Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology

Anokye Acheampong Amponsah [*], Adebayo Felix Adekoya, Benjamin Asubam Weyori

Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, P.O. Box 214, Ghana

ARTICLE INFO

Keywords:
Blockchain technology
Smart contract
Machine learning
Fraud detection
Healthcare fraud

ABSTRACT

Healthcare fraud is a global problem affecting both developing and developed countries. It is the deliberate attempt of the perpetrators to take undue advantage of the inefficiencies in current healthcare systems. Fraud tends to deny legitimate beneficiaries of universal health coverage, especially those under health insurance protection. In this work, we propose using machine learning techniques and blockchain technology to detect and prevent fraud in healthcare, especially in claims processing. A decision tree classification algorithm is adopted to classify the original claims dataset. The extracted knowledge is programmed in the Ethereum blockchain smart contract to detect and prevent healthcare fraud. The comparative experimental results show that the best performing tool achieves a classification accuracy of 97.96% and a sensitivity of 98.09%. This means that the proposed system enhances the blockchain smart contract's ability to detect fraud with an accuracy of 97.96%.

1. Introduction

Blockchain technology (BC) was introduced in 2008 by an unknown sole researcher or group of researchers called Satoshi Nakamoto. It was introduced to provide a permanent solution to the double-spending problem. Blockchain technology was initially tremendously leveraged in the financial sector and since its success, it has been extended to other domains like education, health, engineering, Internet of Things, manufacturing, insurance, energy, agriculture, social media, entertainment, and many other domains. This extension was idealized by Swan [1]. In these domains, cryptocurrency may not be used and in such a case, blockchain technology is leveraged for security, efficiency, privacy, immutability, accountability, ownership, easy auditability, etc. [2]. For example, in the health insurance domain, BC technology can be used to immutably record healthcare services provided and the legitimate claims that have been paid, and other claims-related data [3–6]. In such an example, BC enforces efficient data security and management and not for the transfer and management of cryptocurrency.

Healthcare fraud is an international issue that affects both developed and developing nations, and it is very detrimental to the beneficiaries of quality healthcare services especially those under health insurance [3,7]. All over the world, enormous amounts of money are lost to fraud [8], and this increases the overhead operation costs of the insurance companies and the premiums paid by the insureds [2]. According to the 2015 report on the financial cost of healthcare fraud [8], a total of about £303.8 million was lost to healthcare fraud. This amount can be broken down into £237 million for prescription charge fraud, £43.9 million for dental charge fraud, and £22.9 million for optical charge fraud. These amounts exclude the losses incurred by expenditure and income which will inflate the figure to a range of £3.73 – 5.74 billion. Similarly, according to Thaifur et al. [9], healthcare fraud costs Europe and Korea respectively an estimated €56 billion and 798.2 billion each year. In Africa, the health insurance schemes propounded by Governments help the vulnerable by protecting them from making out-of-pocket payments during healthcare delivery and the purchase of pharmaceuticals. Adequate evidence exists in the literature about the existence of fraud in the health insurance schemes in Africa. For example, according to GenKey SOLUTIONS, B.V. [10], the clinical audit report and claim intelligence data gathered indicate that about 15% to 20% of healthcare expenditure is lost via fraud. This amounts to about $487 billion being lost annually to fraud. In South Africa, fraud adds between R192 ($14) and R410 ($30) per month to every national health insurance member's medical aid contributions. These small overheads in healthcare costs are estimated to be a total of about $882 Million. Regarding this problem, many propositions have been made by several researchers to solve this issue. A classical example is Amponsah et al. [3]. The authors proposed an efficient data management system using blockchain technology and a claims processing system intended to ensure swift payment of service providers. Our work extends the contribution of Amponsah et al. [3] to include a module

[*] Corresponding author.
E-mail addresses: amponsah.acheampong@uenr.edu.gh (A.A. Amponsah), adebayo.adekoyaa@uenr.edu.gh (A.F. Adekoya), benjaminn.weyori@uenr.edu.gh (B.A. Weyori).

https://doi.org/10.1016/j.dajour.2022.100122
Received 4 July 2022; Received in revised form 21 August 2022; Accepted 30 August 2022
Available online 5 September 2022

that predicts and detects fraud in the claims processing system using machine learning (ML).

In Ghana, according to Wang et al. [11] and Amponsah et al. [3], the National Health Insurance Scheme is faced with tremendous financial sustainability threats. These threats are among others composed of fraud and corrupt practices found within the claims processing lifecycle. This has made it practically difficult for the citizens and residents of the country to benefit from Universal Health Coverage (UHC) as defined by the World Health Organization and the United Nations Sustainable Development Goal 3 [12]. Consequently, this work proposes a fraud detection system in a blockchain claims processing system using machine learning knowledge implemented in the smart contract. The paramount benefit of this work is to prevent malicious and disgruntled entities found within the NHIS claims processing lifecycle from executing their fraudulent activities. It will also protect the insurance scheme against financial loss that results from the manipulation of the inefficiencies found within the claims processing lifecycle.

2. Statement of the problem

Referencing the concept of Blockchain 3.0 [1], limited works have been done to test the applicability of Blockchain Technology in the insurance sector. Consequently, this work answers the question "How can Blockchain Technology be leveraged in contemporary insurance sub-business processes?". It contributes to the efforts to expand Blockchain 3.0 to include the insurance sectors and thereby limiting the scope of the work to fraud detection in claims processing. Again, in the same efforts, Amponsah et al. [3] propose a cloud-based blockchain system that efficiently manages claims data and swiftly processes claims, and reimburses healthcare service providers. The system ensures transparent dealings among the major stakeholders however, we are certain that their system can be enhanced when the smart contract is equipped with data-driven decision-making capability. Consequently, this work fills the crucial gap by applying the extracted Decision Tree classification rules in the blockchain smart contract to equip it with decision-making capability. The capability allows the smart contract to detect fraud in health insurance claims using real data. As identified in Section 1, the National Health Insurance Schemes in Africa are entangled with fraud resulting in the loss of funds that could have been used to ensure Universal Health Coverage. . Therefore there is the need for efficient solution to prevent healthcare fraud. The prototype of this work detects and prevents fraud by using complex and sophisticated Decision Tree classification algorithms and health insurance domain-specific data to originally equip the smart contract of the blockchain technology.

The Contribution of the work

Our contribution hinges on the proposition of a novel fraud detection and prevention system that uses machine learning algorithm and blockchain technology. The machine learning algorithm enables domain-specific data to be converted into knowledge and our proposed adaptable framework allows the collection and usage of new data as they become available. The outcome of this work expands the coverage of Blockchain 3.0 to include the insurance sector as conceived by Swan [1]. The Blockchain technology provides enhanced security, privacy, authentication, integrity, and non-repudiation. We apply the extracted machine learning decision rules and implement the framework in the health insurance claim processing system. The health insurance case was studied because healthcare fraud is an international problem that results in the loss of millions of dollars worldwide.

This work is organized as follows. Section 2 discusses the related works and Section 3 presents the concepts and preliminary definitions of blockchain technology and machine learning. Section 4 describes the design of the experiments and the methods employed to successfully achieve the set objectives. Section 4 presents the results from the machine learning experiments and how it was implemented in the development of the Blockchain fraud detection system. Section 5 discusses the results and Section 6 is the conclusion.

3. Related works

This work is related to Momeni et al. [13], Eshghie et al. [14], Bandara et al. [15], Wang et al. [16], Tann et al. [17], and Ashizawa et al. [18]. These authors have demonstrated in numerous ways how to integrate machine learning and Ethereum smart contracts from the perspective of security. Many of these authors used machine learning techniques to identify, monitor, and detect the vulnerabilities found in Ethereum and other platform smart contracts. Table 1 summarizes the related works from a security perspective. In similar works, Merrad et al. [19] created a P2P blockchain system to enable the production and trading of energy among consumers without any centralized authority. The authors used a deep learning model to forecast and predict future energy consumption based on past data, and the K-Means clustering algorithm to make the time of energy use independent and incontestable. Again, Hu et al. [20] have designed an efficient Ethereum traffic identification system with high identification accuracy for an Internet Service Provider to supervise its internal users on the Ethereum platform without Deep Packet Inspection (DPI). The authors used Support Vector Machine, K-Nearest Neighbour, Random Forest, and Logistic regression. Tsoukas et al. [21], have also proposed a blockchain-based system that ensures data integrity, automation of workflows, and interoperability between monitoring systems of different vendors. From Table 2, it can be seen that the majority of the works are done for the sectors of energy, health, food supply chain, and Unmanned Aerial Vehicles, and works in Table 1 focus on security.

In this work, we focus on leveraging machine learning to equip the blockchain smart contract in detecting and preventing fraud in health insurance claims processing. We use Decision Tree algorithms to classify original NHIS claims related data to extract the patterns of fraud. The extracted knowledge is then implemented in Ethereum smart contract. Machine learning methods are proposed because they can turn domain-specific data into knowledge and then provide efficient classification and predictive models that can be used in the future. The principal reason for using the decision tree algorithm is its ability to perform white-box computation and the generation of specific domain data-driven classification rules.

4. Preliminaries

4.1. Blockchain technology

Blockchain is a distributed, publicly verifiable, transparent, immutable, authentic, and decentralized solution for recording the history of transactions and for data management [2,25,26]. Although blockchain is novel, it relies heavily on existing technologies such as cryptography algorithms, consensus mechanisms, distributed ledger, Merkle tree, and certificates to provide immutability, confidentiality, privacy, and security [26]. Blockchain (BC) emerged in 2008 as a financial application [27] pioneered by Bitcoin and Ethereum. BC increased in recognition because the finance sector benefited tremendously and saw a fundamental change in the business processes [28]. BC is used in this work to provide enhanced security, transparency, easy auditability, privacy, and data integrity.

4.2. Types of blockchain technology

There are two main types of blockchain technology — permissionless and permissioned. According to Helliar et al. [29], Permissionless Blockchains, are also referred to as trustless or public blockchains and are open networks where anybody can participate in the consensus process used by the blockchain to confirm transactions and data. They are completely decentralized and distributed among unknown parties. Permissionless blockchain systems are fully transparent of transactions with open source development, participants are completely anonymous, with some exceptions, and lack a central regulatory body. Public

Table 1
Comparison of security propositions regarding smart contracts.

Author(s)	Objective	Perspective(s)	Methods	Target platform	Contribution
Momeni et al. [13]	The proposition of a Machine Learning predictive model to detect patterns of security vulnerabilities in smart contracts.	Security	Support vector machine, Neural network, decision trees, random forest Dataset: 1000 plus operational ethereum smart contracts with known vulnerabilities	Ethereum	A fast approach to analyse the code of smart contract for anomalies.
Eshghie et al. [14]	The proposition of Dynamit — a smart contract monitoring framework to detect reentrancy vulnerability in Ethereum smart contracts.	Security	Random forest, naive bayes, logistic regression, K-Nearest neighbour, and support vector machine. Dataset: Transaction metadata and balance data from the blockchain system.	Ethereum	The proposed system requires no domain knowledge, code instrumentation, or special execution environment.
Bandara et al. [15]	The invention of a new scalable smart contract platform (Aplos)	Security and Performance	Mystiko Dataset: NA	Mystiko	Performance and scalability in the smart contract platforms. The proposed system supports concurrent transactions, high throughput, data analytics, and machine learning.
Wang et al. [16]	Development of contraWard to detect vulnerabilities in smart contracts using machine learning.	Security	Xtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Random forest, support vector machine, and K-Nearest Neighbour. Dataset: 49 502 smart contract dataset	Ethereum but applicable to other platforms like Solidity, Serpent, LLL, etc.	The proposition of ensemble supervised learning models that can be used to detect vulnerabilities in smart contracts.
Tann et al. [17]	Use LSTM to quickly and safely detect vulnerabilities in smart contracts.	Security	Long Short Term Memory Dataset: 620 000 Ethereum dataset	Ethereum	A quick and safer approach to detect vulnerabilities in smart contracts using machine learning
Ashizawa et al. [18]	The proposition of Eth2Vec, a machine-learning-based static analysis tool for vulnerability detection in smart contracts.	Security	Eth2Vec. Dataset: 5000 contract files from Etherscan, which is an open database of Ethereum smart contracts,	Ethereum	The proposition of Eth2Vec – a static analysis method to detect several anomalies in the codes of smart contracts.
Zhang et al. [22]	Ensemble learning-based approach smart contract vulnerability prediction to predict vulnerabilities in Ethereum smart contracts	Security	21,667 smart contract dataset CNN, RNN, RCNN, DNN, GRU, Bi-GRU, and transformer	Ethereum	Accurate smart contract source code vulnerability prediction.
This work	To provide a health insurance fraud prevention system.	Non-security	J48 ADTree BFTree REPTree Dataset — 1323 health insurance claims record.	Platform independent	Provision of a robust blockchain-based claims fraud prevention system that leverages machine learning algorithm.

Table 2
Application of smart contracts and blockchain technology in other domains.

Author	Sector	Methods	Proposition
Merrad et al. [19]	Energy	Deep learning model, K-Means algorithm	Peer to Peer energy trading system
Hassija et al. [23]	Health	Naive Bayes, Logistic regression,	Blockchain-based framework for the recommendation of specialists for minor medical consultations
Hu et al. [20]	Networking	Support vector machine, K-Nearest neighbour, Random forest, and Logistic regression	Ethereum traffic identification system
Tsoukas et al. [21]	Supply chain (Food)	TinyML	A blockchain-based system that ensures data integrity, automation of workflows, and interoperability between monitoring systems of different vendors
Feng et al. [24]	Unmanned Aerial Vehicles (UAVs)	CNN Federated EMNIST dataset	A blockchain-empowered decentralized horizontal Federated Learning framework to authenticate cross-domain UAVs using smart contracts
This work	Insurance	Decision tree Claims processing dataset	The proposition of adaptive self-autonomous domain data-driven decision-making smart contract

blockchain relies heavily on incentivizing participants using tokens and other digital assets. Examples of Permissionless blockchain platforms are Bitcoin and Ethereum [30]. Permissioned blockchains also known as private blockchains or Permissioned sandboxes are closed networks wherein prior designated parties, usually members of a consortium, communicate and participate in the consensus and data validation. They are mostly halfway decentralized in the sense that, unlike permissionless blockchains, resources are shared among already known participants instead of anonymous users. Tokens and digital assets are possible, but not as widespread as they are in permissionless environments. Examples of private blockchains are Hyperledger Fabric, Corda, Multichain, etc. [30]. Lately, there are also cloud-based Blockchain As A Service platforms like Xooa, Microsoft, IBM, Amazon, R3, etc. These systems provide a low or no code Integrated Development Environment in the cloud to allow developers to have a friendly environment for swift development of blockchain solutions.

4.3. Smart contract

Nick Szabo proposed the term smart contract in a paper titled "Formalizing and Securing Relationships on Public Networks". Cryptographic methods, he explained, could make it possible to build computer software that resembles contractual agreements, limiting the ability to cancel its performance duties. Scholars have investigated computer-based contractual languages in the years since. According to Szabo [31], Smart Contracts (SC) are blockchain-based computer programs that are called by BC members. Any system that BCs support can benefit from SC automation and control flow logic. In every way, smart contracts should be viewed as software functions, and the BC engine for the smart contract should be deterministic. SCs' determinism referencing Christidis and Devetsikiotis [32] is the property that keeps the ledger stable and consistent, ensures transaction finality, and prevents soft and hard forks. The developer is usually in charge of the determinism of SC's activities. As a result, the programmer must guarantee that automated activities are carried out as planned and that the data is left in a consistent form, regardless of the node(s) on which they are carried out. Each time the SC is run, the actions must provide the same result. Smart contracts can be classified into three broad categories, according to the authors' empiricism: static, dynamic, and Oracle-driven.

4.4. Machine learning and healthcare fraud detection systems

Machine Learning (ML) is a subfield in Artificial Intelligence that presents computer-assisted tools to turn data/information into knowledge. Machine learning is a type of data analysis that automates the creation of analytical models. It is a subset of artificial intelligence predicated on the premise that computers can learn from data, recognize patterns, and make judgments with little or no human intervention. The two main types of machine learning categorization are supervised and unsupervised [33]. It plays significant roles in data-driven decision-making. Data-driven decision-making methods backed by machine learning models are needed and used everywhere from pharmaceuticals [33], education [34–36], transportation, science, health [37,38], Edge Computing [39], energy [40,41], smart cities [42,43], and many more. As indicated in Tables 1 and 2, machine learning methods have been employed to assess the security status of Ethereum smart contracts. Some researchers have used ML in blockchain smart contracts to propose novel systems and schemes in numerous domains but none in the insurance sector. Consequently, this work proposes the use of decision tree algorithms to make efficient, and accurate fraud-related decisions in healthcare scoping the claims submission and processing.

4.5. Paradigm shift of fraud detection systems

Following the exploration of blockchain in other non-financial domains, the focus has drifted from cryptocurrency to security, efficiency, and integrity in data management, access control, privacy, and others. This implies that blockchain technology is now sought after for any other purpose other than that of the provision of resolution to the double-spending problem [44–47], and in these domains, the use of blockchain technology is to enhance internal operations by recording every transaction securely, transparently, immutably, and to ensure ownership, and accountability. According to Amponsah et al. [2] blockchain can be applied in the insurance sector to enhance the submission and processing of claims, for efficient data management, fraud detection and prevention, data identification and ownership, secured and privacy assured data sharing, ridding of middlemen, and prevention of money laundering. In this work, we use blockchain technology for storage of data securely and immutably. According to Wang et al. [11] the National Health Insurance Scheme processes an average of 2.4 million claims resulting in an estimated 28.8 million claims data points annually. The amount of data generated by the scheme makes it data-rich for the use of machine learning. Hypothetically, with the dawn of machine learning, big data analytics, and modern technologies, the modelling and development of the logic, operations, and workflows of a blockchain smart contract through the traditional static procedural, and error-prone approach may not be efficient in contemporary NHIS claims systems [33–43]. Consequently, we are proposing the use of machine learning algorithms and blockchain technology for the submission and processing of claims to detect and prevent fraud in the NHIS and claims-related systems. Our proposed system can be applied to all the top most fraud-prone insurance sub-domains (Health, Automobile, Marine, Long-Term care, Disability, Etc.) [2].

5. Experimental design and methods

The machine learning classification task was executed to obtain the classification results and then weighed against the benchmark. This was done before the extraction of the decision rules which were implemented in the smart contract. We used four variants of the decision tree — J48, ADTree, BFTree, and REPTree. Section 4 presents the results obtained from the machine learning experiments.

5.1. Machine learning software and blockchain development environment

This work leveraged the exquisite computational and graphical power of WEKA version 3.6 to perform the data mining simulations. Unless otherwise indicated, the simulation used 10-Fold Cross Validation to avoid statistically biased data mining results. Reproducibility is key in machine learning research and to allow that, all the classifiers were used with their default configuration files. Again Remix Integrated Development Environment (IDE) was used to develop the blockchain fraud detection and prevention system. The smart contract was tested using JavaScript VM and MetaMask.

Decision Tree Classification Algorithm

The employed machine learning algorithms are the variants of the Decision Tree (DT) algorithm. We used J48, Alternating Decision Tree (ADTree), Best First Tree (BFTree), and Reduced Error Pruning (REPTree). The Decision Tree classification algorithm works by generating a series of logically connected nodes [48]. The nodes are choices among several alternatives, and each terminal node is denotes a classification [49]. Decision trees offer an explicit concept description for a dataset and have the potential to be very effective predictors. A decision tree algorithm according to Chen et al. [50] starts at the root node, tests the attribute, and then continues along the branch of the tree that corresponds to the value of the attribute to do classification. Up till a terminal node is reached, the procedure is repeated. Each not is split into two and the splitting scheme measures the level of impurity at the node [49]. The Gini Index and the Information Gain are the most popularly used splitting criteria in decision tree. Information Gain is reached using the following:

$$Info\left(\left[P_1, P_2, p_3 \ldots, P_n\right]\right) = entropy\left(p_1, p_2, p_3, \ldots, p_n\right)$$

where $P_1, i = 1, 2, 3, \ldots, n$ represent the number of cases for each label. The estimated value of p_i is the figure of P_i divided by the sum of all P_i [49]. This is how the Gini index is calculated:

$$gini\left(p_1, p_2, p_3, \ldots, p_n\right) = \sum_{j \neq i} p_i p_j,$$

$$gini\left(p_1,p_2,p_3,\dots,p_n\right)=\sum_{j\neq i}p_j\left(1-p_i\right)=1-\sum_j p_j^2,$$

where the probabilities of instances in classes i and j, respectively, are p_i and p_j [49,50].

The J48 algorithm (Java version of C4.8) is an improvement and expansion of Quinlan's ID3 algorithm (C4.5). **Alternating Decision Tree (ADTree)** algorithm is based on boosting and generates classification rules that are often smaller and easier to interpret than the rules generated by the C5.0. It was introduced by Freund and Mason [51]. **Best First Tree (BFTree)** expands the nodes in best-first order instead of a fixed order. This method uses binary split for both numeric and nominal attributes. For missing values, the method of 'fractional' instances is used. The best node is the node that reduces the level of impurity among all the nodes available for splitting. best-first decision trees are constructed in the divide-and-conquer fashion. Reference can be made to Friedman et al. [52] and Shi [49] for more information. **Reduced Error Pruning (REPTree) is** a Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). REPTree only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).

5.2. Hardware specification

The desktop personal computer used for the experiments possesses the following characteristics:

- Operating System: Windows 10 Pro 64-bit
- Processor: Intel (R) Core (TM) i7-6700 CPU @ 3.40 GHz
- Memory: 16 384 MB RAM
- DirectX Version: DirectX 12

5.3. Performance evaluation methods

The work employed the Accuracy, Sensitivity, Specificity, RMSE, MAE, and KAPPA Statistics to test the robustness of the produced DT model. These evaluation methods are used because of their popularity.

Accuracy: According to Nisbet et al. [53], the accuracy of a classification algorithm is found by identifying the total number of correctly classified instances over the total number of instances. The accuracy is calculated using Eq. (1).

$$Accuracy=\frac{TP+TN}{N}\tag{1}$$

Sensitivity: The sensitivity is calculated using Eq. (2). Sensitivity points to the number of positive instances that are correctly classified and it is equal to *TP* divided by *TP* plus *TN* [54,55].

$$Sensitivity=\frac{TP}{TP+FN}\tag{2}$$

Specificity: The specificity is also calculated using Eq. (3) and it refers to the proportion of negative instances that are correctly classified as negative. Specificity is equal to the proportion of TN divided by TN plus FP [53,55].

$$Specificity=\frac{TN}{TN+FP}\tag{3}$$

Root Mean Square Error (RMSE): The Root Mean Square Error sometimes called the Root-Mean-Square Deviation is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. RMSE is calculated using (4).

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}\left\|y(i)-\hat{y}(i)\right\|^2}{N}}\tag{4}$$

Mean Absolute Error (MEA): The Mean Absolute Error is calculated using (5). This is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|\tag{5}$$

Kappa: The Kappa result is used to test the interrater and intrarater reliability. The Kappa Statistics is used to determine the randomness in an agreement between two or more observers. Researchers mostly use this method for the calculation of the level of agreement between two or more viewers [56,57], and in a practical case, is used to measure the agreement of predictions made concerning the true class. A kappa value of 0 means agreement by chance whereas a kappa value of 1 means perfect agreement [56]. Any kappa value between 0.81 and 0.99 is considered an "almost perfect agreement" [56,57].

$$K=\frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}\tag{6}$$

Table 3
Description of the dataset.

S/N	Attribute	Type	Distinct values
1	Medication_1	Nominal	17
2	Medication_2	Nominal	17
3	Medication_3	Nominal	15
4	Medication_4	Nominal	8
5	Medication_5	Nominal	8
6	Claimed_Amount	Numeric	Min (15.71); Max (76.96)
7	Label	Nominal	Y (Yes); N (No)

5.4. Description of dataset

This work uses an original domain-specific dataset to generate the domain knowledge in the classification rules module. The dataset used for the data mining experiment contains 7 attributes (label inclusive), and 1323 instances. Out of the 7 attributes, 6 were nominal and 1 is numeric as shown in Table 3.

5.4.1. Data collection procedure

The data was obtained from the National Health Insurance Scheme after they have been vetted and their respective service providers have been paid. The contemporary claims submission and processing method is manual and allows the service providers to administer at most five drugs hence the medication 1–5. The dataset is deemed to be an absolute representation of reality as it was first submitted to the scheme by the service providers, audited by the fulfilling and the vetting officers, and the legitimate amounts paid accordingly. Reference can be made to Amponsah et al. [3] and Wang et al. [11] for an extensive description of the claims processing lifecycle.

5.4.2. Pre-processing of the data

The data was preprocessed to ensure complete privacy protection of the scheme and the patients. Consequently, identifiers like names, IDs, etc., and quasi-identifiers like Date of Birth, hometown, etc. were all removed to ensure k-Anonymity. Following Kittoe and Asiedu-Addo [58] the data was preprocessed according to the drug list and ailments as identified by the NHIS. For example, a drug like Artemether + Lumefantrine and Amoxicillin are known to be used for the cure of Malaria and infections respectively. Therefore all the drugs were converted to their usage. Preliminary data analysis indicated that the most administered drugs in descending order are pain killer, infection, and malaria. Consequently, the data was encoded as the following — 1 = "Pain Killer", 2 = "Infection", and 3 = "Malaria".

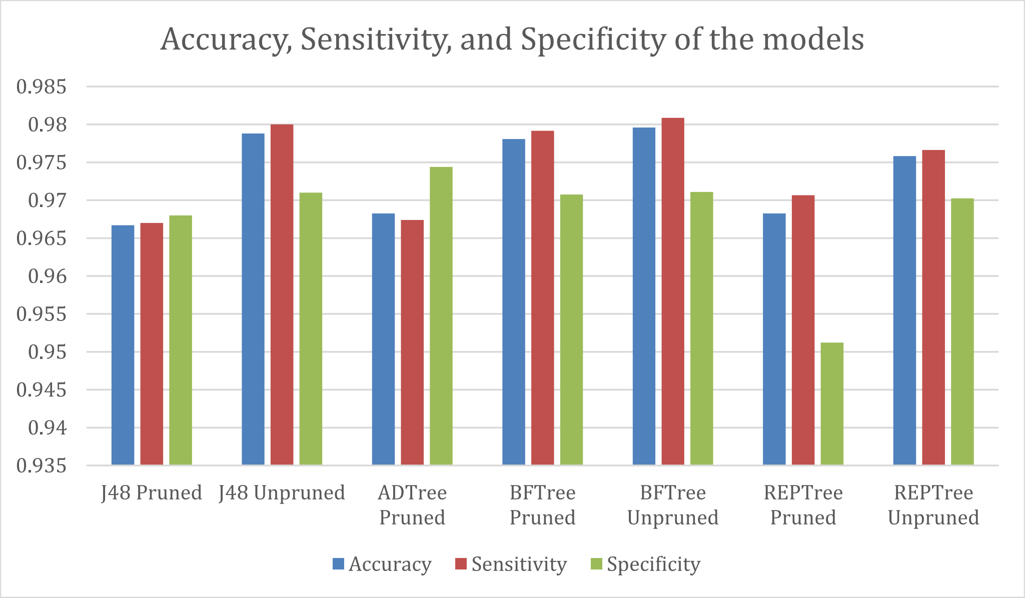


Fig. 1. Accuracy, sensitivity, and specificity of the decision tree algorithms.

Table 4
Accuracy, sensitivity and specificity of the decision tree algorithms.

Classifier		Accuracy	Sensitivity	Specificity
J48	Pruned	0.9667	0.967	0.968
	Unpruned	0.9788	0.98	0.971
ADTree	Pruned	0.968254	0.9674	0.9744
	Unpruned	NA	NA	NA
BFTree	Pruned	0.978080	0.979167	0.970760
	Unpruned	0.97959	0.98087	0.971098
REPTree	Pruned	0.968254	0.970664	0.95122
	Unpruned	0.975813	0.976623	0.970238

Table 5
RMSE, MAE, and Kappa Statistics of the decision tree algorithms.

Classifier		RMSE	MAE	Kappa statistics
J48	Pruned	0.1786	0.0581	0.8539
	Unpruned	0.1422	0.0292	0.9104
ADTree	Pruned	0.1882	0.1148	0.8606
	Unpruned	NA	NA	NA
BFTree	Pruned	0.1498	0.0336	0.907
	Unpruned	0.1443	0.0292	0.9138
REPTree	Pruned	0.1714	0.051	0.8631
	Unpruned	0.152	0.0329	0.8967

6. Experimental results

This section presents the results of the machine learning experiments and model performances, and the proposed implementations in the blockchain fraud detection and prevention system. We present the high-level system abstraction and the extracted classification rules.

6.1. The performance results of the decision tree

6.1.1. Accuracy, sensitivity, specificity

From Table 4, the accuracy, sensitivity, and specificity of the pruned and unpruned DT models are presented from Eqs. (1)–(3). It is seen that the pruned J48 DT achieved the values of 0.9667, 0.967, and 0.968 for the accuracy, sensitivity, and specificity respectively. Similarly, the unpruned J48 DT obtained the values of 0.9788, 0.98, and 0.971 for the accuracy, sensitivity, and specificity respectively. Again, the pruned ADTree achieved accuracy, sensitivity, and specificity values of 0.968254, 0.9674, and 0.9744 respectively. Contrarily, ADTree was not prunable. Pruned and unpruned BFTree obtained accuracy values of 0.978080 and 0.97959, sensitivity values of 0.979167 and 0.98087, and specificity values of 0.970760 and 0.971098 respectively. Similarly, pruned and unpruned REPTree obtained accuracy values of 0.968254, and 0.975813, sensitivity values of 0.970664, and 0.976623, and specificity values of 0.95122, and 0.970238.

It can be deduced from Table 4 and Fig. 1 that the unpruned BFTree model performed better regarding accuracy and sensitivity. However, the pruned ADTree obtained the highest specificity.

6.1.2. Root mean square error, mean absolute error, and kappa statistics

Again, other performance metrics were used to test the performance of the models — thus using Root Mean Squared Error, Mean Absolute Error, and the Kappa Statistics using the formulas indicated in Eqs. (4), (5), and (6). The values of 0.1786, 0.0581, and 0.8539 were obtained for the RMSE, MAE, and Kappa Statistics respectively of the pruned J48 DT model. Similarly, the values of 0.1422, 0.0292, and 0.9104 were obtained for the RMSE, MAE, and Kappa Statistics respectively of the unpruned J48 DT model. Pruned ADTree acquired the values of 0.1882, 0.1148, and 0.8606 for RMSE, MAE, and Kappa statistics respectively. The pruned and unpruned BFTree respectively obtained the RMSE values of 0.1498, and 0.1443, MAE of 0.0336, and 0.0292, and Kappa statistics of 0.907, and 0.9138. Likewise, the pruned and unpruned REPTree accordingly obtained RMSE values of 0.1714, and 0.152, MAE values of 0.051, and 0.0329, and Kappa statistic values of 0.8631, and 0.8967. This is summarized in Table 5 (see Fig. 2).

6.1.3. Decision tree classification rules

The principal reason for using the DT classification algorithm is the ability to produce human-readable, and easy-to-implement classification rules. The classification rules consist of *If-Then-Else* statements that contain the ML discovered pattern/knowledge. This discovered pattern can be implemented in any other system (expert system, robots, embedded system, etc.). For simplicity, ease, and space, the classification rules for the pruned J48 DT model are recommended. The original classification rules are shown in Appendix and the translated classification rules are shown in 6.1.3.2.

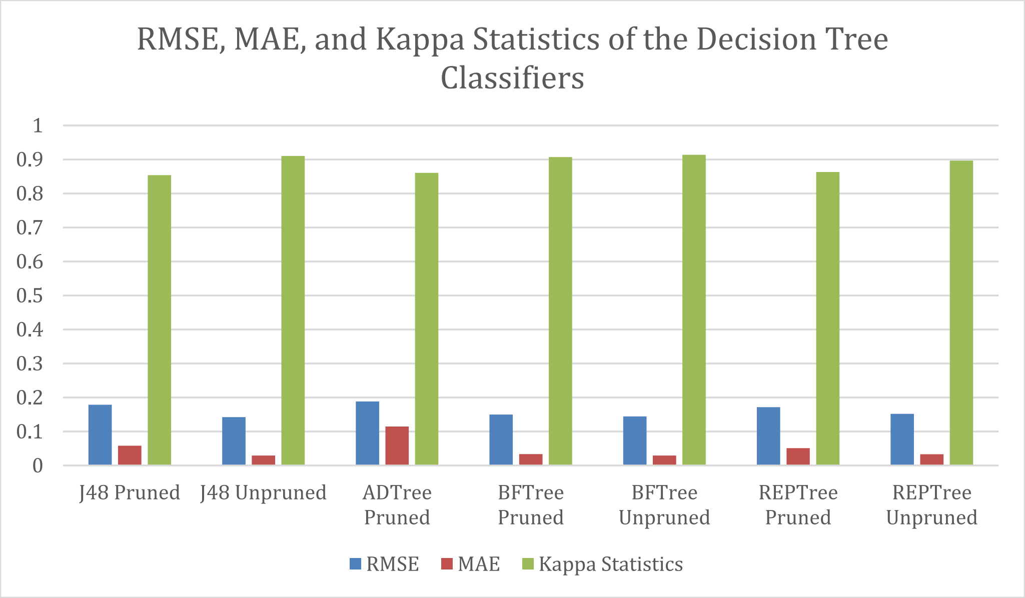
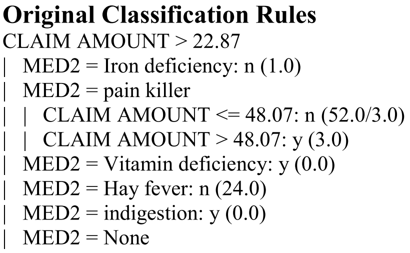


Fig. 2. RMSE, MAE, and Kappa Statistics of the decision tree algorithms.

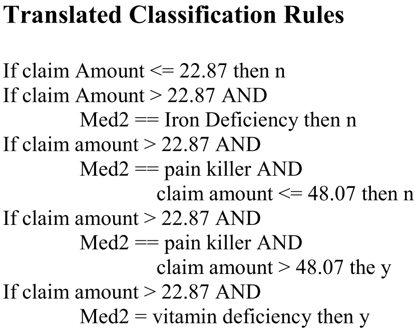
6.1.3.1. Extracted classification rules. These classification rules were produced by the J48 pruned decision tree. The size of the tree is 39 and the number of leaves is 35. The rules of the pruned decision tree were used because the rules are concise yet representative enough to contain the classifications, knowledge, and factors rich enough for the smart contract to make informed decisions. Reference can be made to the appendix for the complete classification rules.

6.1.3.2. Specimen of the decision tree classification rules. The following is a specimen of the programmable classification rules obtained from the J48 pruned DT. Space limitation has resulted in the shortening of the lines. The lines are re-written in the blockchain smart contract.



Original Classification Rules

```
CLAIM AMOUNT > 22.87
|  MED2 = Iron deficiency: n (1.0)
|  MED2 = pain killer
|  |  CLAIM AMOUNT <= 48.07: n (52.0/3.0)
|  |  CLAIM AMOUNT > 48.07: y (3.0)
|  MED2 = Vitamin deficiency: y (0.0)
|  MED2 = Hay fever: n (24.0)
|  MED2 = indigestion: y (0.0)
|  MED2 = None
```



Translated Classification Rules

```
If claim Amount <= 22.87 then n
If claim Amount > 22.87 AND
        Med2 == Iron Deficiency then n
If claim amount > 22.87 AND
        Med2 == pain killer AND
                claim amount <= 48.07 then n
If claim amount > 22.87 AND
        Med2 == pain killer AND
                claim amount > 48.07 the y
If claim amount > 22.87 AND
        Med2 = vitamin deficiency then y
```

6.2. Smart contract implementation

Fig. 3 shows the Solidity Remix interface and the sample source used to collect inputs from the users of the smart contract. Fig. 3a shows the beginning of the smart contract source code and the structure of the claims data. Fig. 3b also shows the fraud detection module which is used to collect and evaluate the user data in step 5 of the system abstraction. In line with the domain dataset, the data entry function accepts 6 different data of which 5 are string data and 1 is an unsigned integer data. The algorithm below represents the pseudocode of the fraud detection module in the smart contract.

6.2.1. Edifice of the proposed system

Steps Involved in the Edifice

These steps are followed to combine the algorithms of machine learning with the blockchain smart contract. The objective of this work used the Decision Tree algorithm for the extraction of the classification rules (extracted knowledge). These rules are incorporated into the smart contract to enhance its efficiency in decision-making.

Step 1 — Machine Learning Algorithm

Machine learning algorithm determines the knowledge/rules to be accepted in place of the people using an efficient and accurate domain data-driven approach. This work processes the use of white-box machine learning algorithms like a decision tree.

Step 2 — Adaptive Decision Rule Generation Module

The extracted knowledge is prepared and incorporated into the blockchain smart contract automatically or by the developer. The developer objectively implements the classification rules in the smart contract avoiding any form of contamination.

Step 3 — Fraud Decision-Making Module

This module is found within the smart contract and it engages when a user enters data into the system using the data collection module. This module contains the ML data-driven classification rules extracted by the machine learning algorithm in steps 1 and 2 and the rules are used to make an enhanced decision in the smart contract.

Step 4 — Data Entry Console

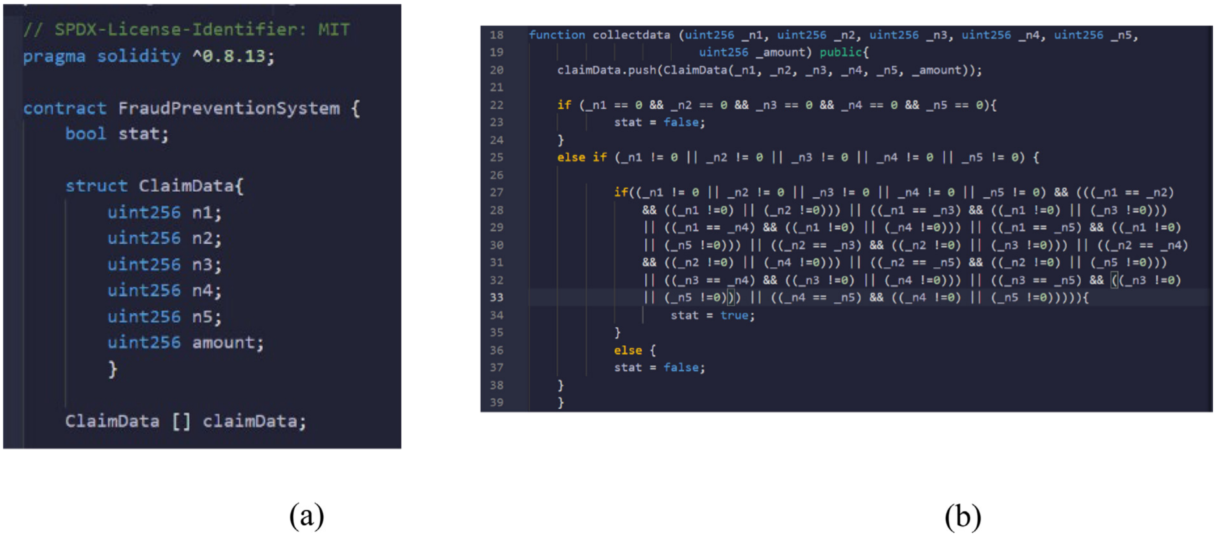
The data entry console is called by the users purely for the entry of data into the proposed system. As indicated in Fig. 5a, the data entry forms are shown to the user with appropriate labels. This console is directly connected to the data collection module (step 4) which stores the entered data and presents them to the decision-making module in step 5.

Step 5 — Data Collection Module

This component is a function within the smart contract that detects the status of the entered data based on the programmed adaptively

Fraud Detection Algorithm

```
WHILE TRUE DO
    Collect ClaimData ($n_1$, $n_2$, $n_3$, $n_4$, $n_5$, Amount);
    IF ($n_1$ == $n_2$ == $n_3$ == $n_4$ == $n_5$ == 0){
    claimDataStatus = False;
        } ELSE IF ($n_1$ != 0 or $n_2$ != 0 or $n_3$ != 0 or $n_4$ != 0 or $n_5$ != 0){
            IF (($n_1$ != 0 or $n_2$ != 0 or $n_3$ != 0 or $n_4$ != 0 or $n_5$ != 0)) &&
            ((($n_1$ == $n_2$) && ($n_1$ = 0 or $n_2$ = 0)) … (($n_4$ == $n_5$) && ($n_4$ = 0 or $n_5$ = 0))){
                claimDataStatus = True;
            }
            ELSE {
            claimDataStatus = False;
            }
        }
    }
```



```
// SPDX-License-Identifier: MIT
pragma solidity ^0.8.13;

contract FraudPreventionSystem {
    bool stat;

    struct ClaimData{
        uint256 n1;
        uint256 n2;
        uint256 n3;
        uint256 n4;
        uint256 n5;
        uint256 amount;
        }

    ClaimData [] claimData;
```

```
function collectdata (uint256 _n1, uint256 _n2, uint256 _n3, uint256 _n4, uint256 _n5,
                        uint256 _amount) public{
    claimData.push(ClaimData(_n1, _n2, _n3, _n4, _n5, _amount));

    if (_n1 == 0 && _n2 == 0 && _n3 == 0 && _n4 == 0 && _n5 == 0){
            stat = false;
    }
    else if (_n1 != 0 || _n2 != 0 || _n3 != 0 || _n4 != 0 || _n5 != 0) {

            if((_n1 != 0 || _n2 != 0 || _n3 != 0 || _n4 != 0 || _n5 != 0) && (((_n1 == _n2)
                && ((_n1 !=0) || (_n2 !=0))) || ((_n1 == _n3) && ((_n1 !=0) || (_n3 !=0)))
                || ((_n1 == _n4) && ((_n1 !=0) || (_n4 !=0))) || ((_n1 == _n5) && ((_n1 !=0)
                || (_n5 !=0))) || ((_n2 == _n3) && ((_n2 !=0) || (_n3 !=0))) || ((_n2 == _n4)
                && ((_n2 !=0) || (_n4 !=0))) || ((_n2 == _n5) && ((_n2 !=0) || (_n5 !=0)))
                || ((_n3 == _n4) && ((_n3 !=0) || (_n4 !=0))) || ((_n3 == _n5) && ((_n3 !=0)
                || (_n5 !=0))) || ((_n4 == _n5) && ((_n4 !=0) || (_n5 !=0))))){
                    stat = true;
            }
            else {
            stat = false;
    }
    }
```

(a) (b)

Fig. 3. Sample Solidity smart contract. (a) Structure of the claims data (b) Fraud detection module of the smart contract.

generated decision rule in step 2. The data together with the determined status are stored in the blockchain (step 6) and the worldstate (step 7) systems respectively.

Step 6 — Blockchain Ledger

The blockchain ledger records all the transactions of the proposed system. These transactions include the deployment of the smart contract, data views, and creation, state mutability, etc. The transactions in the blockchain ledger are secure, immutable, transparent, auditable, and traceable [2]. In a private blockchain system, access to the distributed blockchain ledger is limited to authorized users only [3,59].

Step 7 — WorldState

The world state contains the latest data in the ledger. This component of the system is modifiable (change and delete). In other words, if a user initially stores “112 233” ledger and later updates it to “987 654”, the ledger will show 987 654 when accessed. However, to ensure security, and transparency in the system all these actions are immutably recorded in the blockchain ledger.

Step 8 — Rejection Logs

This entity is introduced in the framework referencing the margin of error obtained during the decision tree classification process, and as such data that are classified as fraud/false/negative are not deleted or prevented from storage. However, they are logged in the rejection logs (step 8), awaiting the action of the system administrator. It is seen from Table 5 that the J48 pruned DT obtained RMSE of 0.1786, MAE of 0.0581, and Kappa value of 0.8539. This implies that the classification rules make data-driven decisions with these error margins and as such, an expert view is required to completely label the entered data as classified by the algorithm.

6.3. System related costs

Table 6 shows the costs associated with running the proposed system. The one-time deployment cost of the smart contract is 0.00079078 ETH ($ 0.82) and the state change cost associated with the collection of claims data is 0.00021753 ETH ($ 0.23).

7. Discussion

This work has applied blockchain technology and decision tree classification rules to detect and prevent healthcare fraud. The Blockchain technology adds a layer of security, decentralization, accountability, auditability, etc. to the existing centralized claims processing systems thereby totally eradicating the reliance on the trust of the claimants (healthcare providers). Also, the decision tree classification rules equip the blockchain smart contract with data-driven decision-making capabilities in which sense fraud is detected and prevented using domain specific data [33–43]. In general, this work contributes to the expansion of blockchain 3.0 [1,60] to include the insurance sector. Several efforts and propositions have been made to mitigate healthcare fraud [3] in which light this work provides a novel approach to avoiding fraud by integrating decision tree classification rules and the blockchain smart contract. Adding the fraud detective and preventive smart contract

Table 6
Costs associated with key system related tasks.

System task	Eth usage	Currency ($)	Transaction hash	Status
Smart contract deployment	0.00079078	0.82	0x752f45b4c172cd00dfa15fc75edea826d5bd3543874b05e852181872faa8134	Success
Data collection	0.00021753	0.23	0x6e256233c5973ed55ad5234111dec6dc5b98e5d4437858b029126bef985cd667	Success

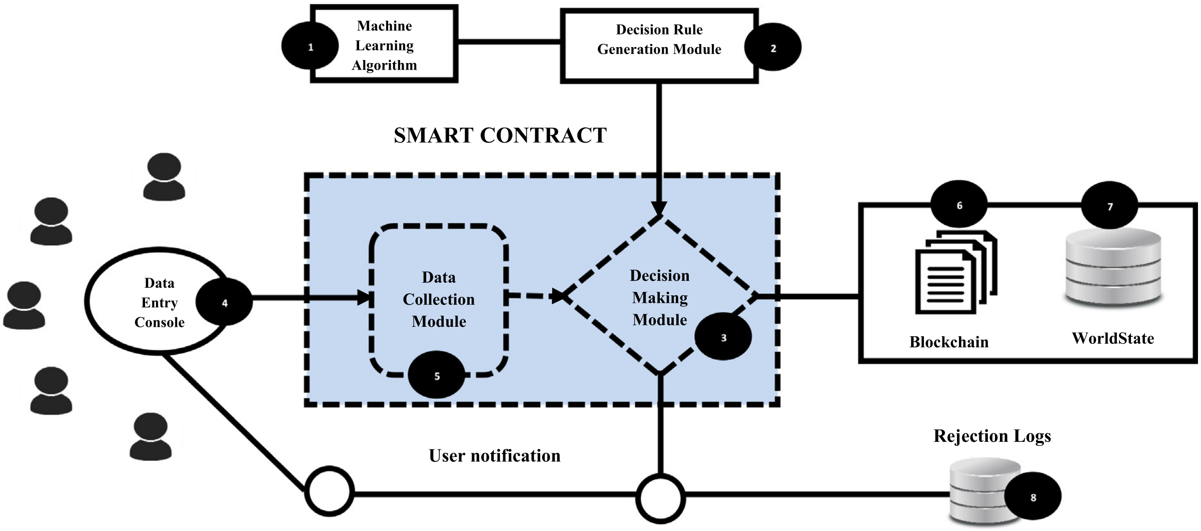
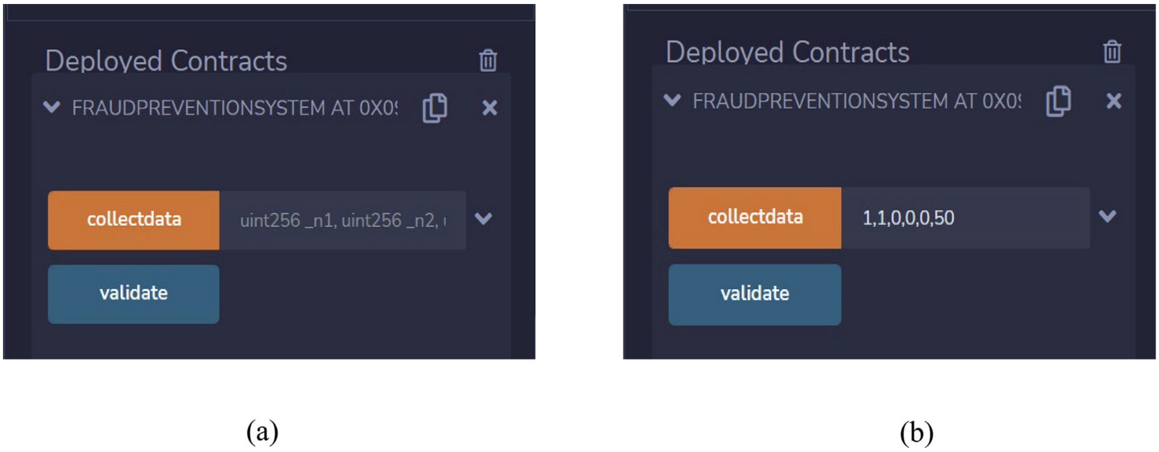


Fig. 4. High level system view of the fraud prevention system.


Deployed Contracts
FRAUDPREVENTIONSYSTEM AT 0X0!
collectdata
uint256 _n1, uint256 _n2, ı
validate
Deployed Contracts
FRAUDPREVENTIONSYSTEM AT 0X0!
collectdata
validate

1,1,0,0,0,50

(a) (b)

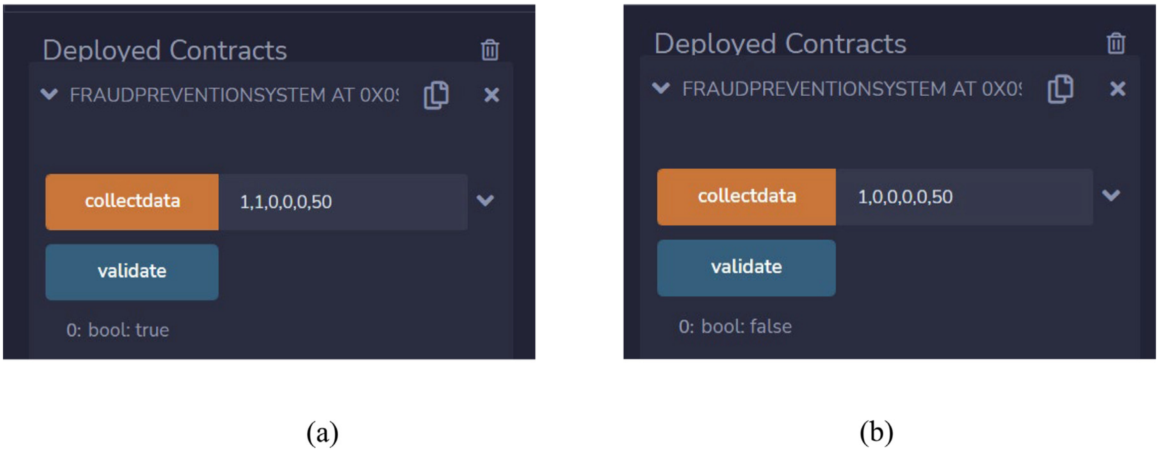
Fig. 5. Interface of the smart contract. (a) Empty data entry form, (b) Populated data entry form.

changes the paradigm shift of healthcare fraud prevention systems and provides a next-level efficient claims processing system for Africa and global insurance schemes.

Reliability in data management and predictive models is key and it is seen from Table 4 that the unpruned BFTree yielded the best accuracy and specificity of 97.96% and 98.09% respectively. However, the ADTree classifier achieved the highest specificity of 97.44%. This implies that once adopted, the BFTree classification rules extracted and implemented detects fraudulent claims with an accuracy of about 98%. The RMSE, MAE, and the Kappa statistics of all the experimented DT achieved exquisite results. For example, in Table 5 the unpruned BFTree achieved RMSE, MAE, and Kappa values of 0.1443, 0.0292, and 0.9138 respectively. According to the benchmarks indicated by Amponsah et al. [3], Viera and Garrett [56], Ojo [61], Hajjar [62], and McHugh [57], these values are excellent and prove that the proposes system is robust. Claims may be marked as "fraudulent" for several reasons — i. truly fraudulent, ii. Human error, and iii. Machine misclassification due to the error margin or programmer blunder. Consequently, we propose that claims officers should review the rejected claims to prove that they are indeed fraudulent.

Healthcare fraud cost lots of funds to be lost worldwide [3,9]. Although the proposed system requires operational cost, it is worthy compared to the annual healthcare costs lost to fraud. From Table 6, an average of $0.82 is spent on the deployment of the proposed smart contract and $0.23 is needed to store a record.

The high-level system abstraction of the system has 8 different components (Machine Learning Algorithm, Adaptive Decision Rule Generation Module, Decision Making Module, Data Entry Console, Data Collection Module, Blockchain Ledger, WorldState, and the Rejection Logs). Any machine learning algorithm can be used in component 1. However, we propose the use of the method that performs whitebox computation like Decision Tree [63,64]. Then the knowledge produced



Deployed Contracts

FRAUDPREVENTIONSYSTEM AT 0X0!

collectdata 1,1,0,0,0,50

validate

0: bool: true

(a)

Deployed Contracts

FRAUDPREVENTIONSYSTEM AT 0X0!

collectdata 1,0,0,0,0,50

validate

0: bool: false

(b)

Fig. 6. Smart contract case based testing using real data.(a) Data evaluated to true (b) Data evaluated to false.

after the machine learning experiments is generated/extracted in component 2 and then implemented in the blockchain smart contract. Components 1 and 2 are performed outside the ethereum smart contract therefore unbiased experiments and programming is obliged.

The successful implementation of the proposed system following the usage of domain-specific data has provided an efficient and cost saving solution for the health insurance schemes in Africa. However, it can be reauthored to suite other schemes worldwide and other domains like Merrad et al. [19], Hassija et al. [23], Hu et al. [20], Tsoukas et al. [21], and Feng et al. [24] (see Figs. 4 and 6).

8. Conclusion

While the effect of healthcare fraud is common knowledge, the world continues to suffer from it. The evidence is crystal from the numerous reports and amounts published in the literature. Blockchain's disruptive nature in the provision of contemporary solutions cannot be over-emphasized. Considering the annual cumulative funds lost to fraud, the need for a novel approach to process health insurance claims is momentous and requires the greatest attention. By developing and testing a blockchain-based system that uses machine learning and domain data to judge the fraudulent nature of health insurance claims, this work has proposed a novel claims submission and processing system that is secured and makes data-driven decisions. The machine learning experiments imply that the proposed system accurately classified the claims data with an accuracy of about 98%. In a similar light, future claims will be classified with an error rate of about 2%. Although there is a cost ramification in adopting the proposed system, the long-term benefits make it a necessary cost compared to the worldwide annual amounts lost to fraud. Transitioning from the centralized system to the decentralized Blockchain-based system will ensure security, efficiency, and high data integrity in claims processing and also tremendously increase the efforts against fraud.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

Data is available upon request.

Appendix

```
  Sample Extracted Classification Rules

  J48 pruned tree
  ------------------

  CLAIM AMOUNT <= 22.87: n (1043.0/26.0)
  CLAIM AMOUNT > 22.87
  |  MED2 = Iron deficiency: n (1.0)
  |  MED2 = pain killer
  |  |  CLAIM AMOUNT <= 48.07: n (52.0/3.0)
  |  |  CLAIM AMOUNT > 48.07: y (3.0)
  |  MED2 = Vitamin deficiency: y (0.0)
  |  MED2 = Hay fever: n (24.0)
  |  MED2 = indigestion: y (0.0)
  |  MED2 = None
|  |  MED1 = stomach pain: y (0.0)
|  |  MED1 = infection: y (1.0)
|  |  MED1 = malaria: y (0.0)
|  |  MED1 = lung disease: y (0.0)
|  |  MED1 = Blood pressure: y (0.0)
|  |  MED1 = Indigestion: y (0.0)
|  |  MED1 = None: y (115.0/1.0)
|  |  MED1 = Deficiency Anemia: n (7.0/1.0)
|  |  MED1 = Pain Killer: y (0.0)
|  |  MED1 = Skin Infections: y (0.0)
|  |  MED1 = Bacteria Infections: y (0.0)
|  |  MED1 = Iron deficiency: y (0.0)
|  |  MED1 = Allergy reliever: y (0.0)
|  |  MED1 = Diabetes: y (0.0)
|  |  MED1 = Hypertension: y (0.0)
|  |  MED1 = Skin irritation: y (0.0)
|  |  MED1 = mental/mood problems: y (0.0)
|  MED2 = cough: y (0.0)
|  MED2 = Tetanus infection: y (0.0)
|  MED2 = skin infection: y (0.0)
|  MED2 = Dehydration: y (0.0)
|  MED2 = Hypertension: y (0.0)
|  MED2 = Infection: n (12.0/1.0)
|  MED2 = malaria: y (36.0/1.0)
|  MED2 = Stomach pain: y (0.0)
|  MED2 = Pain Killer: n (29.0/6.0)
|  MED2 = Diabetes: y (0.0)
|  MED2 = Deficiency Anemia: y (0.0)

Number of Leaves:       35
Size of the tree:   39
```

References

[1] M. Swan, Blockchain: Blueprint for a New Economy, O'Reilly Media, Inc, 2015.

[2] A.A. Amponsah, B.A. Weyori, A.F. Adekoya, Blockchain in insurance: Exploratory analysis of prospects and threats, Int. J. Adv. Comput. Sci. Appl. (2021) 12.

[3] A.A. Amponsah, A.F. Adekoya, B.A. Weyori, Improving the financial security of national health insurance using cloud-based blockchain technology application, Int. J. Inf. Manage. Data Insights 2 (2022) http://dx.doi.org/10.1016/j.jjimei.2022.100081, (2022).

[4] P.K. Meduri, S. Mehta, K. Joshi, S. Rane, Disrupting insurance industry using blockchain, in: International Conference on Intelligent Data Communication Technologies and Internet of Things, Springer, Cham, 2018, pp. 1068–1075.

[5] I. Nath, Data exchange platform to fight insurance fraud on blockchain, in: 2016 IEEE 16th International Conference on Data Mining Workshops, ICDMW, IEEE Computer Society, 2016, pp. 821–825.

[6] J. Xu, Y. Wu, X. Luo, D. Yang, Improving the efficiency of blockchain applications with smart contract based cyber-insurance, in: ICC 2020-2020 IEEE International Conference on Communications, ICC, IEEE, 2020, pp. 1–7.

[7] M. Kirlidog, C. Asuk, A fraud detection approach with data mining in health insurance, Proc.-Soc. Behav. Sci. 62 (2012) 989–994.

[8] J. Gee, M. Button, The financial cost of healthcare fraud 2014: What data from around the world shows, 2014.

[9] A.Y.B.R. Thaifur, M.A. Maidin, A.I. Sidin, A. Razak, How to detect healthcare fraud? A systematic review, Gaceta Sanit. 35 (2021) S441–S449.

[10] GenKey SOLUTIONS, B.V., Reducing healthcare fraud in Africa - What we do. How we do it, 2016, Accessed on July 16, 2022 at https://www.genkey.com/wp-content/uploads/2016/11/Healthcare_ebook_EN-version-2.0.pdf.

[11] H. Wang, N. Otoo, L. Dsane-Selby, Ghana National Health Insurance Scheme: Improving Financial Sustainability Based on Expenditure Review, World Bank Publications, 2017.

[12] F.A. Dake, Examining equity in health insurance coverage: an analysis of Ghana's National Health Insurance Scheme, Int. J. Equity Health 17 (1) (2018) 1–10.

[13] P. Momeni, Y. Wang, R. Samavi, Machine learning model for smart contracts security analysis, in: 17th International Conference on Privacy, Security and Trust (PST), IEEE, 2019, pp. 1–6.

[14] M. Eshghie, C. Artho, D. Gurov, Dynamic vulnerability detection on smart contracts using machine learning, Eval. Assess. Soft. Eng. (2021) 305–312.

[15] E. Bandara, W.K. Ng, N. Ranasinghe, K.D. Zoysa, Aplos: smart contracts made smart, in: International Conference on Blockchain and Trustworthy Systems, Springer, Singapore, 2019, pp. 431–445.

[16] W. Wang, J. Song, G. Xu, Y. Li, H. Wang, C. Su, Contractward: Automated vulnerability detection models for ethereum smart contracts, IEEE Trans. Netw. Sci. Eng. 8 (2) (2020) 1133–1144.

[17] W.J.W. Tann, X.J. Han, S.S. Gupta, Y.S. Ong, Towards safer smart contracts: A sequence learning approach to detecting security threats, arXiv preprint (2018) arXiv:1811.06632.

[18] N. Ashizawa, N. Yanai, J.P. Cruz, S. Okamura, Eth2Vec: learning contract-wide code representations for vulnerability detection on ethereum smart contracts, in: Proceedings of the 3rd ACM International Symposium on Blockchain and Secure Critical Infrastructure (2021) pp. 47–59.

[19] Y. Merrad, M.H. Habaebi, M.R. Islam, T.S. Gunawan, E.A. Elsheikh, F.M. Suliman, M. Mesri, Machine learning-blockchain based autonomic peer-to-peer energy trading system, Appl. Sci. 12 (7) (2022) 3507.

[20] X. Hu, C. Zhu, Z. Tong, W. Gao, G. Cheng, R. Li …, J. Gong, Identifying ethereum traffic based on an active node library and DEVp2p features, Future Gener. Comput. Syst. 132 (2022) 162–177.

[21] V. Tsoukas, A. Gkogkidis, A. Kampa, G. Spathoulas, A. Kakarountas, Enhancing food supply chain security through the use of blockchain and TinyML, Information 13 (5) (2022) 213.

[22] L. Zhang, J. Wang, W. Wang, Z. Jin, C. Zhao, Z. Cai, H. Chen, A novel smart contract vulnerability detection method based on information graph and ensemble learning, Sensors 22 (9) (2022) 3581.

[23] V. Hassija, R. Ratnakumar, V. Chamola, S. Agarwal, A. Mehra, S.S. Kanhere, H.T.T. Binh, A machine learning and blockchain based secure and cost-effective framework for minor medical consultations, Sustain. Comput.: Inform. Syst. 35 (2022) 100651.

[24] C. Feng, B. Liu, K. Yu, S.K. Goudos, S. Wan, Blockchain-empowered decentralized horizontal federated learning for 5G-enabled UAVs, IEEE Trans. Ind. Inf. (2021).

[25] D. Kamboj, T.A. Yang, An exploratory analysis of blockchain: applications, security, and related issues, in: Proceedings of the International Conference on Scientific Computing, CSC, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2018, pp. 67–73.

[26] M.R. Ogiela, M. Majcher, Security of distributed ledger solutions based on blockchain technologies, in: 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications, AINA, IEEE, 2018, pp. 1089–1095.

[27] S. Nakamoto, Bitcoin: A Peer-To-Peer Electronic Cash System, Manubot, 2008.

[28] V.J. Morkunas, J. Paschen, E. Boon, How blockchain technologies impact your business model, Bus. Horiz. 62 (3) (2019) 295–306.

[29] C.V. Helliar, L. Crawford, L. Rocca, C. Teodori, M. Veneziani, Permissionless and permissioned blockchain diffusion, Int. J. Inf. Manage. 54 (2020) 102136.

[30] S. De Angelis, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, V. Sassone, PBFT vs proof-of-authority: Applying the CAP theorem to permissioned blockchain, 2018.

[31] N. Szabo, Smart contracts: building blocks for digital markets, EXTROPY: J. Transhumanist Thought, (16) 18 (2) (1996) 28.

[32] K. Christidis, M. Devetsikiotis, Blockchains and smart contracts for the internet of things, Ieee Access 4 (2016) 2292–2303.

[33] S.A. Kumar, T.D. Ananda Kumar, N.M. Beeraka, G.V. Pujar, M. Singh, H.S. Narayana Akshatha, M. Bhagyalalitha, Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry, Future Med. Chem. 14 (4) (2022) 245–270.

[34] V. Hamilton, Y. Onder, N.R. Andzik, T.D. Reeves, Do data-driven decision-making efficacy and anxiety inventory scores mean the same thing for pre-service and in-service teachers? J. Psychoeduc. Assess. (2022) 07342829211069220.

[35] J.L. Barton, B.A. Akin, Implementation drivers as practical measures of data-driven decision-making: An initial validation study in early childhood programs, Glob. Implement. Res. Appl. (2022) 1–12.

[36] Y. Teng, J. Zhang, T. Sun, Data-driven decision-making model based on artificial intelligence in higher education system of colleges and universities, Expert Syst. (2022) e12820.

[37] M. Alipour-Vaezi, A. Aghsami, F. Jolai, Prioritizing and queueing the emergency departments' patients using a novel data-driven decision-making methodology, a real case study, Expert Syst. Appl. 195 (2022) 116568.

[38] L.J. Basile, N. Carbonara, R. Pellegrino, U. Panniello, Business intelligence in the healthcare industry: The utilization of a data-driven approach to support clinical decision making, Technovation (2022) 102482.

[39] H. Lamaazi, R. Mizouni, H. Otrok, S. Singh, E. Damiani, Smart-3Dm: Data-driven decision making using smart edge computing in hetero-crowdsensing environment, Future Gener. Comput. Syst. 131 (2022) 151–165.

[40] X. Chen, Distributed and Data-Driven Decision-Making for Sustainable Power Systems (Doctoral dissertation), Harvard University, 2022.

[41] F.R. Cecconi, A. Khodabakhshian, L. Rampini, Data-driven decision support system for building stocks energy retrofit policy, J. Build. Eng. (2022) 104633.

[42] I.H. Sarker, Smart city data science: Towards data-driven smart cities with open research issues, Internet Things 19 (2022) 100528.

[43] A.M.S. Osman, A.A. Elragal, A. Ståhlbröst, Data-driven decisions in smart cities: A digital transformation case study, Appl. Sci. 12 (3) (2022) 1732.

[44] R. Guerraoui, P. Kuznetsov, M. Monti, M. Pavlovic, D.A. Seredinschi, The consensus number of a cryptocurrency, Distrib. Comput. 35 (1) (2022) 1–15.

[45] Z. Hatefi, M. Bayat, M.R. Alaghband, N. Hamian, S.M. Pournaghi, A conditional privacy-preserving fair electronic payment scheme based on blockchain without trusted third party, J. Ambient Intell. Humaniz. Comput. (2022) 1–14.

[46] T. Khanna, P. Nand, V. Bali, BEVDS: A blockchain model for multiparty authentication of COVID-19 vaccine beneficiary, in: Innovative Data Communication Technologies and Application, Springer, Singapore, 2022, pp. 857–869.

[47] K.Y. Kang, Cryptocurrency and double spending history: Transactions with zero confirmation, Econom. Theory (2022) 1–39.

[48] B.T. Pham, A. Jaafari, T. Van Phong, H.P.H. Yen, T.T. Tuyen, V. Van Luong ., L.K. Foong, Improved flood susceptibility mapping using a best first decision tree integrated with ensemble learning techniques, Geosci. Front. 12 (3) (2021) 101105.

[49] H. Shi, Best-First Decision Tree Learning (Doctoral dissertation), The University of Waikato, 2007.

[50] W. Chen, S. Zhang, R. Li, H. Shahabi, Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling, Sci. Total Environ. 644 (2018) 1006–1018.

[51] Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: Icml Vol. 99, 1999, pp. 124–133.

[52] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), Ann. Statist. 28 (2) (2000) 337–407.

[53] R. Nisbet, J. Elder, G. Miner, HandBook of Statistical Analysis and Data Mining Applications, Academic Press, Burlington, MA, 2009.

[54] H.A. Elsalamony, Bank direct marketing analysis of data mining techniques, network, 5, 0, Int. J. Comput. Appl. (0975–8887) 85 (7) (2014) 2014.

[55] G.S. Linoff, M.J. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, John Wiley & Sons, 2004.

[56] A.J. Viera, J.M. Garrett, Understanding interobserver agreement: the kappa statistic, Fam. Med. 37 (5) (2005) 360–363.

[57] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (3) (2012) 276–282.

[58] J.D. Kittoe, S.K. Asiedu-Addo, Exploring fraud and abuse in national health insurance scheme (NHIS) using data mining technique as a statistical model, J. Educ. Stud. Math. Sci. 13 (2017) 13–31.

[59] A. Al Omar, M.S. Rahman, A. Basu, S. Kiyomoto, Medibchain: A blockchain based privacy preserving platform for healthcare data, in: International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, Springer, Cham, 2017, pp. 534–543.

[60] P. Grover, A.K. Kar, M. Janssen, Diffusion of blockchain technology: Insights from academic literature and social media analytics, J. Enterpr. Inf. Manage. (2019).

[61] A.I. Ojo, Validation of the DeLone and McLean information systems success model, Healthc. Inform. Res. 23 (1) (2017) 60–66.

[62] S.T. Hajjar, Statistical analysis: Internal-consistency reliability and construct validity, Int. J. Quant. Qual. Res. Methods 6 (1) (2018) 46–57.

[63] E. Pintelas, I.E. Livieris, P. Pintelas, A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability, Algorithms 13 (1) (2020) 17.

[64] C. Rudin, J. Radin, Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition, 2019, Available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/7 accessed May 8, 2022.