Advances in Manufacturing Technology XXXV
M. Shafik and K. Case (Eds.)

doi:10.3233/ATDE220608

Anomaly Detection System for Ethereum Blockchain Using Machine Learning

Njoku ThankGod ANTHONY, Mahmoud SHAFIK, Fatih KURUGOLLU, Hany F. ATLAM

College of Science and Engineering, University of Derby, Derby DE22 1GB, UK

Abstract. Over the past few years, Blockchain technology has been utilized in various applications to improve privacy and security. Although blockchain has proven its worth as a very powerful technology, research has shown that it is not entirely immune to security and privacy attacks. There was a successful 51% attack on Ethereum Classic back in January 2019 which shows that blockchain still facing security and privacy challenges. This paper aims to develop an anomaly detection solution for the Ethereum blockchain to overcome security challenges using Machine Learning (ML). The proposed solution focuses on using a dynamic approach where the normal operational behaviour of the Ethereum blockchain is used to train ML algorithms and any deviation will be tagged as an anomaly and will be detected by the system. Four ML algorithms including K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GaussianNB), Random Forest, and Stochastic Gradient Descent (SDG) were utilized to train and verify the accuracy of the proposed solution. The experimental results demonstrated that the random forest algorithm provided the best accuracy of 99.84% over other ML algorithms.

Keywords. Blockchain technology, Machine Learning, Deep Learning, Anomaly Detection

1. Introduction

Blockchain, as a distributed public ledger, has been attracting lots of attention from different industries and researchers. Blockchain was first introduced by the release of a white paper in 2008 by Nakamoto [1]. Blockchain is a distributed ledger that records, share and synchronize transactions in their respective electronic ledgers. The decentralised nature of blockchain operations, which means the transactions can take place without intermediaries, makes it applicable for digital assets, online payments, remittances and other financial services [2; 3]. Blockchain technology, however, can be prone to security threats and attacks as have been evidenced by some attacks recorded previously on blockchain-empowered currency [4]. A permissionless blockchain platform for smart contracts known as Ethereum, for instance, has suffered two different attacks which were bad enough to affect the functions of the blockchain network [5].

Although blockchain has proven its worth as a very powerful technology, it is not entirely immune to data security challenges or cyber-attacks. These attacks can be seen as a deviation from normal behaviour, and this is what can be described as anomalous behaviour. Anomalies in networks or systems are usually led to fraudulent transactions or activities. For the safety of blockchain networks, these anomalies need to be detected

and removed immediately to ensure the safety of the blockchain and as well maintain the trustworthiness that the blockchain is known for.

This paper aims to propose a security solution to improve security in the Ethereum blockchain by providing an anomaly detection solution using Machine Learning (ML). Four ML algorithms were utilized and implemented including K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GaussianNB), Random Forest, and Stochastic Gradient Descent (SDG). The main purpose of using ML algorithms is to build an automated prediction system where any sign of anomalies can be predicted using the change in behaviour or pattern from the know digital signature or behavioural pattern. This can improve the security of the Ethereum blockchain by detecting and preventing anomalies effectively and promptly.

The rest of this paper is organized as follows: Section 2 summarizes related work on anomaly detection in the blockchain. Section 3 shows security attacks in blockchain, Section 4 presents the proposed framework. Section 5 shows the implementation and results, and Section 6 is the conclusion.

2. Related Work

ML algorithms have been successfully utilized for anomaly detection in the literature, and this has mostly been based on known attacks by building attack logs. Chen et al. [6] proposed an anomaly detection solution from a network behavioural pattern. The authors claimed that this solution is of a generalised nature and can be applied to every blockchain platform. However, the data used in this study was not publicly disclosed. The proposed solution finds the behavioural pattern in the blockchain network and uses it to categorise them. The authors firstly utilized Dynamic Time Warping (DTW), Longest Common Sub-Sequence (LCSS) [7], Euclidean distance [8], and Edit Distance on Real sequence (EDR) for training and testing the proposed solution. Based on the results, the authors recommended DTW as the most efficient algorithm since LCSS and EDR are focused more on handling noise data but blockchain data are noise-less.

Heilman et al. [9] introduced what the author called the blockchain anomaly detection (BAD) system. Their system exploits blockchain metadata. This is used to collect information regarding any malicious activity in the blockchain. The focus was on eliminating eclipse attacks. The main idea was to collect the data of any malicious activity and then use it to create a model for tackling further attacks. The authors proposed building a database of threats that have happened over time and used it as a baseline model.

Monamo et al. [10] proposed a solution for anomaly detection in the blockchain using a dataset provided by a computational biology laboratory from the University of Illinois. The dataset however is still not publicly available. The work was based on using K-means clustering. The authors used a multivariant set-up to detect suspicious users and any form of activity coming from them. The approach was able to detect 5 fraudulent users from the 30 user cases available.

Kumari and Catherine [11] employed the K-Means algorithm to monitor and detect malicious activities. Their focus was on the behaviour of the nodes. The clustering was performed by separating nodes with similar traits, they are then grouped to identify malicious nodes. The use of DTWT (Dynamic Time Wrapping Technique) helps them to calculate the time taken for each transaction, the similarity index, the quantity of the time from one node to another and other behavioural parameters.

3. Security Attacks in Blockchain

Blockchain is a new technology that has been adding more security value to different applications. It is difficult for attackers to manipulate blockchain due to the consensus algorithms and the decentralized nature of the system. Encryption in blockchain also plays a vital role in blockchain and the communication channel in the network. In blockchain systems, cryptographic keys, (public-private keys) are used to communicate. This helps to ensure the privacy and security of the nodes in the blockchain network.

Although blockchain provides countless benefits in various applications, it has some challenges. It has various challenges ranging from scalability, mining, identification, and security challenges. Although blockchain has proven its worth as a very powerful technology, it is not entirely immune to data security challenges or cyber-attacks. There was a successful 51% attack on Ethereum Classic back in January 2019, which shows the security vulnerabilities in one of the popular blockchain platforms. There are several security challenges for blockchain, as shown in Figure 1. Therefore, there is a need to provide a security solution to overcome security issues in blockchain technology.

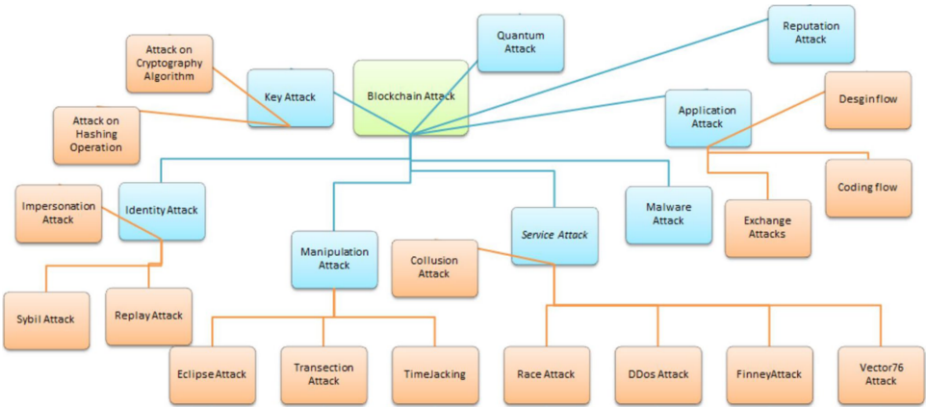

Figure 1. Blockchain security challenges [12]

4. Proposed Framework

This paper aims to solve security challenges in the Ethereum blockchain by developing a framework for anomaly detection using ML algorithms for smart government applications. The proposed framework can be shown in Figure 2. Smart government applications collect their data from different sources like sensors, smart devices, and other data sources. The data collected from these sources are then processed as part of the smart application. As shown in Figure 2, blockchain works as an integral part of these applications. Then, ML algorithms analyse data and provide real-time data analytics. Then, ML algorithms detect the anomalies by comparing the system behaviour against the normal behaviour that has been implemented.

To ensure immutability and decentralisation, the data in the ML models will be stored in the blockchain platform and this will help reduce other simple challenges like missing data values, and data duplication and then the major focus will be achieved. Four ML algorithms including KNN, Gaussian, Random Forest, and SDG were used to build an automated prediction system where any sign of anomalies can be predicted using the

change in behaviour from the know behavioural. This can allow the detection and prevention of anomalies effectively and promptly.

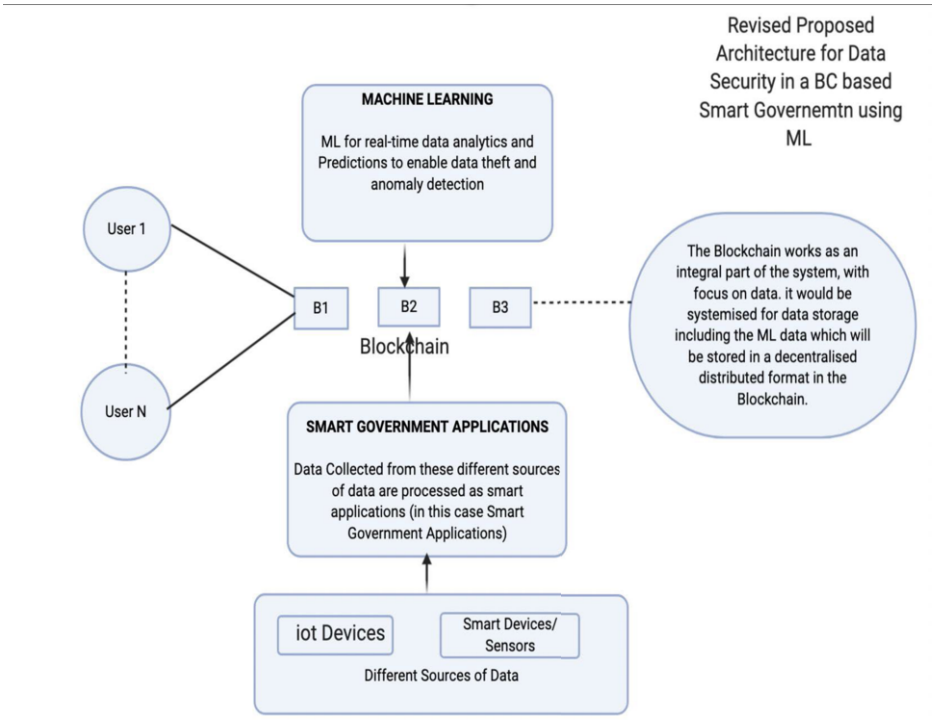

Figure 2. Proposed framework

5. Implementation and Experimental Results

Four ML algorithms were utilized to train and test the proposed framework to verify its accuracy and effectiveness. Python was utilized to implement the experiments. A dataset consisting of 6489 records was obtained from Blockchair [13], a private database dump for different permissionless blockchains. The dataset was split into 70% for training, and 30% for testing. This dataset was used as a set of normal behaviour for the anomaly detection system which the system will be used to compare against to detect anomalies. The accuracy of the proposed framework was computed using the detection accuracy (ACC), True Positive Rate (TPR) and False-Positive Rate (FPR). The equations of ACC, TPR and FPR are as follows:

$$ACC = \frac{TP + TN}{N}$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

where TP denotes True Positive, TN denotes True Negative, FN donates False Negative, and FP donates False Positive.

Analysis of the algorithms was performed to extract the features using the behaviour of the transactions. The addresses in the dataset were labelled where '0' was used for non-malicious addresses, while '1' was used for malicious addresses. For the classification, ML algorithms were used using the python socket-learn library. Table 1 shows the results of applying ML algorithms.

Table 1. Results of four ML algorithms

Algorithm	TPR	FPR	Accuracy
KNN	1.0	0.001636661	0.9799
GaussianNB	1.0	0.00216216	0.9490
Random Forest	1.0	0.00109	0.9984
SGD	0.973684	0.00109	0.9938

Table 1 compares the accuracy of four ML algorithms that were utilized to train and test the proposed anomaly detection solution. KNN algorithm is better suited for classification problems with having a time series dataset. As the data is not separable, the GaussianNB accuracy is low compared to the KNN algorithm. Random forest algorithm performs better than all other three algorithms. It can efficiently perform a non-linear classification using implicitly mapping their inputs into high-dimensional feature spaces. It is one of the most robust and accurate algorithms among the other classification algorithms. Random Forest requires relatively large data for training than the other algorithms. As we have utilized a relatively large dataset, Random Forest has performed very well. The experimental results indicated that the Random Forest algorithm has achieved the highest accuracy of 99.84%, while the GaussianNB algorithm has achieved the lowest accuracy of 94.9%.

6. Conclusion

Blockchain has established itself as a very strong technology, however, it is not totally immune to data security issues or cyberattacks. These attacks are a deviation from typical blockchain behaviour. The paper aimed to develop a security solution to improve security issues in the Ethereum blockchain platform. To achieve this, we proposed a security solution that can detect anomalies in the Ethereum blockchain. The proposed solution builds an automated prediction system where any sign of anomalies can be predicted using the change in behaviour or pattern from the know digital signature or behavioural pattern. This can improve the security of the Ethereum blockchain by detecting and preventing anomalies effectively and promptly. Four ML algorithms including KNN, GaussianNB, Random Forest, and SDG were utilized to train and test the accuracy and effectiveness of the proposed anomaly detection solution. The experimental results demonstrated that the highest accuracy was achieved by Random Forest (99.84%) while the lowest accuracy was achieved by GaussianNB (94.90%). All ML algorithms used have achieved an accuracy of over 90%.

References

[1] Nakamoto, S., 2008. [online] Bitcoin.org. Available at https://bitcoin.org/bitcoin.pdf [Accessed 18 June 2022].
[2] Namecoin.org. 2014. Namecoin. [online] Available at https://www.namecoin.org [Accessed 20 June 2022].
[3] Peters, G., Panayi, E. and Chapelle, A. "Trends in Crypto-Currencies and Blockchain Technologies: A Monetary Theory and Regulation Perspective". Journal of Financial Perspectives, Vol. 3, No. 3, 2015.
[4] A. Bogner. "Seeing is understanding: Anomaly detection in blockchains with visualized features". In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp '17, pages 5–8. ACM, 2017.
[5] Chen, Xuhui & Ji, Jinlong & Luo, Changqing & Liao, Weixian & Li, Pan. "When Machine Learning Meets Blockchain: A Decentralized, Privacy-preserving and Secure Design". In the 2018 IEEE International Conference on Big Data (Big Data), 2018.
[6] Huang, Butian, Z., Liu, J., Chen, A., Liu. Q., Liu, and Q. He. "Behavior pattern clustering in blockchain networks". Multimedia Tools and Applications. 2018.
[7] Vlachos, Michail, George Kollios, and Dimitrios Gunopulos. "Discovering similar multidimensional trajectories". In the Proceedings of the 18th IEEE International Conference on Data Engineering, pp. 673 - 684, 2002.
[8] Morse, Michael D and Jignesh M Patel. "An efficient and accurate method for evaluating time series similarity". In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. ACM, pp. 569-580, 2007.
[9] Heilman, Ethan et al. "Eclipse Attacks on Bitcoin's Peer-to-Peer Network". In: USENIX Security Symposium, pp. 129-144, 2015.
[10] Monamo, Patrick, Vukosi Marivate, and Bheki Twala. "Unsupervised learning for robust Bitcoin fraud detection" In Information Security for South Africa (ISSA), pp. 129-134, 2016.
[11] R. Kumari, and M. Catherine, "Anomaly Detection in Blockchain Using Clustering Protocol", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), vol. 4, no. 12, Dec. 2017.
[12] Iansiti, Marco and Karim R. Lakhani. The Truth About Blockchain. English, 2017.
[13] Blockchair.com. 2022. Ethereum Explorer. [online] Available at https://blockchair.com/ethereum [Accessed 1 May 2022].