BALTIC JOURNAL OF LAW & POLITICS
A Journal of Vytautas Magnus University
VOLUME 15, NUMBER 4 (2022)
ISSN 2029-0454

Cite: *Baltic Journal of Law & Politics* 15:4 (2022): 458-465
DOI:10.2478/bjlp-2022-004049

Detection of Malware in Cloud Storage Data using Naive Bayes Algorithm Comparing K-Nearest Neighbors Algorithm to Reduce False Detection

Borra Madhan Mohan Reddy
Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India: 602105.

P.Sriramya
Project Guide, Corresponding Author, Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India: 602105.

Received: August 8, 2022; reviews: 2; accepted: November 29, 2022.

Abstract

Aim: To enhance the accuracy in detection of Novel Cloud malware in cloud storage data Using K-Nearest Neighbors Algorithm comparing Naive Bayes Algorithm to reduce false detection. **Materials and Methods:** This research work we are considering two groups,one group is K-Nearest Neighbors Algorithm (KNN) comparing group 2 Naive Bayes Algorithm (NB). Each group consists of a sample size of 30. Their accuracies are compared with each other using different sample sizes also. **Results:** By running algorithms for various iterations the following results are obtained. SPSS was used to calculate the sample size. The pre-test analysis was maintained at 80%. G-power is used to calculate sample size. K-Nearest Neighbors Algorithm is 99.4% more accurate than the Naive Bayes Algorithm of 62.8% in detection of malware in cloud storage data which reduces the false detection rate (p=0.001). **Conclusion:** Through this, we are able to prove that the prediction novel cloud Malware Analysis done using K-Nearest Neighbors (KNN) model is significantly better than the Naive Bayes in identifying Malware detection in cloud storage data. It can be also considered as a better option for the classification of malware detection.

Keywords

Malware Detection, Cloud Storage, Novel Cloud Malware Analysis, Machine Learning, Naive Bayes Algorithm, K-Nearest Neighbors Algorithm.

INTRODUCTION

Antivirus software programs are one of the most broadly used popular tools for detection, preventing malicious and undesirable scripts. However, the future impact of common host primarily based antivirus programs is questionable (El-Khouly and El-Seoud 2017). Antivirus software programs neglect to see various contemporary threats and their growing intricacy has ended in vulnerabilities which might be being taken advantage of by malware scripts (Yadav 2019). This document supports a replacement approach for host

malware recognition based on the implementation of antivirus as a network service in the cloud (Supriya et al. 2020). This mannequin permits the detection of malicious and undesirable by software using more than one detection engine severally (Win, Tianfield, and Mair 2015; "Malware Detection in Cloud Computing Infrastructures" 2018). We additionally argue the advantages of a couple of detection for the duration of the cloud and give a brand latest technique to work detection throughout the cloud (Watson et al. 2016).

Most cited articles, The websites visited reference are IEEE and Google Scholar. IEEE has 90 citations and Google scholar has about 170 citations. "Malware detection in cloud computing infrastructures" (Win, Tianfield, and Mair 2015) has been cited by 161, "Analyzing CNN based behavioral malware detection techniques on Cloud Iaas" (Christodorescu et al. 2007) was Cited by 15, "Effective analysis of malware detection in cloud computing" was Cited by 22. This paper consolidates detection methodologies, static signature analysis and dynamic evaluation detection. Utilizing this component, We discover that Novel Cloud Malware discovery presents 35% higher discovery inclusion against the latest threats using this method compared with an individual antivirus machine then a 98% discovery dimension throughout the cloud environment (Salam, Maged, and Mahmoud 2014). Malware safety of pc structures is a totally crucial assignment in Cyber-Security (McDole et al. 2020). Even one unmarried assault is enough to lose our data (Nancy et al. 2016).

Previously our team has a rich experience in working on various research projects across multiple disciplines (Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020). The research gap identified from the literature survey is that classification models adopting Naive Bayes are not appropriate for handling massive datasets. It doesn't operate properly when the dataset has extra target instructions with greater noise and overlapping. In these cases the volume of highlights for every data factor surpasses the volume of making ready information tests, The accuracy of Naive Bayes will fall short of expectations. The study's goal is to implement novel malware detection and improve the classification accuracy by incorporating Naive Bayes Algorithm comparing K-Nearest Neighbors Algorithm to reduce false detection (Hegedus et al. 2011).

MATERIALS AND METHODS

The research work was performed in the Data analytics Lab in the Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences. The sample size taken for conducting the experiment was 10. Two groups are considered as classifiers algorithms in order to classify prediction of fare amount, machine learning classification algorithms are used. The work was carried out on 100000 records from a data-master dataset (Joslin 2010). The accuracy in classifying the blood cells was performed by evaluating two groups. A total of 10 iterations were performed on each group to achieve better accuracy. The study uses a dataset-master image dataset downloaded from kaggle.

Naive Bayes (NB) Algorithm:

Naive Bayes is a probabilistic ML algorithm that can be utilized in a wide assortment of grouping tasks. The name naive is utilized on the grounds that it accepts the provisions that go into the model are free of one another. Equation (1) gives the numerically given the Bayesian calculation is addressing a class variable and the arrangement of qualities, Conditional probability of A given B can be registered as:

$$P(A \mid B) = P(A \cap B) / P(B) \quad (1)$$

K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors algorithm is a simple, supervised machine learning algorithm that can be used to resolve both regression and classification problems. It's easy to set up and operate, but has an important downside of becoming significantly slower as the quantity of that information in use increases.

Numerical Example of K-Nearest Neighbor Algorithm

1. Here, the Step-By-Step instruction on a way to cypher the KNN algorithm.
2. Based on Parameters K = Number of Nearest Neighbors.
3. Compute the difference distance between the query-instance and every one of the preparation tests.
4. Sort the distance and get nearest neighbors predicting on the K-th minimal distance.
5. Y is the category for the Nearest Neighbors.
6. Use the easy majority of the order of Nearest Neighbors as the query-prediction instance's value.

Statistical Analysis

Statistical Package for the Social Sciences Version 23 software tool for statistical analysis, a software application was used. For accuracy, an independent sample T-test was used. The SPSS Software programme was also used to calculate standard deviation and standard mean errors. Group statistics and independent sample t-tests were performed on the experimental results and the graph was built for two groups with two parameters under study.

RESULTS

The proposed algorithm Naive Bayes and existing algorithm K-Nearest Neighbors (KNN) algorithm were run at a time in an Anaconda-Jupyter. Fig. 1 shows an Architecture diagram Table 1 has the sample sets executed for a number of iterations, the accuracy values of the accuracy grouping of Naive Bayes and K-Nearest Neighbors Algorithm classifiers differs.

Analysis of the overall classification of Detection of Novel Cloud Malware in Cloud storage Data by Naive Bayes and K-Nearest Neighbors Algorithm models shows the classification of the detecting malware. K-Nearest Neighbors (99.4%) shows better accuracy than Naive Bayes (62.7%). Statistical Analysis of Standard Error, Standard deviation, Mean and Accuracy of Naive Bayes and K-Nearest Neighbors Algorithm is done. The group statistics for the t-test is shown in Table 2 and an independent sample test for the given samples is shown in Table 3. There is a genuinely massive contrast in Accuracy values between the algorithms. K-Nearest Neighbors had obtained higher accuracy compared to Naive Bayes (NB) which is shown in Fig. 2.

Figure 2 the Bar chart suggests the evaluation of Accuracy and loss of Naive Bayes Algorithm and K-Nearest Neighbors. It’s easily observed that K-Nearest Neighbors gives more accuracy and low mistakes when varied with Naive Bayes as shown in Fig. 2.

DISCUSSION

The Naive Bayes and K-Nearest Neighbors (KNN) algorithm classifiers on a dataset acquired from diverse sources like Kaggle, Github, et al. are compared during this section. After completing preprocessing and extraction on the dataset, the dataset was separated into portions for training and testing. The accuracy is calculated using both K-Nearest Neighbors Algorithm and Naive Bayes. Surprisingly, the Naive Bayes outperformed the KNN in every way. The accuracy of a classifier is critical in determining the efficacy of Detection of Novel Cloud Malware in Cloud storage to reduce false detection.

Machine learning algorithms for cloud-based malware detection are being investigated and give the similar findings as discussed in this paper. (Thomas, Vijayaraghavan, and Emmanuel 2020; Kimmell, Abdelsalam, and Gupta 2021). This paper

gives the analysis of varied Models of machine learning that can be utilized as a starting point for further study that focuses just on one machine learning model. The opposite findings are specified in the K-NN Classification of Malware in Cloud Traffic Using the Metric Space Approach (Lokoč et al. 2016). Malware monitoring in cloud environments using k-NN classification is explored in this paper. The measurement space strategy for estimated k-NN look over a dataset of meager high-layered descriptors is the focus of this paper.

There are limitations with various cloud conditions in identifying malware and furthermore proposes a cloud-based malware identification structure, which utilizes a hybrid way to distinguish malware. Cloud malware analysis tools are developing new and advanced features, which will probably be able to resolve such uncertainties. These findings are being provided to an interface that will display and populate a Machine Learning (ML) algorithm that identifies and simplifies the principles underlying the data it encounters. Despite the actual fact that the presented methodology yielded good results, the approach's shortcoming is that it needs to be enhanced to reduce false detection of malware. This may be avoided in the future by combining Naive Bayes with other approaches.

CONCLUSION

The studies on prediction are completed using the device getting to machine learning algorithms. Naive Bayes algorithm compared with K-Nearest Neighbors Algorithm (KNN) are giving the accuracy of 62.8% and 99.4% respectively. The studies can be in addition prolonged with diverse datasets and diverse attributes for the ensemble of the device getting to know algorithms.

DECLARATION

Conflicts of Interest

The author declares no conflict of Interest.

Authors Contributions

Author BMMR was involved in data collection, data analysis, manuscript writing. Author PSR was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Qbec Infosol Pvt. Ltd., Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. "A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres." *Materials Today: Proceedings* 33 (January): 798–805.

Christodorescu, Mihai, Somesh Jha, Douglas Maughan, Dawn Song, and Cliff Wang. 2007. *Malware Detection*. Springer Science & Business Media.

El-Khouly, Mahmoud M., and Samir Abou El-Seoud. 2017. "Malware Detection in Cloud Environment (MDCE)." *International Journal of Interactive Mobile Technologies (iJIM)*. https://doi.org/10.3991/ijim.v11i2.6575.

Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. "Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition." *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.

Hegedus, Jozsef, Yoan Miche, Alexander Ilin, and Amaury Lendasse. 2011. "Methodology

for Behavioral-Based Malware Analysis and Detection Using Random Projections and K-Nearest Neighbors Classifiers." *2011 Seventh International Conference on Computational Intelligence and Security*. https://doi.org/10.1109/cis.2011.227.

Joslin, Ann. 2010. "Regional Fare Policy and Fare Allocation, Innovations in Fare Equipment and Data Collection." https://doi.org/10.5038/cutr-nctr-rr-2006-04.

Kimmell, Jeffrey C., Mahmoud Abdelsalam, and Maanak Gupta. 2021. "Analyzing Machine Learning Approaches for Online Malware Detection in Cloud." *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*. https://doi.org/10.1109/smartcomp52413.2021.00046.

Lokoč, Jakub, Jan Kohout, Přemysl Čech, Tomáš Skopal, and Tomáš Pevný. 2016. "K-NN Classification of Malware in HTTPS Traffic Using the Metric Space Approach." *Intelligence and Security Informatics*. https://doi.org/10.1007/978-3-319-31863-9_10.

"Malware Detection in Cloud Computing Infrastructures." 2018. *International Journal of Recent Trends in Engineering and Research*. https://doi.org/10.23883/ijrter.conf.20171201.044.wsqfb.

McDole, Andrew, Mahmoud Abdelsalam, Maanak Gupta, and Sudip Mittal. 2020. "Analyzing CNN Based Behavioural Malware Detection Techniques on Cloud IaaS." *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-030-59635-4_5.

Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. "Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia." *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.

Nancy, Nancy, Sanjay Silakari, Uday Chourasia, and Uit Rgpv. 2016. "A Survey Over the Various Malware Detection Techniques Used in Cloud Computing." *International Journal of Engineering Research and*. https://doi.org/10.17577/ijertv5is020388.

Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. "Dental Caries Profile and Associated Risk Factors among Adolescent School Children in an Urban South-Indian City." *Oral Health & Preventive Dentistry* 18 (1): 379–86.

Salam, Safaa, Maged, and Mahmoud. 2014. "Malware Detection in Cloud Computing." *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2014.050427.

Sathish, T., and S. Karthick. 2020. "Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite." *Materials Today: Proceedings* 33 (January): 2555–58.

Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. "Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid." In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.

Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. "Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies." *Materials Today Communications* 25 (December): 101603.

Supriya, Yamiala, Gautam Kumar, Dammu Sowjanya, Deepali Yadav, and Devarakonda Lakshmi Kameshwari. 2020. "Malware Detection Techniques: A Survey." *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. https://doi.org/10.1109/pdgc50313.2020.9315764.

Thomas, Tony, Athira P. Vijayaraghavan, and Sabu Emmanuel. 2020. "Support Vector Machines and Malware Detection." *Machine Learning Approaches in Cyber Security Analytics*. https://doi.org/10.1007/978-981-15-1706-8_4.

Venu, Harish, and Prabhu Appavu. 2021. "Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine." *International Journal of Ambient Energy* 42 (14): 1588–94.

Watson, Michael R., Noor-Ul-Hassan Shirazi, Angelos K. Marnerides, Andreas Mauthe, and David Hutchison. 2016. "Malware Detection in Cloud Computing Infrastructures." *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/tdsc.2015.2457918.

Win, Thu Yein, Huaglory Tianfield, and Quentin Mair. 2015. "Detection of Malware and Kernel-Level Rootkits in Cloud Computing Environments." *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*. https://doi.org/10.1109/cscloud.2015.54.

Yadav, Ram Mahesh. 2019. "Effective Analysis of Malware Detection in Cloud Computing." *Computers & Security*. https://doi.org/10.1016/j.cose.2018.12.005.

Tables and Figures

Table 1. Comparing Accuracy and Sensitivity achieved during the evaluation of Naive Bayes and K-Nearest Neighbors models for classification with different iterations.

No. of Iteration	Naive Bayes	K-Nearest Neighbors
1	62%	98%
2	64%	96%
3	60%	99%
4	63%	97%
5	61%	94%

Table 2. Standard Error, Standard Deviation, Mean, and Accuracy of Naive Bayes and K-Nearest Neighbors Statistical Analysis In the algorithms, there is a statistically significant difference in accuracy values. K-Nearest Neighbors had the highest Accuracy (99%) and Sensitivity (62%) compared with Naive Bayes. The Standard error is also less in Naive Bayes in comparison to K-Nearest Neighbors.

Accuracy Group	N	Mean	Std. Deviation	Std.Error Mean
Naive Bayes	5	62.5420	.44757	.20016
KNN	5	97.7240	.63787	.28526

Table 3. Comparing the significance level for Naive Bayes and K-Nearest Neighbors algorithms with value p = 0.001. Both Naive Bayes and K-Nearest Neighbors have a significance level less than 0.05 in terms of accuracy with a 95% confidence interval.

	Levene's Test for Equality of Variances		T-test for Equality of means						
Accuracy	F	Sig.	t	df	Sig(2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
								Lower	Upper
Equal variances assume	.093	.768	-100.958	8	.001	-35.18200	.34848	-35.98560	-34.37840

d Equal variances not assumed			-100.958	7.17	.001	-35.6000	.34848	-36.00208	-34.36192

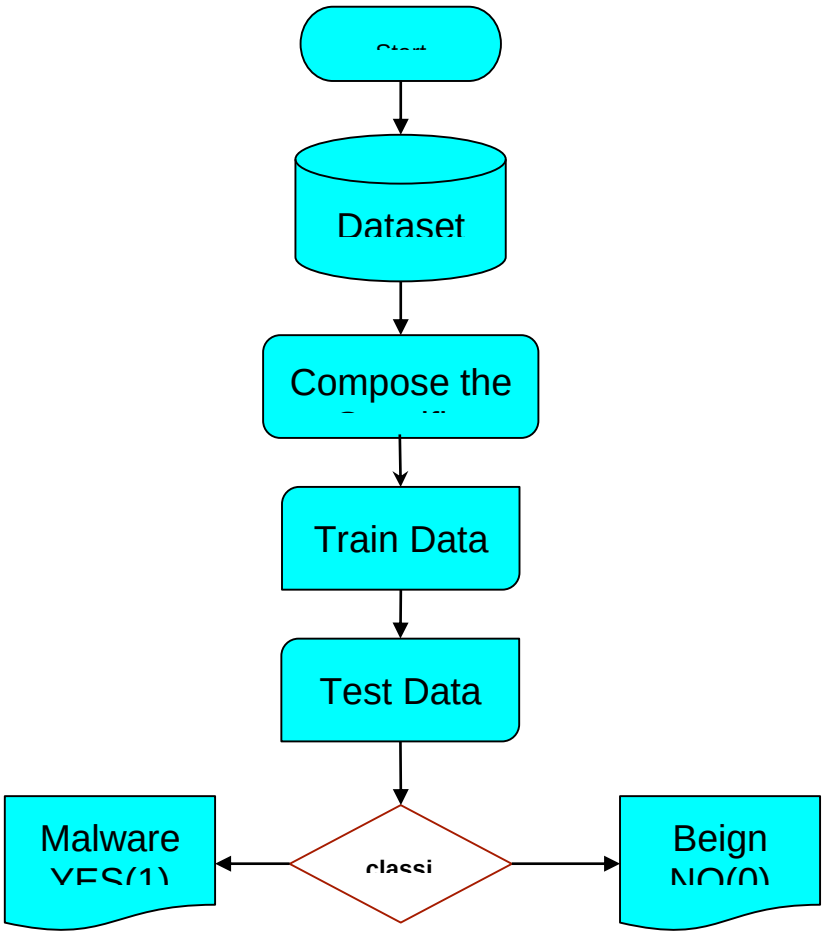

Fig. 1: Architecture Diagram Malware Analysis

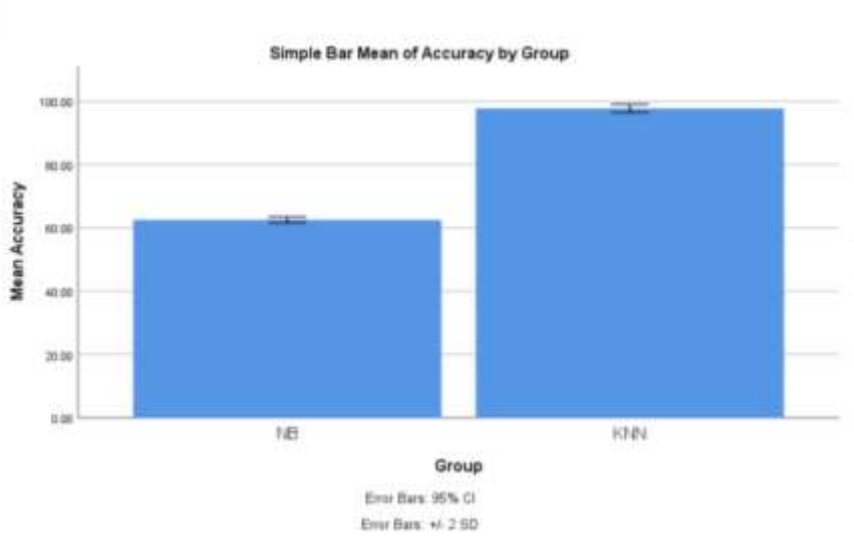

Fig. 2: Comparison of mean accuracy of Naive Bayes and K-Nearest Neighbors algorithm. The standard errors appear to be less in K-Nearest Neighbors (KNN) compared to Naive Bayes. K-Nearest Neighbors appear to produce more consistent results with higher sensitivity. X-Axis: K-Nearest Neighbors vs Naive Bayes algorithm. Y-Axis: Mean sensitivity of detection +/- 2 SD, Error Bars 95% CI.