Big Data-Based Autonomous Anomaly Detection Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing

P. M. Diaz and M. Julie Emerald Jiju

1 Introduction

Virtualization is the potential of operating various control systems nearly unique physical machines and dividing the fundamental hardware resources. It is applicable to enhance everywhere along with costs by utilizing the physical resource quality group. Virtualized cloud-based computing deployed the essential platform to operating systems and storage units. Cloud virtualization handles the workload by converting usual data processing to make better hierarchal, cost-effective and powerful environments. One of the principal aspects of virtualizations is its capability to facilitate the combined use of applications toward different clients [1, 2].

For example, a recent study [3] proposed an unverified learning-based autonomous anomaly detection method that causes precise outcomes anomaly detection regarding a cost-effective method. It was not a unique impact through structures data, while may be able to manipulate alike common and separate anomalies successfully. In this method, [4] applies Autonomous anomaly detection to determine the statistics approach with an intention to seek particular rare occurrence techniques. Numerous enrolments encompass Autonomous anomaly detection; for instance, detect illegal activity, conflagration, human body monitoring, etc. The study in [5] described that virtualization is one of the enormous advances in cloud computing. Another example, [6] focuses only on the way to enlarge particular capability or else append distinctly feasible to the current configuration in recent infrastructure. Negi et al. [7] reported

P. M. Diaz (✉)
Department of Mechanical Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India
e-mail: pauldiaz@sreyas.ac.in

M. Julie Emerald Jiju
Department of MCA, CSI Institute of Technology, Thovalai, Kanyakumari, Tamil Nadu, India

R. Agrawal et al. (eds.), *Cyber Security in Intelligent Computing and Communications*,
Studies in Computational Intelligence 1007,
https://doi.org/10.1007/978-981-16-8012-0_6

that Cloud Computing is a comparatively a new approach, which, moreover, provides a significant range suitable to anticipate end-user.

This [8] describes a file management system configured to store a large number of data throughout numerous nodes of commodity hardware. Hadoop was originally established in 2007, meanwhile open source Implementation of the MapReduce processing engine joined with a distributed file system. In this category, it [9] used Machine learning techniques to detect the unusual posture of services. Virtualization takes complex reliability and high availability process to meet the high solution for standard essential. For example, [10] described that Cloud computing leads to numerous securities to the organization. The component which affects cloud computing adoption and attacks is an applicable solution to build up surveillance and solitude within the Cloud environment.

It is considerable to refer that our research is diverse in accordance to the above stated investigation in different usages. On the other hand, most of the current analyses commonly assessed a large amount of data processing, anomaly detection, unsupervised learning mostly emphasized clustering processing as opposed to real-time processing. Conversely, we have mainly focused on autonomous anomaly detection security analytics for protecting virtualized infrastructures in cloud computing containing clustering and classification methods. Datasets collected from virtual machines are stored in Hadoop Distributed File System. A new method of autonomous anomaly detection centered on k-means clustering is used. It applies four different classification methods, such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) and Artificial Neural Network (ANN). Classification methods individually show the accuracy of ROC. Eventually, the Receiver Operator Characteristics Curve (ROC) shows the graphical way connection Specificity and Sensitivity.

The remainder of this paper is organized as follows. Section 2 briefly brings out the Literature review. Section 3 explains the methodology offered. Section 4 defines the investigational outcomes and discussion. Section 5 describes the conclusion.

2 Literature Review

Security data analysis for Virtualized Infrastructures in Cloud Computing by Autonomous Anomaly Detection depends on a dependable circumstance building approach; many academics have suggested various methods to create background models [11, 12].

An essential allusion for our research is the work on collocating [13]. This work showed that huge quantities of data processing and machine learning in anomalous detection have the explication to examine. The system is suggested for clustering/classification algorithms, big data processing technologies and analytics for anomaly detection. The author also shows accuracy and efficiency in the detection rate with the algorithm. The security cloud introduced by Wei et al. [14] states an accumulator security log for cloud customer’s data. Furthermore, it merely obtains the stored information besides that states security on computer-based information. Shantanu et al. [15] stated that security is a reasonable major challenge through storing data in the cloud computing system. This [16] described cloud computing, virtualization increases the security by securing both the integrity of guest virtual machines and the cloud infrastructure components. For example, the study in [17] contemplated the security problems in massive data, cloud computing and the Hadoop environment. The fundamental of this study emphasized the security issues, thereby processing big data in the data processing. It [18] examined the issue related to the security in virtualized infrastructure for identifying the current threats and complexity in the cloud platform. In which, [19] indicates that the detection of anomalies secures the environment in cloud computing based on the concept of Big Data, although it uses only one classification method to prove a highly accurate algorithm. In this analysis, Mahendiran et al. [20] put forward the K-Means clustering algorithm, one of the very popular and high-performance clustering algorithms was compared with other algorithms. A present study by Ahmad et al. [21] uses many unsupervised machine learning algorithms for anomaly detection. The recent study by Win et al. [22] figured out a model, which interpolates the attribute data items and assures that it is well generalized to detect unknown attacks.

3 Proposed Methodology

The fundamental principle of the proposed approach is to detect the autonomous anomaly detection analytics for protecting virtualized infrastructures in cloud computing, which, occasionally, collect the large dataset from Comodo Cloud Security Centre. Attributes, Pe-Files-Malwares, Malware Good ware and Malware-Analysis datasets are collected from virtual machines. Clustering and Classification method is preowned for security analysis. In the process, K-means algorithms perceive anomalies from the dataset. Later, Big Data-Based Autonomous Anomaly Detection Security Analytics applies four different classification methods like KNN, SVM, RF and ANN. Each classification separately checks the accuracy by Receiver Operator Characteristics (ROC).

3.1 Autonomous Anomaly Detection Using K-Means

Anomaly detection is the unity of outstanding events, module or identification of rare occurrences or events of concern due to their differing characteristics from majority of the processed data. It is used in applications such as fraud and intrusion detection, system health monitoring and ecosystem disturbance monitoring. The proposed method uses a novel method of Autonomous Anomaly Detection (AAD) to detect independent anomalies from a large dataset. The AAD method describes the anomaly detection approach, which is implemented using MATLAB.

3.2 K-Means

Clustering is one of the most common probing data analysis techniques used to get a hunch about the structure of the data. Clustering analysis is generally used in many applications in particular image processing, pattern recognition, data analysis and market research. K-means is one of the unsupervised learning algorithms that solves the common clustering problem. The algorithm recurrences to assign each data point to one of the K groups based on the features that are provided. The main idea is to define k centers, one for each cluster. Suppose, the required dataset has ‘n’ objects and the partitioning method (k-means) constructs ‘k’ partitions of data, each and every partition will signify as a cluster. It means that it will classify the data into k groups that satisfy the $k \leq n$ requirement. Using K-means, unrelated data will be modeled as clusters. Subsequently, it is described as similar and dissimilar groups. That is, in some of the rare cases, one or two separate data can be detected, in a way, that cases initiate it as a cluster. The K–Means clustering uses distance-based measurements to determine the similarity between data points. K-medoids clustering method is more similar to K-Means method. Compared to other cluster methods, AAD clustering method is well organized.

3.3 Classification Methods

In supervised learning, both input and output data are provided. Input and output data are labeled for classification to provide a learning basis for data processing. Two techniques used in supervised learning are linear regression and classification techniques. Linear regression model gives the relationship between quantitative data. It is a statistical method that is used for predictive analysis. Classification is used to group the uniform data points into different components to classify them. Particularly, this method consists of four classifications: (i) K-Nearest Neighbor, (ii) Support Vector Machine, (iii) Random Forest and (iv) Artificial Neural Network.

3.4 K-Nearest Neighbor

The k-nearest neighbor algorithm is a simple supervised machine learning algorithm. It can be used to solve both classification and regression problems. All the accessible data are stored and arranged in a new discrete unit of information based on the similarity. The nearest neighbor selects K training cases that have the smallest distance where ‘K’ denotes the required value whether it is maximum or minimum to the nearest neighbor.

3.5 Support Vector Machine

A support vector machine is a supervised machine learning algorithm that analyzes and recognizes patterns. Data are used to solve both classification and regression analysis. A support vector machine is also known as a support vector network. The algorithm creates a line or a hyper-plane, which separates the data into classes.

3.6 Random Forest

Random forest classifier is an entity classifier that produces multiple decision trees. It can be used for both classification and regression. A number of m input variables are used to regulate the decision at a node of the tree. Random forest algorithm gives a more accurate estimate of error rate when compared with the decision tree.

3.7 Artificial Neural Network

An Artificial Neural Network is an arithmetic model based on the structure and functions of biological neural networks. There are hundreds or thousands of artificial neurons called processing units, which are interlinked by nodes. These processing units are devised by input and output units. The artificial neural network is used as a random function estimate tool.

4 Experimental Result

MATLAB is a programming environment for algorithm development, data analysis, visualization and numerical computation developed by Math Works. The proposed Big Data-Based Autonomous Anomaly Detection Security Analytics was implemented using the MATLAB R2018a platform.

The succeeding datasets: (i) Attribute (ii) PE-Files-Malware (iii) Malware Good ware and (iv) Malware Analysis data sets are taken into consideration for execution. The pictorial imageries of both the feature (Dataset-1) and PE-Files-Malware (Dataset-2) datasets are provided in this article. Figures 1 and 2 depict Autonomous Anomaly detection using K-means clustering. In each data for classification, there is a positive and negative class. Positive class denotes blue and negative class denotes black. In 1 attribute dataset graph, x denotes x1 and y denotes x2. In PE-Files Malware dataset graph, x denotes E cblp *104 and y denotes E cp *104.

Figures 3 and 4 depict the ROC curve for KNN in two datasets. It measures the accuracy in the ROC of the system. The suggested techniques provide improved

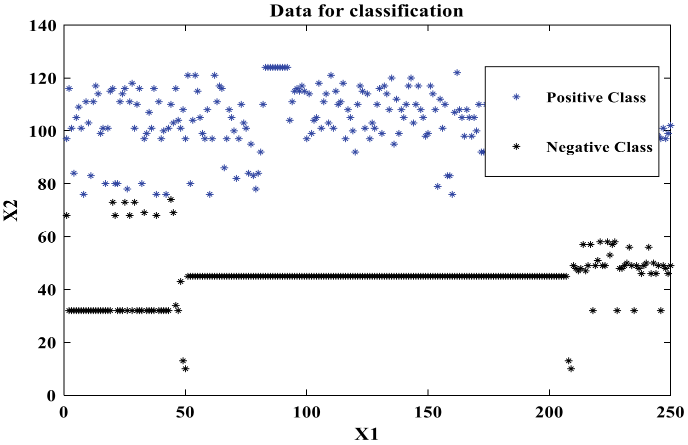

Fig. 1 Dataset 1—autonomous anomaly detection using K-means clustering

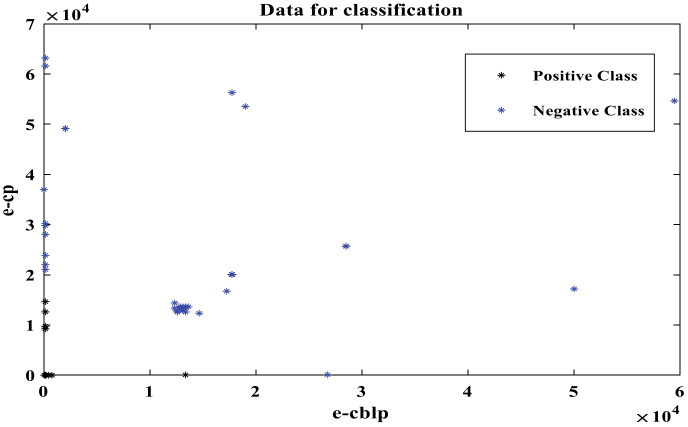

Fig. 2 Dataset 2—autonomous anomaly detection using K-means clustering

accuracy results for data classification. From the result, it is observed to have 1 sensibility and specificity. The effectiveness of a classifier is to identify positive labels and negative labels. The red circle denotes 1 sensibility and specificity.

Figures 5 and 6 illustrate the ROC curve for SVM in datasets. In accordance result, it determines accuracy in ROC of the method. The proposed method ensures enhanced accuracy in the classification of the datasets. From the result, it is found to have 1 sensibility and specificity. The performance of a classifier is to distinguish positive labels and negative labels. The red circle implies 1 sensibility and specificity.

Figures 7 and 8 demonstrate the ROC curve for RF in datasets. The implementation measures accuracy in the ROC of the system. The implied process affords more effective accuracy in the classification of the datasets. Through a result, it is observed

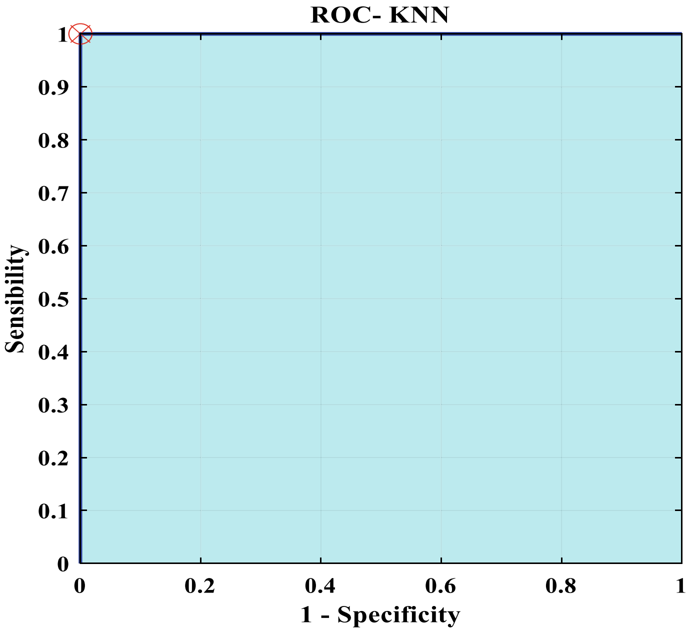

Fig. 3 Dataset 1—ROC curve for KNN

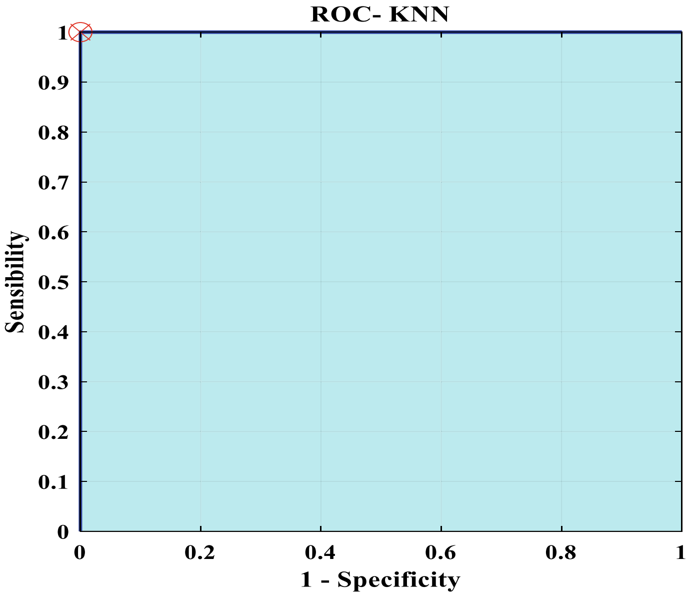

Fig. 4 Dataset 2—ROC curve for KNN

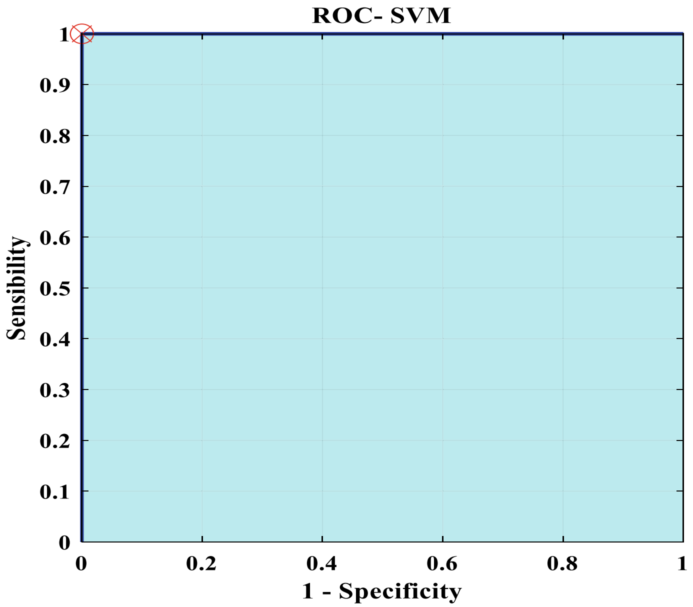

Fig. 5 Dataset 1—ROC curve for SVM

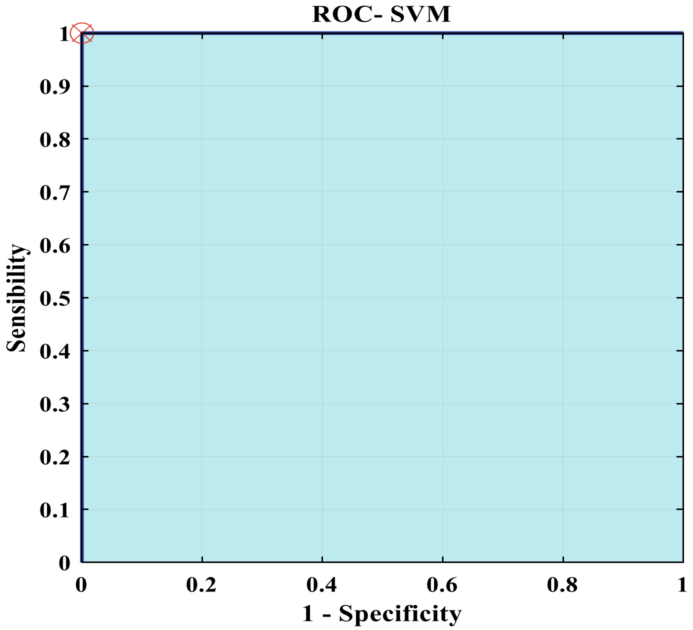

Fig. 6 Dataset 2—ROC curve for SVM

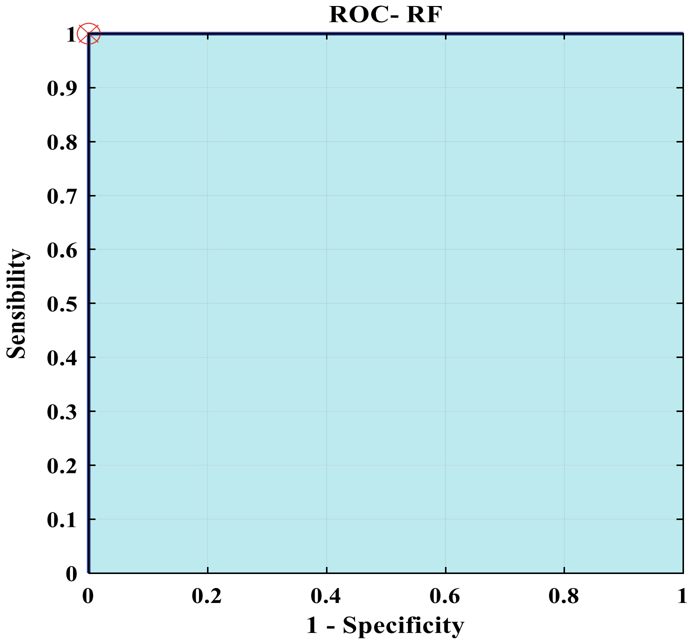

Fig. 7 Dataset 1—ROC curve for RF

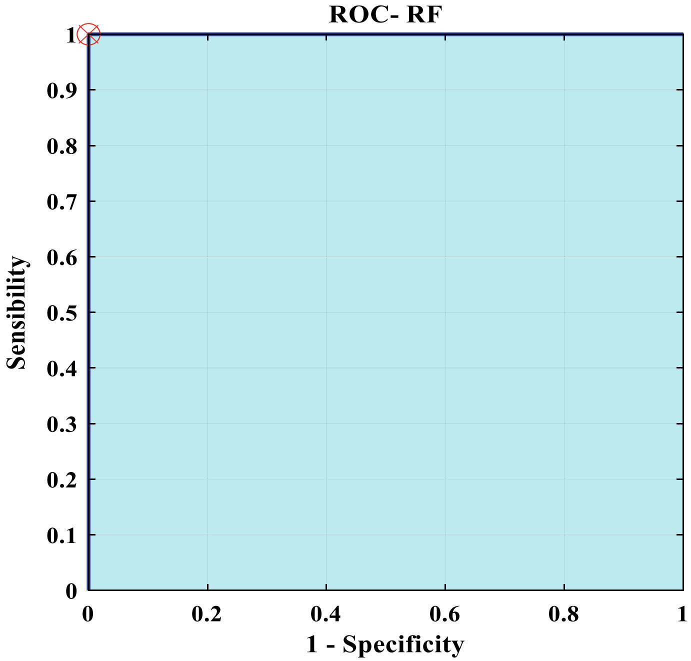

Fig. 8 Dataset 2—ROC curve for RF

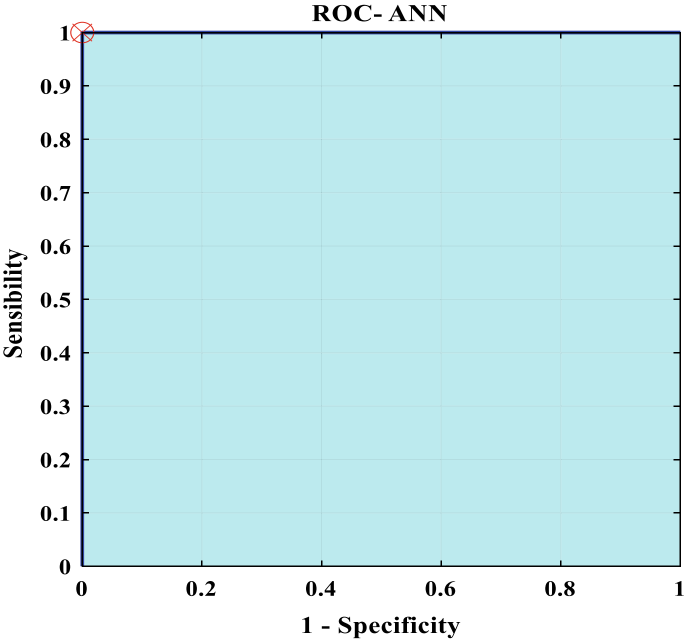

Fig. 9 Dataset 1—ROC curve for ANN

to have 1 sensibility and specificity. The influence of a classifier is to determine positive labels and negative labels. The red circle signifies 1 sensibility and specificity.

Figures 9 and 10 elucidate the ROC curve for ANN in datasets. From the result, it measures accuracy in the ROC of the system. The indicated expertise contributes to rectified accuracy in the classification of the datasets. Among results, it is observed to have 1 sensibility and specificity. The effectiveness of a classifier is to specify positive labels and negative labels. The red circle designates 1 sensibility and specificity.

Datasets are collected from the virtual machine and stored in Hadoop Distributed System, and K-means clustering is done with four classifications. In each dataset for classification, there is a positive and negative class. In the Roc curve, there is a graphical way connection between sensitivity and specificity. We discuss each of the classifications with accuracy 1. The proposed approach can detect anomalies in high accuracy by the datasets. In some cases, the best accuracy is produced at the cost of high computational processing and time. ROC curve is a graph airing the performance of a classification model at all classification thresholds. It contains two parameters: True Positive and True Negative. Sensitivity identifies positive labels and specificity identifies negative labels. The result highlights four datasets. An attribute dataset is an important type of semantic property shared among different activities. In attribute dataset, the accuracy and ROC of four classifications are 1. It gives more information about the functionality of the malware and how the malware interacts with OS. In the four classifications, PE-file's malware dataset contains the accuracy 0.9998 and ROC

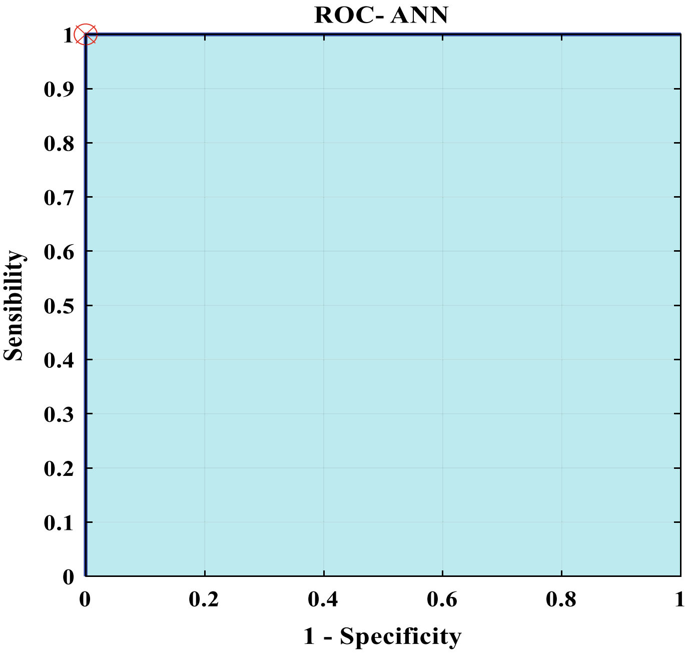

Fig. 10 Dataset 2—ROC curve for ANN

1. Malware analysis dataset contains static analysis data, the accuracy is 0.9946 and ROC is 1. In the Malware Good Ware dataset, the accuracy is 0.9946 and ROC is 1. In the approach, four classifications calculate the Receivers Operating Characteristic Curve, which represents 1—Specificity and Sensitivity.

5 Conclusion

This paper proposed a novel autonomous anomaly detection technique for securing the virtualized infrastructure in cloud computing based on clustering and classification. In this approach, the datasets are initially filtered by using K-means clustering. The clustered data are then directed into classifiers to obtain accurate classification outcomes. Different classifiers considered for evaluation include KNN, SVM, RF and ANN. The performance of classification can be observed through ROC parameters of individual models. The proposed method is found to achieve the best results in all the datasets. Therefore, it can be applied in virtual computer networks to provide full data security with attack detection. The future work is to integrate a DaaS software-defined networking with the present security analytic system to establish advanced security to the cloud networks. It is also significant to improve the framework through the experience acquired from the developers.

References

1. V. Ratten, Cloud computing technology innovation advances: a set of research propositions, in *Disruptive Technology: Concepts, Methodologies, Tools, and Applications* (2020), pp. 693–703
2. H. Shukur, S. Zeebaree, R. Zebari, D. Zeebaree, O. Ahmed, A. Salih, Cloud computing virtualization of resources allocation for distributed systems. J. Applied Sci. Tech. Trends. **1**(3), 98–105 (2020)
3. P.P. Angelov, X. Gu, Applications of autonomous anomaly detection. Stud. Comput. Intell. 249–259 (2018)
4. X. Gu, P. Angelov, Autonomous anomaly detection, in *Evolving and Adaptive Intelligent Systems (EAIS)* (2017)
5. B. Rohit, C. Rituparna, C. Nabendu, S. Sugata, A survey on security issues in cloud computing. Acta Tehnica Corviniensis – Bull. Eng. Tome. 160–177 (2014)
6. K. Rakesh, Applications of cloud computing in academic libraries. Library Waves **3**(1) (2017)
7. A. Negi, M. Singh, S. Kumar, An efficient security farmework design for cloud computing using artificial neural networks. Int. J. Comput. Appl. **129**(4), 17–21. November 2015. Published by Foundation of Computer Science (FCS), NY, USA
8. S. Liu, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J. Big Data **2**(1) (2015)
9. A. Gulenko, M. Wallschlager, F. Schmidt, O. Kao, F. Liu, Evaluating machine learning algorithms for anomaly detection in clouds. IEEE Int. Conf. Big Data (Big Data) (2016)
10. C. Modi, D. Patel, B. Borisaniya, A. Patel, M. Rajarajan, A survey on security issues and solutions at different layers of Cloud computing. J. Supercomput. **63**(2), 561–592 (2012)
11. B. Asvija, R. Eswari, M.B. Bijoy, Security in hardware assisted virtualization for cloud computing—state of the art issues and challenges. Comput. Netw. **151**, 68–92 (2019)
12. M. Jouini, L.B.A. Rabai, A security framework for secure cloud computing environments, in *Cloud Security: Concepts, Methodologies, Tools, and Applications* (2019), pp. 249–263
13. H.R.A. Ariyaluran, F. Nasaruddin, A. Gani, H.I.A. Targio, E. Ahmed, M. Imran, Real-time big data processing for anomaly detection: a Survey. Int. J. Inf. Manag. (2018)
14. L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, A.V. Vasilakos, Security and privacy for storage and computation in cloud computing. Inf. Sci. **258**, 371–386 (2014)
15. K. Shantanu, J. Hiteshkumar, U. Kaushiki, Providing classification and security of Big Data in Cloud computing. Int. J. Tech. Res. Appl. **4**(2), 302–304 (2016)
16. F. Lombardi, R. Di Pietro, Secure virtualization for cloud computing. J. Netw. Comput. App. **34**(4), 1113–1122 (2011)
17. V.N. Inukollu, S. Arsi, S.R. Ravuri, Security issues associated with big data in cloud computing. Int. J. Netw. Secur. App. **6**, 39–45 (2014)
18. A.S. Ibrahim, J. Hamlyn-Harri, J. Grundy, Emerging security challenges of cloud virtual infrastructure (2016)
19. H. Zhengbing, G. Sergiy, K. Oksana, G. Viktor, B. Serhii, Anomaly detection system in secure cloud computing environment. Int. J. Comput. Netw. Inf. Secur. **4**, 10–21 (2017)
20. A. Mahendiran, N. Saravanan, S.N. Venkata, N. Sairam, Implementation of K-means clustering in cloud computing environment. Res. J. App. Sci. Eng. Tech. **4**(10), 1391–1394 (2012)
21. S. Ahmad, A. Lavin, S. Purdy, Z. Agha, Unsupervised real-time anomaly detection for streaming data. Neurocomputing **262**, 134–147 (2017)
22. T.Y. Win, H. Tianfield, Q. Mair, Big data based security analytics for protecting virtualized infrastructures in cloud computing. IEEE Trans. Big Data **4**(1), 11–25 (2018)