than the existing algorithm.

Keywords: Apache spark, Big data, ChiSqSelector, Intrusion detection, Support vector machine.

Introduction

The rise of the internet has resulted in the increased use of large data sets in the medical field. This has resulted in the creation of electronic health records (Hou *et al*., 2020). According to experts, the amount of data that will be collected and stored in the medical industry by 2023 will be over 35 ZB, which is 44 times higher than the 2009 statistics (Ly *et al*., 2020). The way public health care is delivered has gradually transformed as a result of big data. It has resulted in the creation of novel pharmaceuticals and treatments that are customized to the individual (Swan, 2012). Big data is a huge collection of information that can be used to analyze and improve a wide range of situations. It is also referred to as a pool of complex data sets. Due to the immense amount of data that can be collected and stored in a big data set, traditional database systems can't handle it. Instead, it's generated by various streams (Viji *et al*., 2017). Due to rapid technological development, the amount of data that can be collected and stored in a big data set has increased. This has raised the importance of securing the data and keeping it private (Jain *et al*., 2016). People must always consider their data's security when it comes to protecting it. This is because the information that they share through social media can be used to identify them (Tewari *et al*., 2020).

Data security is different from data privacy. Privacy is about protecting the individual's data only. This is done in the right way (Xu *et al*., 2014). A security approach that focuses on caring for the data that's collected by an organization instead of abusing it for financial gain is called data security (Tariq *et al*., 2020). Due to the complexity of big data analytics, it is required that privacy protection measures are upgraded to handle the increasing volume of data. This section aims to provide an overview of the various privacy protection options available in the market (Jee *et al*., 2013). Big data is constantly being spread across the globe due to the increasing number of data processing and storage systems. Various technologies, such as cloud computing and Hadoop map reduction, are being used to store and process this massive amount of data (Chen *et al*., 2016). Big data is a type of data that is stored and processed

PG & Research Department of Computer Science, Government Arts College (Grade-I), (Affiliated to Bharathidasan University) Ariyalur, Tamil Nadu, India

***Corresponding Author:** Mohamed Azharudheen A., PG & Research Department of Computer Science, Government Arts College (Grade-I), (Affiliated to Bharathidasan University) Ariyalur, Tamil Nadu, India, E-Mail: azhar.scas@gmail.com

How to cite this article: Mohamed Azharudheen, A., Vijayalakshmi, V. (2024). Improvement of data analysis and protection using novel privacy-preserving methods for big data application. The Scientific Temper, **15**(2):2181-2189.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.30

Source of support: Nil

Conflict of interest: None.

Received: 20/04/2024 **Accepted:** 25/05/2024 **Published**: 15/06/2024

in cloud computing. Its privacy is important when it comes to handling it (Stergiou *et al*., 2018). Big data requires massive amounts of storage and computation to analyze and interpret it properly. Cloud computing is the ideal solution for this type of data analysis (Singh *et al*., 2019). Despite the various issues related to data privacy, cloud computing still has advantages over traditional methods of computing.

Outsourcing

Outsourced data is a common issue that cloud users face when it comes to protecting their privacy. This is because they do not have control over the data that they are storing (Inukollu *et al*., 2014).

Multi-tenancy

The sharing of the same storage location among various cloud users makes it easy for a third party to access the data without any connection to it. This can lead to security issues and privacy concerns (Gupta *et al*., 2022).

Massive Computation

Due to the immense amount of computation that goes into managing a massive amount of data, traditional systems are not able to protect individual privacy (Rasi *et al*., 2022).

A hardware or software intrusion detection system (IDS) detects and prevents illegal entry to a network or system (Othman *et al*., 2018). When dealing with big data, traditional methods of detecting and preventing illegal access to a network or system become more complicated and ineffective (Najafabadi *et al*., 2015). The difficulty of the data analysis process can prevent an IDS from alerting its users immediately. Big data tools and techniques can be used to improve the efficiency of IDSs by reducing the training time required and the computation requirements (Lawal *et al*., 2020; Singh *et al*., 2019; Hagar *et al*., 2020).

The first generation of intrusion detection systems were mainly based on a signature that was predefined and configured to detect malicious activities. Unfortunately, this method is not very effective since attackers can easily exploit the database of attack signatures to carry out their activities. Machine learning eventually became a useful tool for detecting unknown attacks. It can analyze the activities of a user based on their normal behavior. Several techniques have been developed to improve the detection rate and reduce false positives. This review explores the use of various hybrid, single, and ensemble ML techniques in intrusion detection systems.

Literature Review

A cloud-based healthcare system was proposed by Rani *et al*., (2019), which would restrict access to data by unauthorized users. The system uses a statistical method known as SVM to predict the conditions and expected diseases of patients. It does so using an ML approach. Chakraborty *et al*., (2019) proposed framework did not perform any benchmark tests on the system. Despite the fact that blockchain technology is known to be secure, they did not investigate the framework's security features.

Alabdulatif *et al*., (2018)developed a system that can detect and predict abnormalities in a patient's vital signs using cloud-based technology. The system consists of three main blocks: The smart community resident, the cloud storage, and the data protection. The former collects and stores the data in an encrypted format, while the latter distributes the data to other users. The main component of this system is a smart prediction model, which can detect abnormal changes in the data. This approach does not consider the use of machine learning techniques for analyzing the data.

Tao *et al*., (2018), introduce a hardware-based approach to secure the internet of things (IoT)-based healthcare monitoring systems. The proposed solution involves the implementation of a secret cipher algorithm on a hardware platform. It is designed to provide robust security while collecting data. On the KDD 1999 dataset, Zhang *et al*., (2008) offer a security architecture that detects abnormal traffic using the RF approach. When it comes to finding anomalies in data, the RF approach is 95% accurate. Due to the KDD dataset (2007), which is utilized in many competitions, it has a 1% false-positive rate.

The following environments may be breaking the users' privacy in big data technology:

During the transmission of data over the internet, the information is shared with various external resources. This can lead to speculations about the users.

The data trickling occurred during the various phases of the data life cycle. In the first, second, and third phases, the importance of protecting the privacy of the individual is more important than the amount of data collected.

Traditional DBMS techniques are not designed to handle the immense amount of data that big data platforms can collect. Due to the complexity of the data, high velocity is required in order to manage its growth.

Methodology

Proposed Methodology for Intrusion Detection

This research proposes an efficient and effective intrusion detection system that utilizes a deep learning technique known as the DBN, as well as a parameter optimization technique known as PSO (Figure 1). In the hidden layer of the system, the PSO method increases the number of nodes, while the RBM network is trained according to the top-up approach, which significantly lowers the data's overall dimensionality. The goal of this research is to create a resilient and efficient RBM network that eliminates redundancy. Through a back-propagation method, the network is tuned to ensure that it operates properly.

The detection of anomalous network behavior through a DBN method is carried out by storing various forms

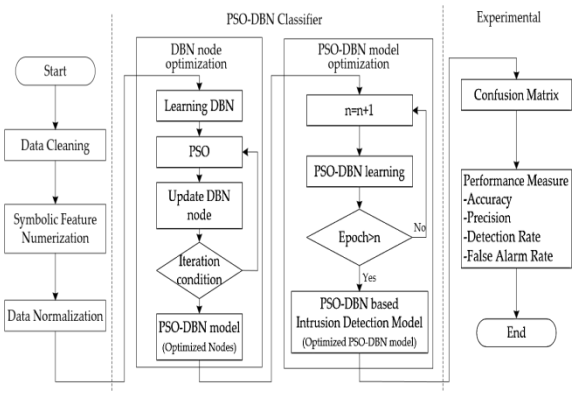

Figure 1: Spark-DBN-PSO architecture for intrusion detection

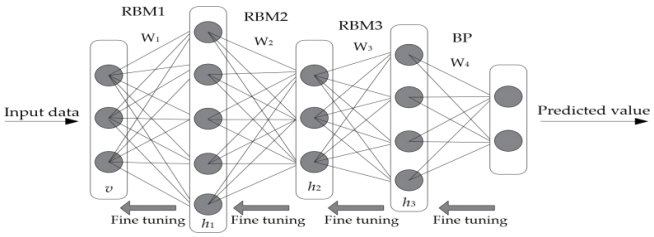

Figure 2: Detection model based on the deep belief network (DBN)

of network data in the input layer. This type of model is made up of RBMs. When utilized in an intrusion detection framework, the network's link weight and neuron bias must be trained first. Reverse fine-tuning and unsupervised learning are some of the processes used in the training phase of the DBN method. Each RBM layer is then trained independently to ensure that the collected features are kept. The RBM's input eigenvector is subsequently routed to the DBN's last layer. After that, an entity relationship classification training session is done under supervision. Each RBM layer has its own weights that are optimized for its feature vector mapping. The BP network is then used to spread error information across the RBM. This ensures that the network is tuned properly. The training of the DBN model is regarded as the starting point of the deep BP network's weights. This ensures that the network is tuned properly and overcomes the shortcomings of the BP network.

Apache Spark

A cluster of Spark is composed of several workers and master nodes. Before an individual can start with a machine learning project, they must first get their local models and datasets into a storage backend that's supported by Apache Spark, which is a popular open-source software. After the data has been loaded, the user can submit their machine-learning tasks to the Spark client. This client then forwards these tasks to the Master node. The Spark driver generates a spark context that enables the user to access the cluster's resources. This allows the cluster to distribute the tasks to the various Worker nodes. An independent worker can split a task into smaller steps. They can also launch a process that handles local processing simultaneously.

When a task requires the use of a storage backend, the Worker nodes typically use an abstraction called the resilient distributed datasets. The resilient distributed datasets representation helps the cluster's storage I/O latency by allowing the cluster's various Worker nodes to collocate data into shards. The data stored at each executor process is also kept in an in-memory cache to improve storage performance. As each step of a given task is performed by different workers, it's necessary that they can easily transfer information among themselves. The output of the workers is then sent to the master node, which is responsible for forwarding and merging the data.

Intrusion attack Recognition model based on the DBN-PSO

The model for intrusion detection is shown in Figure 2. It features a DBN input layer composed of various types of network data. When implementing a DBN network in an analysis, the first step is to train the network structure. The training phase of the DBN involves various steps, such as pre-training and reverse tuning. Each of the RBM network's layers is independently trained. To ensure that the necessary details are retained, a pre-training process is carried out. The mapping of vectors to the various feature spaces is carried out. The BP network is then set up in the last layer. The output eigenvector of RBM is then sent as the input eigenvector to the DBN. The training process for the classification of entity relationships is then carried out. Ensure that the weights in the respective layer are optimal for the mapping of its feature vectors. The BP network must also be used to distribute error information in the entire DBN. In the top-to-bottom layers of the RBM network, the training process is performed to fine-tune the DBN structure. The training of the model is regarded as the start of the weights of the deep BP network, which will allow the DBN to overcome its disadvantages.

A classification problem is usually associated with the detection of intrusions on a healthcare platform. The first step in the process is to preprocess the data. The paper presents a method that converts symbolic attributes into numerical ones. For instance, the protocol type attribute feature can be processed by the three values in column 2 of the data set.

The obtained data are then normalized according to equation 1, which ensures that the attributes are in the same order of their magnitude.

$$\bar{x}(i) = \frac{x(i) - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

The value of $\bar{x}(i)$ is the normalized form of the variable's original value. Its value is expressed as x(i). The minimum and maximum values of the variable are, respectively x_{min} and x_{max}.

The DBN network structure has been pre-established after the intrusion detection data has been processed. The PSO algorithm is then used to optimize the network structure in each layer. The learning rate, number of nodes, and the DBN's hidden layer's number of nodes are some of the parameters that are commonly considered when it comes to establishing a good network structure. The learning rate is a measure of the progress that the model makes in its learning process. It can be set by users through various learning materials. For instance, the optimal learning rate can be determined by analyzing experience values. The number of network layers, on the other hand, is a more complicated calculation. In this paper, the dataset used is not very large, and the network layers that were selected meet the needs of intrusion detection. One of the most important factors that can affect the performance of a DBN network model is the number of nodes in its various layers. The formulas used in most literature for calculating the number of nodes in a given layer are usually based on large training samples. The results of these formulas are not optimal, and the variations in the obtained number significantly affect the performance of the network model. In order to avoid overfitting during training, it is important that the number of network nodes in each layer is optimized.

The number of nodes in the DBN's hidden layer is optimized through the selection of the fitness function for the model. The condition of the PSO ensures that the number of nodes is constantly updated. The supervised learning process for the PSO-DBN model is performed after the model has been completed. It involves updating the weights of the nodes using BP. This learning is carried out after a set of epochs has been assigned.

Security model based on Spark DBN-PSO

The goal of this model is to provide a secure environment for the outsourcing of inference and training workloads for machine learning. This is typically done in scenarios where the owner of a private dataset has sensitive information. This type of setting is commonly used in health-related fields, such as identifying rare diseases. Even if a cloud service can process large amounts of patient data, healthcare facilities are reluctant to transfer this data to it due to security concerns. In contrast, insurance firms can use this data to predict traffic accidents. Although a company can greatly benefit from using the cloud resources of a third-party service, doing so could expose the predictive model to unauthorized access.

This paper aims to provide a framework for protecting the data owner from a malicious adversary that is trying to target its data. It assumes that the client will not perform malicious queries or put false data into the model training. The data owner is responsible for enforcing the access control mechanisms that are used to manage its data. These are widely used and are documented in previous research. The primary goal of this security measure is to ensure that the data collected by clients is protected from unauthorized access. We want our system to act as a black box for Apache Spark ML scripts so that it can perform proper inference and training. The security model we use considers both semi-honest and honest adversaries. Semi-honest adversaries are usually considered when there is a breach in the confidentiality of the model and data. Malicious entities can exploit this type of vulnerability to access the processing data. Nevertheless, we also provide additional countermeasures against the exploitation of adversarial queries. These include the injection of samples into the training model or the backend.

We assume that the data collected by SparkNodes, including the duration and number of steps taken, are explicit leaks. This is because, even though they use secure channels, an adversary can still observe their network communication and deduce these parameters. To minimize the privacy of this data, implementing fixed-size outputs and constant-round execution would require additional steps.

Result and Discussion

The data of patients is stored on a network. This means that the actions that happen in the data can be monitored if you have the right system. An IDS can monitor the activities in your network to see if they are suspicious. If someone logs in with administrative credentials at 3:00 a.m. It's possible that they are trying to access your sensitive information. The actions that happen in the network around your sensitive health information can be used to identify what's happening inside it.

Dataset Description

We consider the MIMIC-III database (https://mimic.physionet.org/gettingstarted/dbsetup/), which was collected from the intensive care unit of a hospital in Boston, Massachusetts, USA. The MIMIC-III database is the world's largest repository of hospital data. It contains information about over 52,726 critically ill patients. It is an open-source data platform that researchers can use to collect and analyze patient data.

Environment

The experiments were carried out on a Cloudera 6.3 cluster consisting of eight Dell Optiplex 3070 Small-Form desktop computers. They were powered by an Intel Core i5-9500 CPU and 16 GB RAM. The cluster's host operating system is Ubuntu 18.04.4 LTS, with a Linux kernel version of 4.15.0. Each machine has a 10Gbps Ethernet card. The experiments were performed on one server, which houses the Spark Master and the client. The remaining seven servers are used for the Spark Workers.

Simulation Results

In today's world, the security of scientific papers is of utmost importance. This is because, in order to prevent unauthorized

access, organizations must implement effective security auditing mechanisms. An intrusion detection system is used to monitor and detect anomalous activities. This system is designed to prevent unauthorized access to a system. This system can be categorized into two parts: the surveillance domain and the control domain. The latter can monitor a wide variety of applications and corporate networks. Figures 3 to 11 shows the implementation results of the proposed framework.

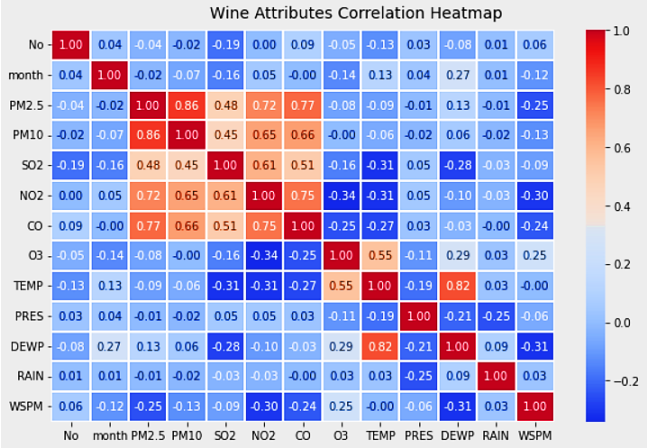

Figure 3: Heat map for attributes

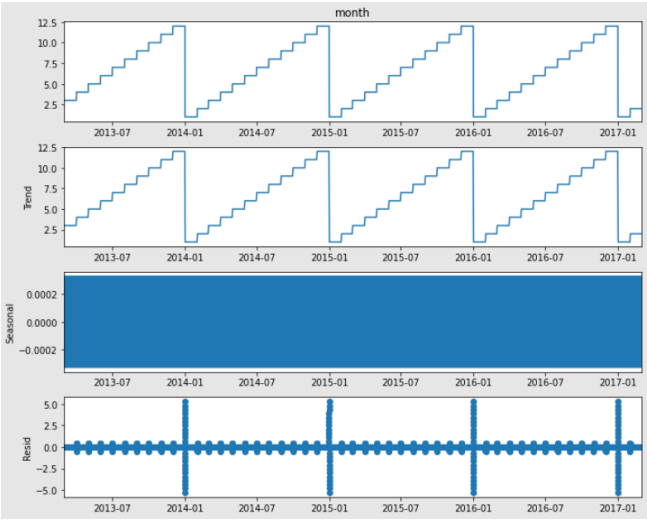

Figure 4: Intrusion identification based on test id, seasonal and trend

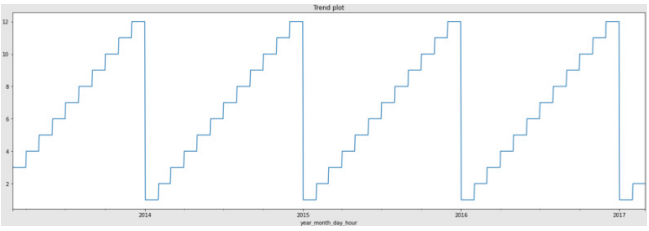

Figure 5: Trend plot

Normal traffic

Traffic from normal hosts that have been validated previously and is particularly valuable for determining actual performance.

Traffic from malicious hosts or robots is known as malware or botnet traffic.

Background traffic is defined as unidentified traffic that is used to saturate algorithms in order to test their speed.

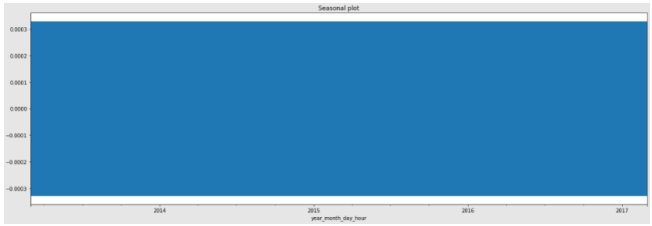

Figure 6: Seasonal plot

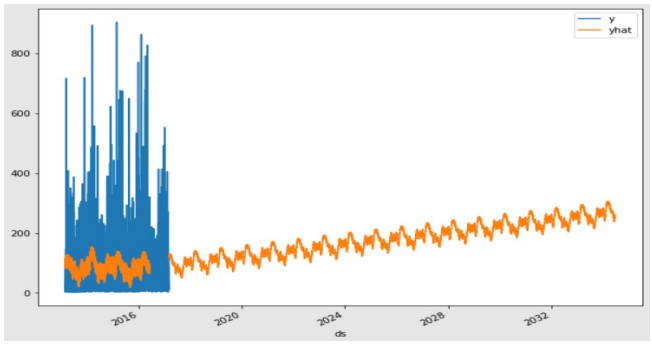

Figure 7: Intrusion prediction plot

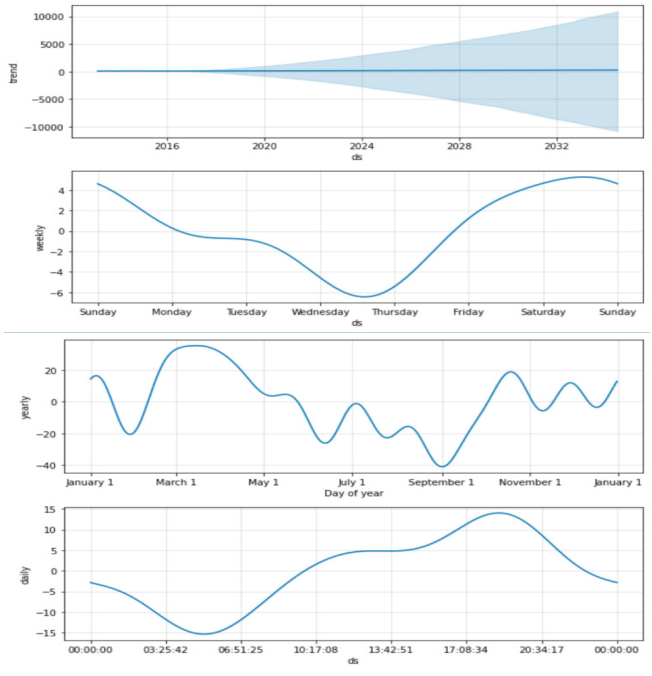

Figure 8: Intrusion detection based on year, week, month and daily hours

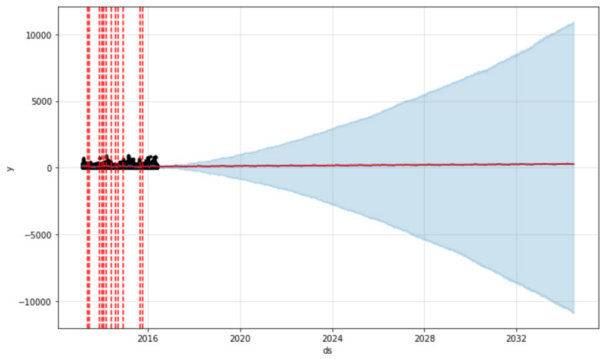

Figure 9: Intrusion change point

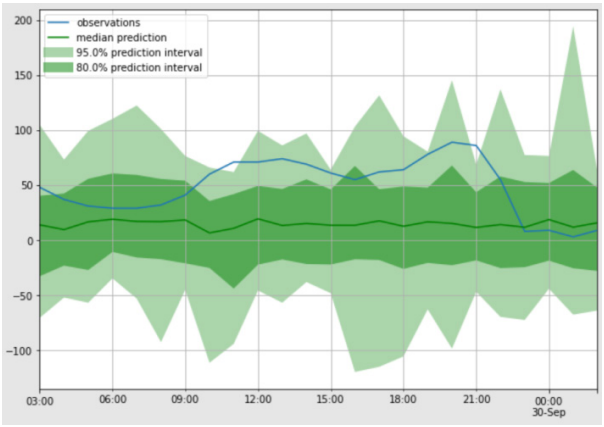

Figure 10: Prediction performance analysis

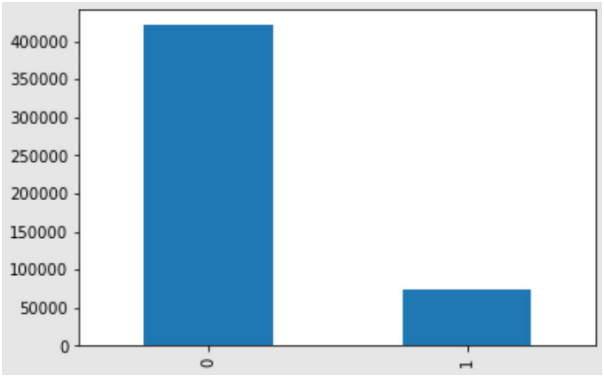

Figure 11: Normal *vs* abnormal action

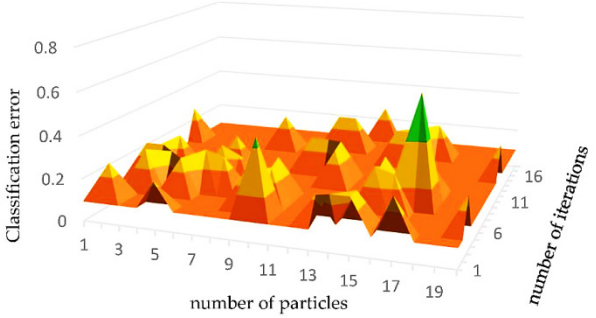

Figure 12: Optimization results of the PSO algorithm

To increase the number of hidden layer nodes, the PSO algorithm is used. The findings of this algorithm are presented in this publication. By optimizing the classification error, the PSO method can increase the number of hidden layer nodes. The study's findings (Figure 12) suggest that adjusting the number of hidden layer nodes at a minimum error of 0.0923 yields the best outcomes.

Comparison of the Proposed System with Existing Techniques

The suggested system is compared to current classifiers such as decision trees, random forests, and RBF in this section. The following metrics are taken into account when evaluating the proposed system: Precision, recall, F-measure, and accuracy (Tables 1 and 2).

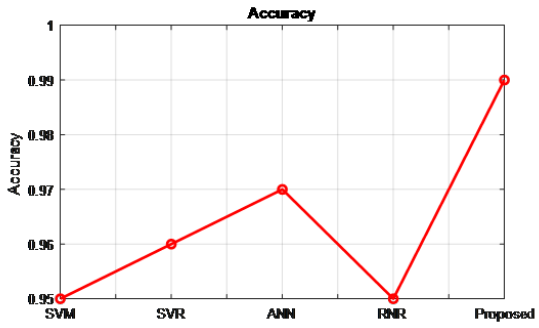

Figure 13: Accuracy of the proposed system along the existing techniques

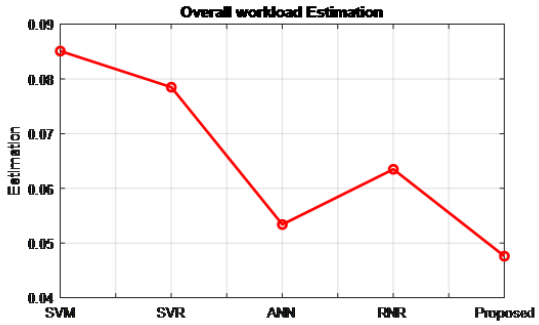

Figure 14: Overall workload estimation of the proposed system along the existing techniques

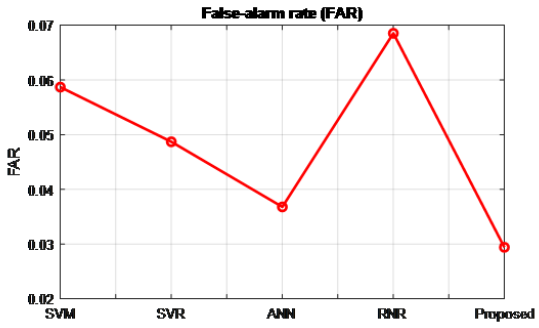

Figure 15: False alarm rate

Table 1: Threat analysis comparison

Methods	*Accuracy*	*Overall workload estimation*	*false-alarm rate (FAR)*	*various infection rates*	*efficiency*	*number of iterations*
SVM	0.95	0.0851	0.0587	0.0245	1.025	358
SVR	0.96	0.0785	0.0487	0.02854	0.987	350
ANN	0.97	0.0534	0.0368	0.0355	1.025	320
RNR	0.95	0.0635	0.06854	0.0246	1.069	250
Proposed	0.99	0.04761	0.02941	0.0119	1.1904	234

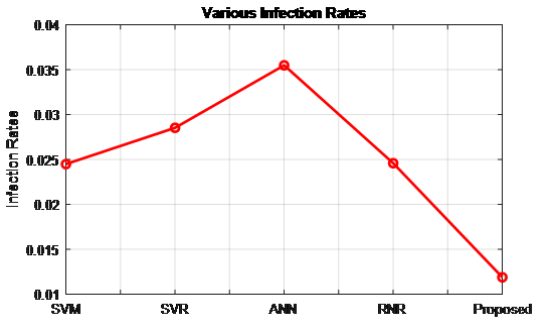

Figure 16: Various infectionrate

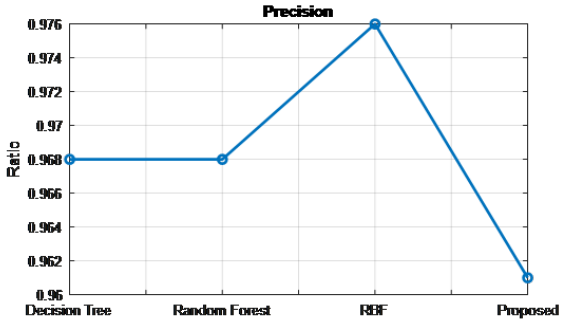

Figure 19: Comparison graph for precision

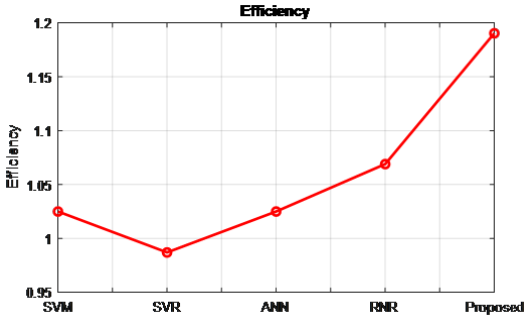

Figure 17: Efficiency

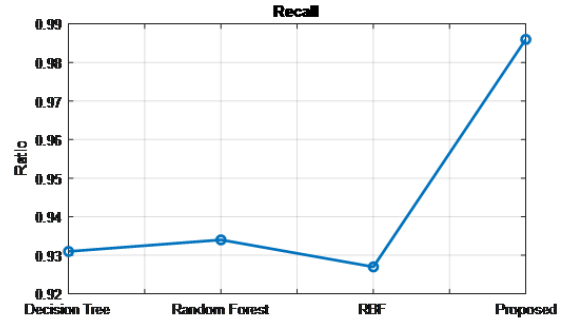

Figure 20: Comparison graph for recall

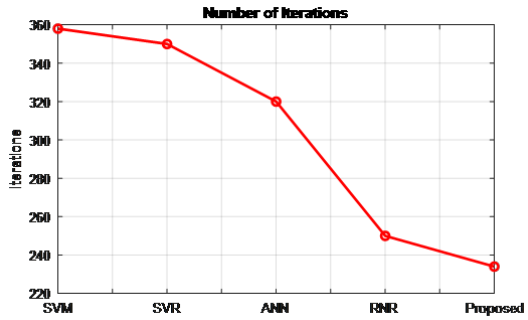

Figure 18: Number of iterations

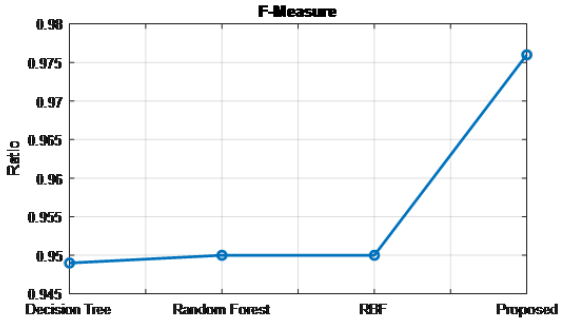

Figure 21: Comparison graph for F-measure

Table 2: List of classifiers with proposed system

Classifiers	*Precision*	*Recall*	*F-measure*	*Accuracy*
Decision tree	0.968	0.931	0.949	96.5333
Random forest	0.968	0.934	0.95	96.667
RBF	0.976	0.927	0.95	96.5333
Proposed	0.961	0.986	0.976	99.04

In comparison to existing classifiers such as RBF, decision tree, and random forest, the suggested system has an ideal precision value of 0.961. It was used to count the number of events that occurred in a certain time frame using a regression approach.

The suggested system has a recall value of 0.986, which is higher than RBF, decision tree, and random forest classifiers.

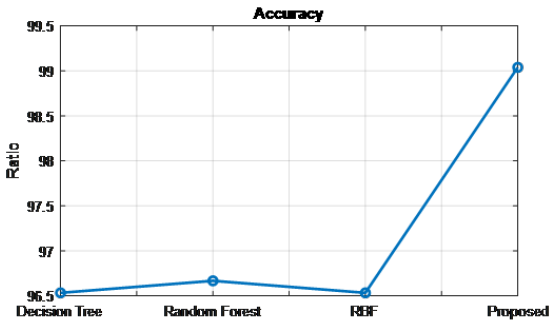

Figure 22: Comparison graph for accuracy

It has also correctly segregated the events using both top-down and bottom-up clustering techniques.

The proposed system has been able to record all the rare events that happened in the multi-cloud environment. It has gained a better F-measure value of 0.976 as against other classifiers such as RBF, decision tree, and random forest (Figures 13-21).

In the above, Figure 22 is used to capture unusual occurrences, and an ensemble-learning approach is used to segregate assaults based on their properties in the proposed system. When compared to existing classification methods such as random forest, decision tree, and RBF, the suggested system has a higher accuracy of 99.04.

Conclusion

One of the most critical factors in protecting medical data is the detection of unauthorized access. Deep learning techniques can be used to identify potential threats in complex data sets. The ability to perform tasks with unsupervised learning can also help the DBN identify breaches in large and diverse databases. The classification and extraction capabilities of the DBN are ideal for medical security. In addition, it can be utilized to enhance the network structure by identifying hidden layer nodes. A study revealed that the PSODBN algorithm, which is a part of deep learning, has an accuracy rate of 99.04%, making it superior to other techniques such as ANN and SVM. The PSO-DBN algorithm's optimization effect is better than that of other deep learning methods. In addition, it can be utilized to perform tasks such as detecting intrusions and extracting feature vectors.

Acknowledgment

I am extremely grateful to my supervisor, Dr. V. Vijayalakshmi, for her invaluable advice, continuous support, and patience during my studies. Her immense knowledge and plentiful experience have encouraged me in all the time of my academic research.

References

Alabdulatif, A., Khalil, I., Forkan, A. R. M., & Atiquzzaman, M. (2018). Real-time secure health surveillance for smarter health communities. *IEEE Communications Magazine*, *57*(1), 122-129.

Chakraborty, S., Aich, S., & Kim, H. C. (2019, February). A secure healthcare system design framework using blockchain technology. In *2019 21st International Conference on Advanced Communication Technology (ICACT)* (pp. 260-264). IEEE.

Chen, H. M., Chang, K. C., & Lin, T. H. (2016). A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs. *Automation in Construction*, *71*, 34-48.

Gupta, M., Ahuja, L., & Seth, A. (2022). A Study on Cloud Environment: Confidentiality Problems, Security Threats, and Challenges. In *Soft Computing for Security Applications: Proceedings of ICSCS 2021* (pp. 679-698). Springer Singapore.

Hagar, A. A., Chaudhary, D. G., Al-Bakhrani, A. L. I. A., & Gawali, B. W. (2020). Big Data Analytic Using Machine Learning Algorithms For Intrusion Detection System: A Survey. *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)*, *10*, 6063-6084.

Hou, R., Kong, Y., Cai, B., & Liu, H. (2020). Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning. *Neural Computing and Applications*, *32*, 5399-5407.

Inukollu, V. N., Arsi, S., & Ravuri, S. R. (2014). Security issues associated with big data in cloud computing. *International Journal of Network Security & Its Applications*, *6*(3), 45.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*, 1-25.

Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthcare informatics research*, *19*(2), 79-85.

Lawal, M. A., Shaikh, R. A., & Hassan, S. R. (2020). An anomaly mitigation framework for iot using fog computing. *Electronics*, *9*(10), 1565.

Lv, Z., & Qiao, L. (2020). Analysis of healthcare big data. *Future Generation Computer Systems*, *109*, 103-110.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. *et al.* (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data* **2**, 1.

Othman, S. M., Ba-Alwi, F. M., Alsohybe, N. T., & Al-Hashida, A. Y. (2018). Intrusion detection model using machine learning algorithm on Big Data environment. *Journal of big data*, *5*(1), 1-12.

Rani, A. A. V., & Baburaj, E. (2019). Secure and intelligent architecture for cloud-based healthcare applications in wireless body sensor networks. *International Journal of Biomedical Engineering and Technology*, *29*(2), 186-199.

Rasi, D., & Mahaveerakannan, R. (2022, February). A Survey on Algorithms in Game Theory in Big Data. In *International Conference on Computing, Communication, Electrical and Biomedical Systems* (pp. 155-166). Cham: Springer International Publishing.

Singh, S. P., Nayyar, A., Kumar, R., & Sharma, A. (2019). Fog computing: from architecture to edge computing and big data processing. *The Journal of Supercomputing*, *75*, 2070-2105.

Singh, U. K., Joshi, C., & Kanellopoulos, D. (2019). A framework for zero-day vulnerabilities detection and prioritization. *Journal of Information Security and Applications*, *46*, 164-172.

Stergiou, C., Psannis, K. E., Gupta, B. B., & Ishibashi, Y. (2018). Security, privacy & efficiency of sustainable cloud computing for big data & IoT. *Sustainable Computing: Informatics and Systems*, *19*, 174-184.

Swan, M. (2012). Health 2050: The realization of personalized medicine through crowdsourcing, the quantified self, and the participatory biocitizen. *Journal of personalized medicine*, *2*(3), 93-118.

Tao, H., Bhuiyan, M. Z. A., Abdalla, A. N., Hassan, M. M., Zain, J. M., & Hayajneh, T. (2018). Secured data collection with hardware-based ciphers for IoT-based healthcare. *IEEE Internet of Things Journal*, *6*(1), 410-420.

Tariq, N., Qamar, A., Asim, M., & Khan, F. A. (2020). Blockchain and smart healthcare security: a survey. *Procedia Computer Science*, *175*, 615-620.

Tewari, A., & Gupta, B. B. (2020). Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework. *Future generation computer systems*, *108*, 909-920.

UCI KDD Archive. (Oct. 2007). KDD Cup 1999 Data. Accessed: Mar. 7, 2020. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Viji, D., Saravanan, K., & Hemavathi, D. (2017, June). A journey on privacy protection strategies in big data. In *2017 international conference on intelligent computing and control systems (ICICCS)* (pp. 1344-1347). IEEE.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: privacy and data mining. *Ieee Access*, *2*, 1149-1176.

Zhang, J., Zulkernine, M., & Haque, A. (2008). Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(5), 649-659.