IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2020, accepted March 16, 2020, date of publication March 25, 2020, date of current version April 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983300

A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement

NAJMEDDINE DHIEB[1], (Student Member, IEEE), HAKIM GHAZZAI[1], (Senior Member, IEEE), HICHEM BESBES[2], AND YEHIA MASSOUD[1], (Fellow, IEEE)
[1]School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[2]Higher School of Communications of Tunis, University of Carthage, Ariana 2083, Tunisia

Corresponding author: Hakim Ghazzai (hghazzai@stevens.edu)

ABSTRACT The private insurance sector is recognized as one of the fastest-growing industries. This rapid growth has fueled incredible transformations over the past decade. Nowadays, there exist insurance products for most high-value assets such as vehicles, jewellery, health/life, and homes. Insurance companies are at the forefront in adopting cutting-edge operations, processes, and mathematical models to maximize profit whilst servicing their customers claims. Traditional methods that are exclusively based on human-in-the-loop models are very time-consuming and inaccurate. In this paper, we develop a secure and automated insurance system framework that reduces human interaction, secures the insurance activities, alerts and informs about risky customers, detects fraudulent claims, and reduces monetary loss for the insurance sector. After presenting the blockchain-based framework to enable secure transactions and data sharing among different interacting agents within the insurance network, we propose to employ the extreme gradient boosting (XGBoost) machine learning algorithm for the aforementioned insurance services and compare its performances with those of other state-of-the-art algorithms. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. Furthermore, we propose an online learning solution to automatically deal with real-time updates of the insurance network and we show that it outperforms another online state-of-the-art algorithm. Finally, we combine the developed machine learning modules with the hyperledger fabric composer to implement and emulate the artificial intelligence and blockchain-based framework.

INDEX TERMS Blockchain, data analysis, fraud detection, insurance, machine learning.

I. INTRODUCTION

Nowadays, insurance companies are deprived of a tremendous amount of financial gain due to claims leakage. Fraudulent claims present a huge and a costly problem for insurance companies, potentially leading to billions of dollars of unnecessary expenses for the industry yearly. Insurance fraudsters will often exaggerate or fabricate situations to provide the basis for fraudulent claims. Insurers have historically relied on mathematicians to measure risk and formulate premium rates for policy underwriting that would ensure rational levels of payouts without endangering the company's financial health. Traditional insurance fraud detection methods are complex and time-consuming. They mainly depend on expert scrutiny, adjusters, and special investigation services. Added to that, manual detection results in additional costs and inaccurate results. Moreover, late decisions might lead to extra losses for the insurance companies.

The Association of Certified Fraud Examiners (ACFE), a worldwide anti-fraud organization and a major provider of

The associate editor coordinating the review of this manuscript and approving it for publication was Berdakh Abibullaev.

educational and training programs against fraud, identifies fraud as an act of deception or mistakes made by a person or an entity that knows that the mistake could result in some benefits that are not good to the individual or others [2]. Insurance fraud has led to significant negative impacts to the insurance sector for several decades. It is the second-largest white-collar crime in the US. Insurance fraud leads to an estimated $80 billion in economic losses annually according to the Federal Bureau of Investigation (FBI) [2]. Additionally, it is reported that 21% of bodily injury and 18% of personal injury claims finishing with a full refund are fraudulent [3]. The excessive number of fraudulent claims paid out by auto insurance companies has lead to premiums that have been increased by hundreds of dollars to offset the fraudulent payouts, which harms the competitiveness and quality of services offered by insurance firms. Therefore, there is a pressing need to devise fast and efficient solutions to build fraud detection, risk measurement, and secure data management solutions that maintain a perfect balance between client personal data preservation, loss prevention savings, and investment of false alert detection. From that perspective, we propose to develop an effective framework for insurance companies to help confront such challenges.

In this paper, we present a novel Smart Insurance System based on Blockchain and ARtificial intelligence (SISBAR) to codify business rules, automate claims processing, estimate clients risk levels, and detect fraudulent claims. To the best of our knowledge, there is no previous work using blockchain technology and AI for insurance applications. Nevertheless, there are some contributions proposing the join use of AI and blockchain in healthcare systems [4], smart energy grids [5], and the internet of smart things to make connected devices autonomous [6]. Moreover, unlike existing and suggested systems, SISBAR aims to decentralize the insurance company network and convert it to an efficient system to cope with fraudulent claims and risky customers. In this context, we propose the use of permissioned blockchain as it is highly recommended for applications that require authenticated participants, limit participants authorities, and protect sensitive and personal information. Shared blockchain records can protect insurance companies from fraudsters clients, double claim submission, and improve fraud detection efficiency.

In addition, we propose to employ two machine learning methods to build the fraud detection and risk measurement modules. The first method is based on a batch learning strategy where the algorithm trains the whole dataset at once. The second method is based on an online learning strategy which dynamically trains, updates, and upgrades the learning weights as new data enters the system, without the need to retrain the whole model from scratch as new information arrives. For the offline learning, we propose to employ a novel machine learning algorithm, namely extreme gradient boosting, *aka* XGBoost [7], to detect, classify fraudulent auto insurance claims, predict suspected customers, and estimate their next claim amount based on their risk levels. As an online learning method, we propose the use of a Very Fast Decision Tree (VFDT) algorithm to dynamically train the fraud detection and classification model for insurance companies. Although using batch machine learning algorithms for auto insurance fraud detection have been investigated in literature [8]–[10], there is no previous study used XGBoost for this purpose. Moreover, classifying claims into different fraud types and predicting risky customers have not been investigated using machine learning algorithms. We advocate the use of XGBoost not only for its computational speed and model performance but also for its capability of efficiently resolving diverse problems across several disciplines, such as medicine and cybersecurity [11], [12].

As it will be discussed in the next section, there is no previous work that integrated online classification training and blockchain technology for the purpose of real-time fraud detection. In this study, the online machine learning algorithm uses the data stored on the blockchain to actively adjust and upgrade the model. Data analysis techniques are employed to clean the insurance data and analyze the features to enhance the machine learning algorithm results. Then, using the feature-engineered dataset, tests are conducted to evaluate the performance of the proposed fraud detectors and risk estimators. Next, comparisons with other state-of-the-art algorithms, such as decision tree, naive bayes, and nearest neighbor for offline classification problems, ridge, elastic-net, and gradient boosting algorithms for off line regression problems, and the Stochastic Gradient Decent (SGD) with linear Support Vector Machine (SVM) loss function for online classification, are performed. Results show that the two proposed machine learning algorithms XGBoost and VFDT considerably outperform the other machine learning algorithms. Finally, to develop and emulate the blockchain network, we use hyperledger fabric composer module. Also, a representational state transfer (REST) server is developed to insure the communication between the blockchain, AI, and other application servers via REST application programming interfaces (APIs).

The rest of the paper is organized as follows: Section II provides a literature review about advances in smart insurance systems. Section III presents the proposed insurance network architecture based on AI and blockchain technologies. In Section IV, we describe the AI modules including the machine learning algorithms and learning strategies. Implementation and performance evaluation are carried out in Section V. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

Blockchain has recently attracted much research interest, as it is a breakthough database technology that may aid in the solution of complicated problems across many sectors [13]. Indeed, blockchain technology is no longer associated exclusively to finance and banking applications. This technology has the potential to be applied to a diverse set of sectors including, but not limited to: information security, healthcare, logistics, and insurance. In cybersecurity, blockchain is applied as a method of mitigating Distributed Denial of

Service (DDoS) attacks as described in [14]. The latter study presents a method to reduce DDoS attacks by implementing a private blockchain that uses decentralized Content Delivery Networks (CDNs) with trusted node participants authorized by military or government agencies. In healthcare, a framework based on blockchain, Internet of Things (IoT), and machine learning technologies were introduced in [15] to intercept, fetch, analyze, and store the data collected from IoT devices attached to patients into the blockchain network. The blockchain system was used not only to store and maintain the patient data but also to support access from different stakeholders subscribed into this system. In the insurance domain, few studies have investigated the use of blockchain to ensure transparency and automation [16], [17]. For health insurance companies, blockchain networks ensure a proof of integrity and validation of health data record. In fact, it is used for health data collection, personal health data access and policies access control. A mobile healthcare system is suggested in [18] for personal health data gathering and sharing to assist collaboration between individuals, healthcare providers, and health insurance companies. Another work is realized in [19], where a prototype system for on-demand insurance used smart contract and sensors data, to dynamically modify insurance coverage based on car/environment conditions measured by the implemented sensors, which can help to reduce policy modification costs and limit insurance fraud. Another study has designed a distributed framework based on a blockchain technology to process insurance transactions using smart contracts [20]. The study investigated the use and scalability of smart contracts for automatic execution of processes for insurance companies.

Artificial Intelligence (AI) and machine learning systems have the capability to be integrated into the claims processing, customer service, and fraud detection sub-sectors of the insurance sector. A case study of fraud and premium prediction in automobile insurance was presented in [8]. A data mining-based method was applied to calculate the premium percentage and predict suspicious claims. Three different classification algorithms were applied to predict the likelihood of a fraudulent claim along with the percentage of premium amount: J48, Naive Bayes, and Random Forest. The study presented in [9] employed a fuzzy logic approach by framing fuzzy rules for the machine learning algorithm to improve fraud detection. The latter technique was used for big and high dimensional datasets to predict fraud by using fuzzy logic membership functions. A healthcare insurance fraudster detection method was suggested in [21] in order to detect fraudulent patients. Moreover, detecting fraud in auto insurance based on nearest neighbor models utilized in concert with traditional statistical methods was investigated in [10] where distance-based, density-based, and statistics methods were used to detect fraud occurrence. Accuracy and F-measure performance metrics were employed to evaluate the proposed model. However, this method is not suitable for large datasets and unbalanced data.

Another work suggested an application for motor insurance to predict clients risk levels based on artificial neural network [22]. This study aims to create a prediction model which could be able to evaluate motor insurance clients. A deep-learning based framework was developed in [23] for individuals' payment behavior analysis. In this work, recurrent neural network architecture was proposed in order to analyze individuals' payment history as well as predict their long-term possible social insurance payment behaviors. In [24], an automated deep-learning based architecture was proposed for vehicle damage detection and localization. This work suggests deep transfer and learning techniques for auto insurance companies to detect damage in vehicles, locate them, and classify their severity levels. In addition, Mask R-CNN techniques were applied to visualize the damages in vehicles.

To the best of our knowledge there is no previous work that have suggested an AI and blockchain based solution for auto insurance companies so as to improve their services, increase their benefits, and enhance their systems security levels.

III. PROPOSED BLOCKCHAIN-BASED AND AI-DRIVEN INSURANCE NETWORK ARCHITECTURE

Blockchain and smart contracts which define the rules between different participants in the insurance industry could be used to improve claims processing speed and decrease their costs as well as errors caused by human intervention and inattention compared to manual processing. In this context, smart contracts could encode the rules for the claims processing and enable the refund process from the company. The blockchain facilitates a highly distributed ledger for recording transactions and ordering them in time. We propose a unified architecture, as shown in Fig. 1, where we use permissioned blockchain as a secure network for insurance companies to store and share different information about customers and claims. In this type of blockchain, participants should have permission to partake in the network. As its name applies, permissioned blockchains can be configured to restrict access or permissions to only approved users. Hence, administrators have the right to both control access and verify the validity of transactions. Unlike a decentralized system where achieving consensus could take time, the decision-making process in a private network is more centralized and therefore much faster. In addition, since only trusted nodes are authorized and responsible to manage the data, the network is able to support and process much higher transaction rates. In fact, most of the private blockchains do not allow network branching using distributed consensus algorithms. Instead, they use algorithms without hash competition, such as Practical Byzantine Fault Tolerance (PBFT) consensus algorithm, which was developed to improve upon original BFT consensus mechanisms against Byzantine faults [25]. The PBFT model primarily focuses on providing a practical Byzantine state machine replication that tolerates Byzantine faults through an assumption that there are independent node

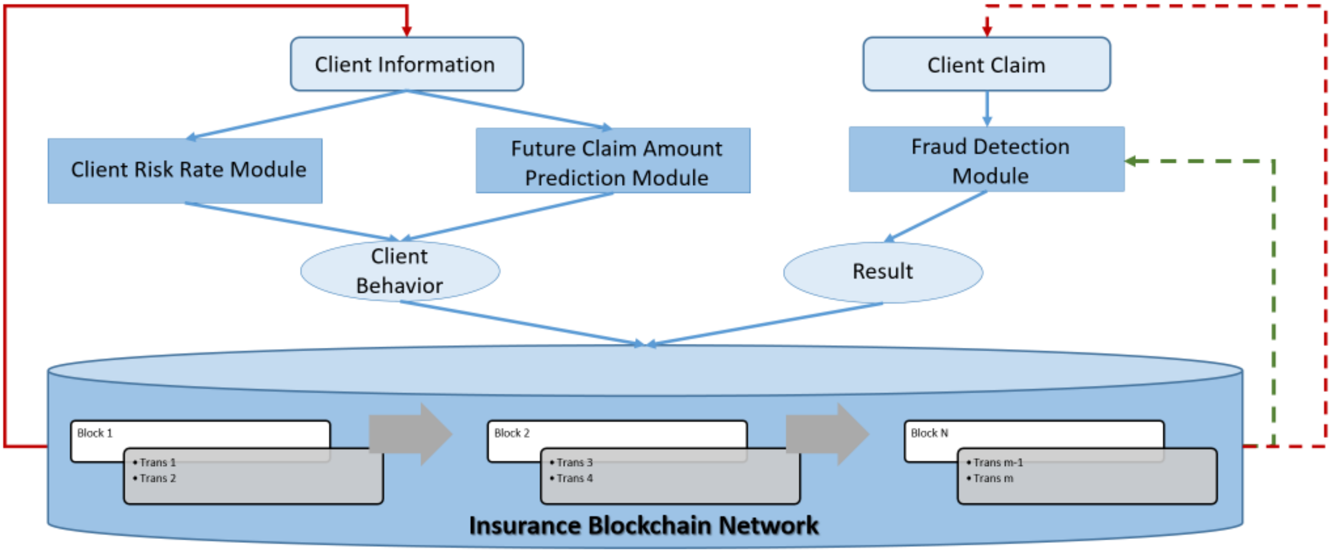


FIGURE 1. Architecture of SISBAR. The solid line refers to the offline learning strategy. The dashed lines indicate the online learning strategy where data is continuously fed to the machine learning model. In this figure, the red color indicates the feeding data for both offline and online machine learning modules and the green one indicates the data, labeled manually, that is fed to the online learning model in order to update its weights and improve its accuracy.

failures and manipulated messages propagated by specific and independent nodes [26]. The algorithm is designed to work in asynchronous systems and provide high-performance with an impressive overhead run-time and only a slight increase in latency [27].

Based on peer-to-peer networks, the system excludes any intermediate or third party controlling the network transaction, which means that no single stakeholder can hack, manipulate, close the chain of blocks, or shut it down. For transactions among insurance company' members, all the data collected from customers and claims are shared and permanently recorded into blocks between all peers taking part of the blockchain network. This will facilitate the employment of online machine learning algorithms for various services such as real-time fraud detection.

Moreover, in the proposed blockchain network, participants have the rights to make decisions, here the trust in the system is not forced but is implicitly guided by user intuition [28]. In fact, the system allows all participants in the blockchain network to accept and verify algorithmically the submitted transactions, e.g., submitting a new claim, before being recorded cryptographically on the block. A block is a record of at least some of the most recent transactions that have not yet been validated by the consensus process. When a block is added to the blockchain network, a digital signature (i.e. a hash value) based on whatever asymetric cryptographic protocol chosen is used to validate whether or not a transaction is recorded. In other words, all clients information and submitted claims must be verified prior to their addition to the blockchain network.

Regarding the digital signature, each user owns a pair of cryptographic keys - a public one and a private one. The private key is used to sign the transactions. The digital signed transactions are spread throughout the whole network and then are accessed by public keys, which are visible to every member of the network. Fig. 2 shows an example of digital signature used in a blockchain. There are two phases in a typical digital signature process: the signing phase and the verification phase. When a client wants to sign a transaction, first they generate a hash value derived from the transaction. Then, they encrypt this hash value by using their private key and send the encrypted hash with the original data to the agency. The agency verifies the received transaction through the comparison between the decrypted hash via the client's public key, and the hash value derived from the received data by the same hash function as the client's. The typical digital signature algorithms used in blockchains include the elliptic curve digital signature algorithm to ensure secure communications between different participants in blockchain network.

It is more advantageous for insurance firms to host a private blockchain over a public blockchain for numerous reasons. Firms that host a private blockchain can easily change the blockchain rules, revert transactions, and modify information. In addition, through reliance on an access control list (ACL) and rules defined in the network, private blockchains can provide a greater level of privacy and security by restricting permissions and access to data shared between different agencies participating in the network. Indeed, contrary to public blockchains where anyone can join the network, private blockchains accept only pre-approved agencies and clients. As a result, only approved participants can submit transactions and share information into the blockchain. Moreover, as mentioned earlier, transactions are faster since they only

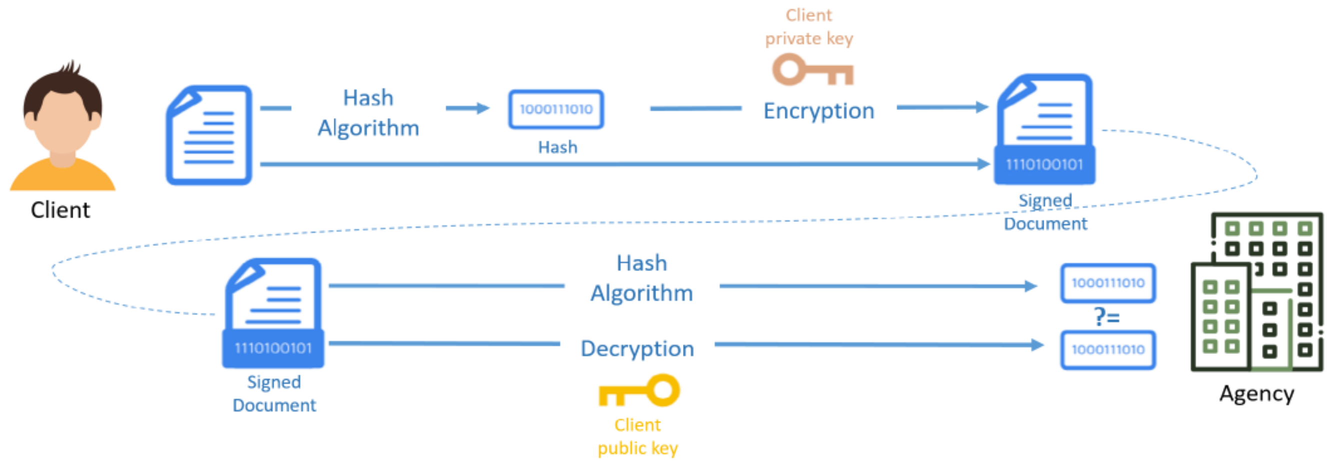


FIGURE 2. Digital signature used in blockchain.

TABLE 1. Generalized features comparison: Public vs. private blockchains.

Features	Public Blockchain	Private Blockchain
New member	Anyone can join	Pre-approved members only
Transaction creation	Anyone can create transactions	Members only
Transaction speed	Low	Fast
Transaction cost	High	Low
Trust	Requires no trust between members	Nodes need to trust each other

rely on a few nodes that are controlled by the firm and contain sufficient processing power. Table 1 highlights the differences between public and private blockchain.

In the proposed architecture shown in Fig. 1, client information and records are collected from the blockchain network and used as features to predict client risk rate and other metrics such as next claim amount. As a result, we can estimate the client's future behavior and vulnerability. Claims are also analyzed and verified by the fraud detection module to detect and classify different fraud types. Submitted claims are then classified and stored into the blockchain network. Since the fraud detection module is based on an online machine learning algorithm, *aka* incremental machine learning algorithm, this model can be updated and upgraded dynamically without retraining the model with the entire dataset in each update. Extracted claims from the shared ledgers can then be manually verified. Afterward, the verified data may be utilized as training data for future iterations of the classifier and improve its accuracy.

In the development of this architecture, we focus on the core entities for auto insurance, clients and claims. We assume that our network is maintained by a collaborating set of auto insurance firms, where client data is all recorded on the shared blockchain network. The proposed system is capable of including an arbitrary set of collaborating insurance firms that can share fraudulent claim data, with the aim of improving their fraud detection and pricing models. This collaborative behavior can minimize collective loss for all of the firms, while maintaining confidentiality of the information of the clients for each firm. Essentially, this system could serve as the basis for collaboration in the sector while minimizing the risk of de-anonymizing their client information, and also maintaining some intellectual property control over their data by only revealing relevant features to the problem. Moreover, since the proposed architecture is modular and scalable, other modules can be integrated into this framework to improve the system functionality.

To summarize, blockchain technology brings with its innovative principles of sharing, verifying, and securing the data between different nodes participating in the network new features that can improve the insurance industry. Added to that, pairing blockchain technologies and AI exhibit big potentials to transform the global insurance industry. In fact, the recent revolution of those technologies and their fast adoption in various sectors allow to rethink about rescheduling traditional processes and integrating those technologies in one solid architecture for the insurance companies and the insurance industry in general.

IV. MACHINE LEARNING-DRIVEN FRAUD DETECTION AND RISK MODELLING

In this section, we first introduce the methodology to detect fraudulent claims, predict client risk, and future behavior. Afterwards, we investigate different learning strategies used to build AI models. Finally, we present the different proposed predictors.

A. METHODOLOGY

The proposed methodology aims to clean, explore, and preserve data privacy using data mining techniques. The workflow, presented in Fig. 3, summarizes this methodology.

1) DATA CLEANSING

Data cleansing or cleaning is the process of detecting, rectifying, and/or removing inaccurate and corrupted information from the dataset or database. More precisely, it encompasses the process of recognizing incomplete data, filling missing

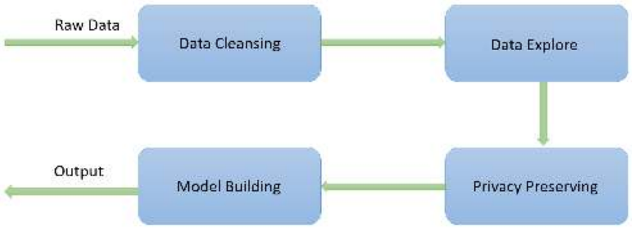


FIGURE 3. Proposed machine learning module pipeline.

values, and removing incorrigible data. Unclean data is typically generated by human, transmission, processing, or storage errors. The data cleaning process provides benefits by reducing the computational expenditures required to train models, improves dataset quality, and expedites the data exploration and feature engineering processes. To handle samples with minor errors, we estimate the missing values using statistical and interpolation techniques as well as running predictive models that impute the missing values. Imputing missing values can be effective for minor errors because the values were originally missing but those techniques lead to an information loss, no matter how sophisticated the imputation method is. For this reason, samples with major missing values are often dropped. In our study, we remove all duplicated samples in addition to those with major errors or missing important information in order to maintain only samples that bring reliable information.

2) DATA EXPLORING

Exploratory data analysis is a way of analyzing and summarizing the main information of the dataset using statistical and visualization methods. In order to have a global view of the dataset and extract the most important features for the machine learning model, we analyze the frequency of features and calculate the correlation between them. Fig. 4 illustrates the correlation matrix of relevant features in the dataset that we used in Section IV. The correlation matrix is a powerful tool that helps with the removal of features that do not contribute much information to the model.

3) PRIVACY PRESERVING

In order to preserve client's privacy, we employ categorization and generalization to anonymize the data. With categorization, most of the attributes of client data and claims are transformed to a binary format. Otherwise, personal information and general attributes are converted into other formats. With generalization, low-level data are replaced to high-level concepts. For instance, customer personal information will not be shown or used in their row formats.

4) MODEL BUILDING

Fraud detection can be framed as a classification problem, where each discrete category corresponds to a distinct type of fraud, or to 'no fraud' detected. Multi-class classification is required due to the multiple methods of fraud that exist in practice. Moreover, the prediction of future client claim amounts based on their riskiness can be modelled as a regression problem.

B. OFFLINE LEARNING: EXTREME GRADIENT BOOSTING ALGORITHM

In the literature, most machine learning algorithms are batch-based. It is a static method of model training, where the input data is fed in one batch to train and build a strong model. In the batch process, after the training phase ceases, the model's hyperparameters are fixed, and a portion of the dataset not used to train the model is used to validate the effectiveness of the model and test its general predictive ability.

XGBoost is one of the most efficient implementations of gradient boosted decision trees and it has been selected as one of the best offline machine learning algorithms used in Kaggle competitions [7], [29]. Specifically designed to optimize memory usage and exploit the hardware computing power, XGBoost decreases the execution time with an increased performance compared to many machine learning algorithms. The main idea of boosting is to sequentially build sub-trees from an original tree such that each subsequent tree reduces the errors of the previous one. In such a way, the new sub-trees will update the previous residuals in order to reduce the error of the cost function.

Consider the following dataset D:

$$D = \{(x_i, y_i) \text{ where } x_i \in \mathbb{R}^m, y_i \in \mathbb{R})\}, \tag{1}$$

where m is the number of features in x_i and y_i is the ground-truth of the sample i. We denote by n the number of samples such that $|D| = n$ where $|.|$ denotes the cardinality of a set (i.e. the number of rows in the dataset). We define the predicted value of the entry i, $\hat{y}_i$, as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), \quad f_k \in F, \tag{2}$$

where f_k indicates an independent tree in F, the space of regression trees, and $f_k(x_i)$ refers to the predicted score given by the i-th sample and k-th tree. The XGBoost cost function $\mathcal{L}$ can be expressed as follows:

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right). \tag{3}$$

The training loss function $l\left(y_i, \hat{y}_i\right)$ evaluates the difference between prediction $\hat{y}_i$ and the actual value y_i while $\Omega(\dot{})$ is the regularization factor of the cost function; the aim of this factor is to reduce the risk of the model overfitting to the data. The factor $\Omega(\dot{})$ can be expressed as follows:

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \left\|w\right\|^2, \tag{4}$$

where γ and λ are two regularization parameters and T and w are the numbers of leaves and the scores of each leaf, respectively. By minimizing the objective function $\mathcal{L}$, the regression tree model functions f_k can be learned. A second degree

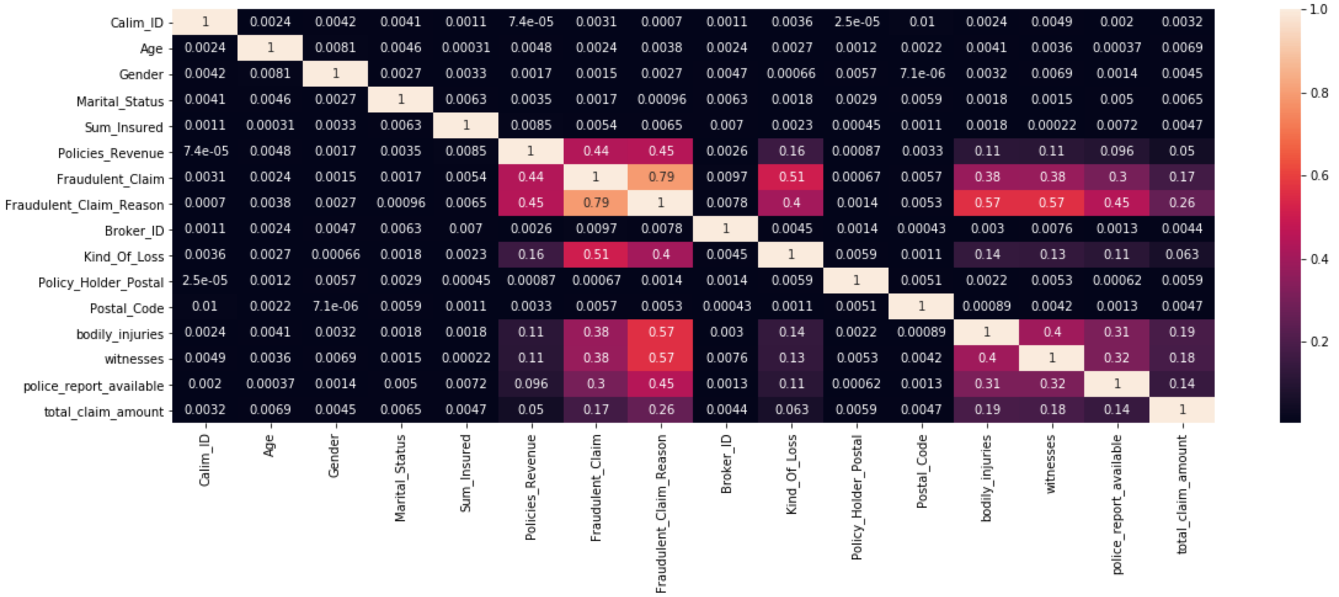


FIGURE 4. Feature correlation matrix. Lighter colors demonstrate higher correlations between two features.

Algorithm 1 Exact Greedy Algorithm for Split Finding

```
Input: Instance set of current node $I$
$score \leftarrow 0$
$G \leftarrow \sum_{i\in I} g_i, H \leftarrow \sum_{i\in I} h_i$
for $k \in \{1, \ldots, m\}$ do
    $G_L \leftarrow 0, H_L \leftarrow 0;$
    for $j \in sorted(I, by\ feature\ values)$ do
        $G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j;$
        $G_R \leftarrow G - G_L, H_R \leftarrow H - H_L;$
        $score \leftarrow max(score, \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G^2}{H+\lambda});$
    end
end
Output: Split with max score
```

Taylor series can be used to approximate the objective function [30]. Let $I_j = \{i|q(x_i) = j\}$ an instance set of leaf j with $q(x)$ a fixed structure. The optimal weights w_j* of leaf j and the corresponding optimal value can be obtained by the following equations:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}, \quad \text{and } \mathcal{L}^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T, \tag{5}$$

where g_i and h_i are the first and the second gradient orders of the loss function $\mathcal{L}$.

The loss function $\mathcal{L}$ can be used as a quality score of the tree structure q. The smaller the score is, the better the model is. As it is not possible to enumerate all the tree structures, a greedy algorithm can solve the problem by starting from a single leaf and iteratively adding branches to the tree. The exact greedy algorithm is given in Algorithm 1. Let I_R and I_L denote the instance sets of right and left nodes after split. Assuming $I = I_R \cup I_L$, the loss reduction will be shown as follows [7]:

$$\mathcal{L}_{\text{split}} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{6}$$

This formula is usually used in practice for evaluating the split candidates. The XGBoost model uses many simple trees and score leaf nodes during splitting. The first three terms of the equation represent the score of the left, right, and original leaf, respectively. In addition, the term γ is a regularization parameter applied to the additional leaf and it is used in the training process to reduce the overfitting problem.

C. ONLINE LEARNING: VERY FAST DECISION TREE ALGORITHM

Unlike the batch machine learning approach, incremental learning algorithms allow not only to train dynamically the data but also to update the model without the need to retrain with the whole dataset. The algorithm updates its parameters after each training instance. In addition, online learning is a common technique used in machine learning fields as it is computationally unfeasible to train over the entire dataset. The training dataset can be split into mini batches and trained one by one since the model supports dynamic sequence learning. It is also used in situations where it is necessary for the algorithm to dynamically adapt to new patterns in the data, or when the data itself is generated as a function of time especially for applications with the insurance industry

Algorithm 2 Hoeffding Tree Algorithm

```
Input: $S$ is the sequence of examples $\mathbf{X}$ is a set of
       discrete attributes $G(.)$ is the heuristic evaluation
       function $\delta$ equal to one minus the required
       probability to select the right attribute at a given
       node
Let $X_1 = \mathbf{X} \cup X_0$
Let $HT$ be a tree a single leaf $l_1$ (root node)
Let $\bar{G}_1$ be the $\bar{G}$ obtained by predicting the most
  frequent fraud class in $S$
for each fraud class $y_k$ do
    for each value $x_{ij}$ of each attribute $X_i \in \mathbf{X}$ do
        Let $\mu(l_1) = 0$ ;
    end
end
for each example $(\mathbf{x}, y_k)$ in $S$ do
    Sort $(\mathbf{x}, y)$ into a leaf $l$ using $HT$ ;
    for each $x_{ij}$ in $\mathbf{x}$ such that $X_i \in X_l$ do
        Increment $\mu(l)$ ;
    end
    Label $l$ with the majority class among the examples
      seen so far at $l$
    if the examples seen so far at $l$ are not all of the
      same class then
        Compute $\bar{G}_l(X_i)$ for each attribute $X_i \in$
          $X_l - \{X_0\}$ using the counts $\mu(l)$ ;
        Let $X_1$ be the attribute with the highest $\bar{G}_l$ ;
        Let $X_2$ be the attribute with the second-highest
          $\bar{G}_l$ ;
        Compute $\epsilon$ using Eq (7) ;
        if $\bar{G}_l(X_1) - \bar{G}_l(X_2) > \epsilon$ and $X_1 \neq X_0$ then
            Replace $l$ by an internal node that splits on
              $X_1$ ;
            for each branch of split do
                Add a new leaf $l_m$ and let $X_m = \mathbf{X}$
                  $-\{X_1\}$ ;
                Let $\bar{G}_m(X_0)$ be the $\bar{G}$ obtained by
                  predicting the most frequent class at $l_m$ ;
                for each class $y_k$ and each value $x_{ij}$ of
                  each attribute $X_i \in X_m - \{X_0\}$ do
                    Let $\mu(l_m) = 0$
                end
            end
        end
    end
end
Output: $HT$ the Hoeffding Tree
```

TABLE 2. Performance of the future claim amount prediction module.

Predictor	Error ($)
ElasticNet	402
Gradient Boosting	162
Ridge	401
XGBoost	**137**

where data is added to the blockchain network continuously by different participants.

VFDT or Hoeffding Tree is an online machine learning algorithm based on decision trees, designed around the principles of the Hoeffding Bound (HB) [31]. The HB supposes that we have M independent random variables $r_1, \ldots, r_M$, with range R and mean $\hat{r}$. The HB states that with a probability $1 - \delta$, the true mean, is at least $\hat{r} - \varepsilon$ where

$$\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2M}}. \tag{7}$$

The pseudo-code of the VFDT algorithm is given in Algorithm 2. The algorithm applies the HB to evaluate if a given leaf should be split during the training phase according to a gain ratio $G(.)$ of all attributes. Let us assume that G to be maximized. If X_1 is the attribute with the highest gain ratio $\bar{G}$ after seeing n example, X_2 is the attribute with the second highest ratio, and $\Delta\bar{G} = \bar{G}(X_1) - \bar{G}(X_2)$ is the difference of information gains. If $\Delta G > \epsilon$ then the HB ensures that the true $\Delta G \geq \Delta\bar{G} - \epsilon > 0$ with probability $1 - \delta$ and X_1 is the best attribute to split with the same probability. This assumption is valid as long as the value of $\bar{G}$ for a given node can be seen as an average of G values for the examples of that node. Therefore, nodes need to accumulate examples until $\Delta\bar{G} - \epsilon > 0$. At this point, nodes can be split using the current best attribute and next examples will be passed to the new leaves. Counts μ are the sufficient statistics needed so as to estimate most heuristic measures. Also, pre-pruning is carried out by considering at each node a 'zero' attribute X_0 that consists of not splitting node. As a result, the split will be made only if the best split found is better than the remaining the same actual state.

Assuming l is the number of leaves in HT, the run time of the tree based algorithm is sub-lineal in regard to the model size $\mathcal{O}(\log l)$ which makes them extremely fast and efficient compared to other incremental machine learning algorithms with a linear relation between the model complexity and the running time [32]. As this algorithm shows a high performance in different topics such as massive data [33] and dynamic and streaming problem resolution [34], we will investigate its performance against fraudulent claims detection and classification.

V. EXPERIMENTAL RESULTS AND IMPLEMENTATION

In this section, we start by discussing the efficiency of the XGBoost algorithm to predict risky clients and their future claims. Then, we evaluate the performance of the online and offline machine learning classifiers to detect and classify different types of fraud. Finally, we present selected results of the implementation of the blockchain framework.

A. FRAUD DETECTION AND RISK MEASUREMENT

For risk and claim prediction models, we train, validate, and test the proposed models using dataset obtained from a real

TABLE 3. Performance measurement obtained for fraud detection Dataset.

Fraud type	$T(0)$	$T(1)$	$T(2)$	$T(1+2)$	$T(3)$	$T(1+3)$	$T(2+3)$	$T(1+2+3)$
Invalid kind of loss	No Fraud	X		X		X		X
No premium but has claim	No Fraud		X	X			X	X
Fraudulent claim amount	No Fraud				X	X	X	X

TABLE 4. Confusion matrix table.

PredictedObserved	True	False
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

TABLE 5. Performance of the fraud detection module.

Classifier	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**Training Time (ms)**
Decision Tree	92.99	0.870	0.929	0.892	471
Naive Bayes	52.06	0.373	0.520	0.425	**155**
Nearest Neighbor	42.70	0.223	0.427	0.255	1254
XGBoost	**99.25**	**0.9928**	**0.992**	**0.9926**	995

TABLE 6. Performance of the client risk rate module.

Classifier	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
Decision Tree	74,44	0.6473	0.5953	0.6005
SVM	73.21	0.6696	0.5652	0.4841
Nearest Neighbor	73.80	0.6696	0.5256	0.4841
XGBoost	**76,81**	**0.6828**	**0.6295**	**0.6392**

insurance company. Since the future claim amount is a regression problem, we use different regression machine learning algorithms to solve it. In literature, elasticNet [35], gradient boosting [36], and ridge [37] algorithms are generally used to solve regression problems. However, in this paper, we propose to use the XGBoost regression algorithm and compare its performances with the ones of the aforementioned regression algorithms. There are several metrics that can be used to measure accuracy for continuous variables and evaluate the model performance. However, the mean absolute error, *aka* MAE, is one of the most efficient metrics for summarizing and assessing the quality of a machine learning model. Its expression is given as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{8}$$

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The regression performance obtained for the ridge, elasticNet, gradient boosting, and XGBoost regressors are shown in Table 2. It should be noted that all the displayed results in the simulation result section are averaged over 50 testing iterations where, at each iteration, different testing sets are evaluated.

For fraud detection, a dataset with more than 64 thousand claims are used to train, validate, and test the classifier. Eight classes are considered and detailed in Table 3 where $T(0)$ refers to non-fraud claims, $T(1)$, $T(2)$, and $T(3)$ represent the three types of auto fraud claims, specifically "Invalid kind of loss", "No premium but has claimed", and "Fraudulent claim amount". The classes $T(1+2)$, $T(2+3)$, $T(1+3)$, $T(1+2+3)$ are obtained from different combinations of the three types of fraud.

The XGBoost performances are compared to those of three other classifiers used in literature for fraud detection in insurance applications, namely the decision tree [38], the naive bayes [8], and the nearest neighbor [10] algorithms. To tune our machine learning models, we use different hyperparameters. For the Decision Tree model, we apply the Gini function to measure the quality of split and we fix the maximum depth of trees to eight. As for the Naive Bayes, the likelihood of features is assumed to be Gaussian. Also, for the Nearest Neighbor, we set the number of neighbors to the same number of fraud classes and we apply a uniform weight distribution for all points in each neighborhood. We provide all the classifiers the same data and we fixed the ratio to 0.3 for the training and testing sets. We use the confusion matrix shown in Table 4 to compute the performance of the classifiers [39]. This matrix provides the values of the following metrics such as accuracy, precision, recall, the F1-score expressed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{11}$$

$$\text{F1-score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

The classification performances obtained for the naive bayes, nearest neighbor, decision tree, and XGBoost classifiers are shown in Table 5 and Table 6.

Table 5 shows that XGBoost provides the best results based on the metrics defined in equations (9) through (12), although it was the second-slowest converging model that was tested. The performance of the decision tree algorithm is better than the ones of naive bayes and nearest neighbor algorithms. However, XGBoost achieves the best results for our data.

As for the effectiveness of the proposed incremental learning algorithm in detecting fraudulent claims, we perform an experiment where at each time, we update the model with

TABLE 7. Comparison of the VFDT performance to those of the SGD for the fraud detection module.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VFDT	98.2	0.982	0.982	0.981
SGD	64.7	0.532	0.642	0.551

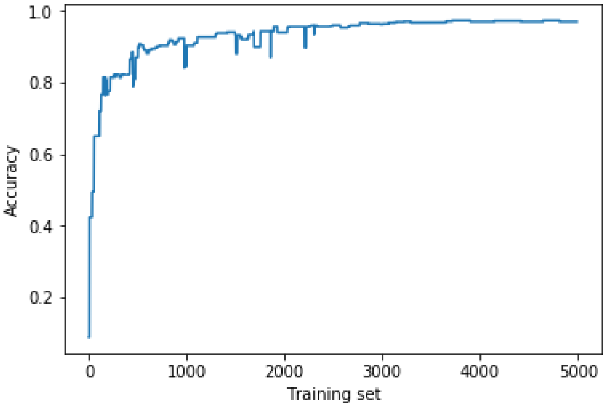


FIGURE 5. Performance of the VFDT algorithm: accuracy (%) vs. training set size.

new data and evaluate its performance. As mentioned earlier, to train the online model, data can be either fed in one block or by sequences. The model will dynamically update its weights and increase its performance. We compare the efficiency of the VFDT to the one of a state-of-the-art incremental machine learning algorithm, namely SGD with linear SVM loss function in detecting fraudulent claims. Table 7 shows that the VFDT significantly outperform the incremental SGD algorithm. This result can be explained by the fact that the tree-based models are eminently efficient in high-dimensional spaces due to their compressing representation where split nodes are only added when it is necessary for the classification of the data seen so far. The opposite is true for distance-based methods along one or two dimensions. The obtained results corroborates the ones obtained in [32] where it is shown that VFDT achieves significantly better results in high dimensional space detection for many datasets.

Fig. 5 demonstrates the evolution of the model accuracy as more data is used to train the data. We see that the accuracy grows as more training data is added to the model, however the rate of improvement decreases as the training set grows. The model reaches 90% accuracy within 300 samples passed into the training algorithm, and takes nearly 5000 more samples to reach 98% accuracy, showing that the VFDT algorithm converges to a stable solution in a short period.

The notable performance of the proposed machine learning algorithms can be validated by the confusion matrices provided in Fig. 6 and Fig. 7, which summarize the normalized predicted values of each class of fraud using the XGBoost and VFDT. We can notice that even with unbalanced dataset fraudulent and non-fraudulent claims detected very accurately. The obtained results can be explained by the fact that XGBoost and decision tree algorithms can find statistical relationships between input features and output results. Hence, they can pick and assign optimized weights for the best features of the input data. Moreover, they are flexible and perform well with a large number of samples. Unlike the tree-based methods, the naive bayes and nearest neighbor algorithms behave quite well with big datasets.

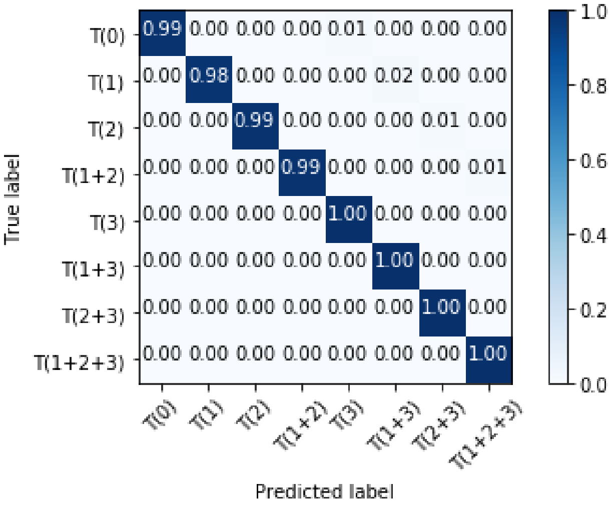


FIGURE 6. XGBoost normalized confusion matrix.

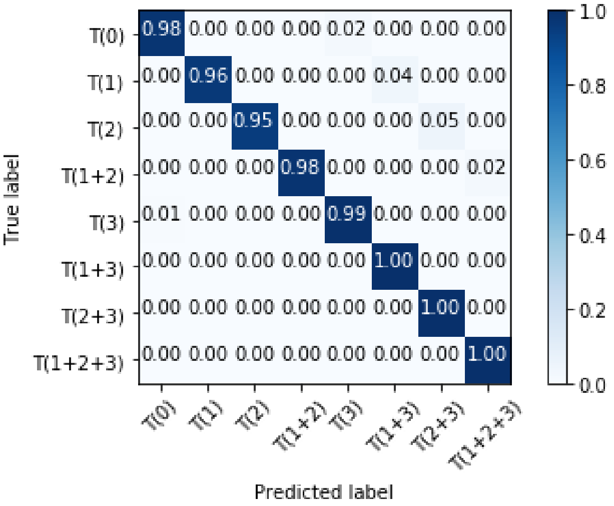
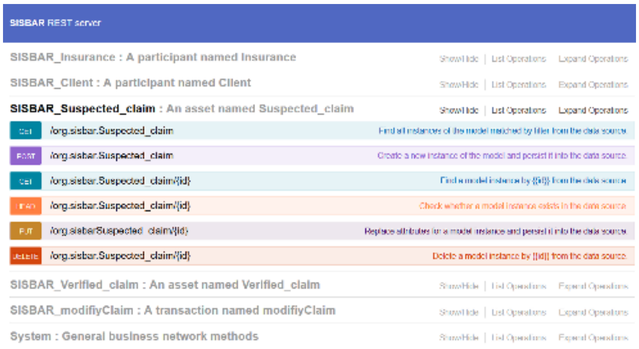


FIGURE 7. VFDT normalized confusion matrix.

B. BLOCKCHAIN IMPLEMENTATION

In this work, we propose the use of Hyperledger Fabric [40]. It is an existing permissioned blockchain implementation that has unique properties that make it well-suited for enterprise-class applications. In fact, the Fabric network consists of different types of entities, peer nodes, ordering service nodes, and clients belonging to different organizations. Each one of them has an identity on the network which is provided by a membership service provider (MSP), typically associated with an organization and all entities in the network have



SISBAR REST server

SISBAR_Insurance : A participant named Insurance
Show/Hide | List Operations | Expand Operations

SISBAR_Client : A participant named Client
Show/Hide | List Operations | Expand Operations

SISBAR_Suspected_claim : An asset named Suspected_claim
Show/Hide | List Operations | Expand Operations

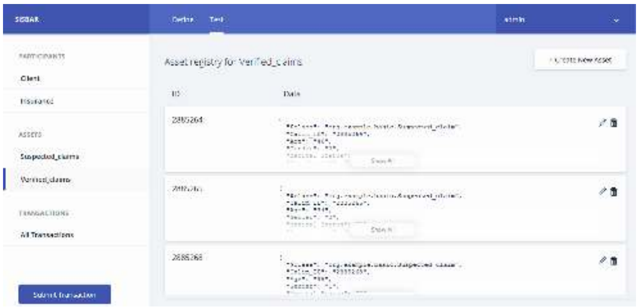
GET	/org.sisbar.Suspected_claim	Find all instances of the model matched by filter from the data source
POST	/org.sisbar.Suspected_claim	Create a new instance of the model and persist it into the data source
GET	/org.sisbar.Suspected_claim/{id}	Find a model instance by {{id}} from the data source
HEAD	/org.sisbar.Suspected_claim/{id}	Check whether a model instance exists in the data source
PUT	/org.sisbarSuspected_claim/{id}	Replace attributes for a model instance and persist it into the data source
DELETE	/org.sisbar.Suspected_claim/{id}	Delete a model instance by {{id}} from the data source

SISBAR_Verified_claim : An asset named Verified_claim
Show/Hide | List Operations | Expand Operations

SISBAR_modifiyClaim : A transaction named modifiyClaim
Show/Hide | List Operations | Expand Operations

System : General business network methods
Show/Hide | List Operations | Expand Operations

FIGURE 8. SISBAR REST server interface.



Asset registry for Verified_claims

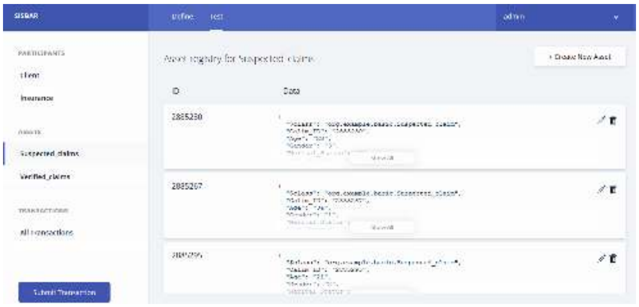
ID	Data
2885264	[illegible]
[illegible]	[illegible]
2885268	[illegible]

FIGURE 9. SISBAR verified claims interface.

the ability to identify all organizations and hence, eases verification. In this context, insurance companies present the organization and their agencies refer to entities participating in the network and associated to those organizations. Claims present assets in the blockcahin network where each submission or modification of any claim refers to a transaction which will be verified and validated before being added into blockchain.

Blockchain applications and simulations are developed using Hyperledger Composer module for Hyperledger Fabric framework. Fig. 8 shows an interface of the REST server used to ensure the communication between the AI and blockchain server through the REST API. In Fig. 8, we can distinguish the services and APIs used to interrogate and communicate with the REST server to enable the interaction between SISBAR servers and modules. Fig. 9 and Fig. 10 show the verified claims and the suspected claims saved on the blockchain network respectively. For example, claim 2885264, a verified claim that before being saved in the blockchain, has been automatically verified by the fraud detection module and successfully passed the test. However, the claim 2885230 was been identified as a suspicious claim by the AI module and has been saved under suspected claims for further investigation.

Saved claims as shown in Fig. 9 and Fig. 10 are accessed by authorised parties in order to check them manually and reintegrate those claims into the network. Moreover, after the manual check, submitted claims can be reused and fed into the online machine learning model so as to update its weights and improve its performance in detecting fraudulent claims.



Asset registry for Suspected claims

ID	Data
2885230	[illegible]
2885267	[illegible]
[illegible]	[illegible]

FIGURE 10. SISBAR suspected claims interface.

VI. CONCLUSION

In this study, we have developed SISBAR, a novel fraud detection system for insurance firms based on permissioned blockchain and machine learning algorithms. We selected two learning strategies for detection and classification of fraudulent claims submissions out of a pool of learning techniques based on experimental performance on a real insurance firm's data. We have also investigated the use of XGBoost and VFDT algorithms for batch and incremental learning strategies to detect and classify different types of fraudulent auto insurance claims and measure risk level of customers. The performances of the proposed algorithms are compared to those of other state-of-the-art solutions. It is shown that the proposed classifiers ensure not only the best accuracy in detecting fraudulent claims but also can classify different types of fraud for insurance unlike the existing solutions. Moreover, XGBoost proved its efficiency in predicting customers' future behavior and their future claims amount. An implementation of the blockchain, AI modules, and essential structural nodes were developed to perform tests and simulations on different actors of the proposed framework. SISBAR presents a basis for insurance companies to decrease their claim refund losses, improve their performance, and their competitiveness. This in turn leads to savings for the insurance clients that act lawfully. As future work, we will focus on enhancing the proposed architecture and implementing AI solutions that are tailored to other insurance services.

ACKOWLEDGMENT

This article was presented at the IEEE International Conference on Vehicular Electronics and Safety (ICVES) 2019, Cairo, Egypt, September 2019.

REFERENCES

[1] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "Extreme gradient boosting machine learning algorithm for safe auto insurance operations," in *Proc. IEEE Int. Conf. Veh. Electron. Saf. (ICVES)*, Cairo, Egypt, Sep. 2019, pp. 1–5.

[2] *Insurance Fraud Handbook*, Association of Certified Fraud Examiners, Austin, TX, USA, Oct. 2018.

[3] D. Corum, "Insurance research council finds that fraud and buildup add up to $7.7 billion in excess payments for auto injury claims," Insurance Res. Council, Malvern, PA, USA, Tech. Rep., Feb. 2015. [Online]. Available: https://www.insurance-research.org/sites/default/files/downloads/IRC%20Fraud%20News%20Release.pdf

[4] S. Wang, J. Wang, X. Wang, T. Qiu, Y. Yuan, L. Ouyang, Y. Guo, and F.-Y. Wang, ''Blockchain-powered parallel healthcare systems based on the ACP approach,'' *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 942–950, Dec. 2018.

[5] A. Pieroni, N. Scarpato, L. Di Nunzio, F. Fallucchi, and M. Raso, ''Smarter city: Smart energy grid based on blockchain technology,'' *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 8, no. 1, pp. 298–306, 2018.

[6] M. Samaniego and R. Deters, ''Internet of smart Things–IoST: Using blockchain and clips to make things autonomous,'' in *Proc. IEEE Int. Conf. Cognit. Comput. (ICCC)*, Honolulu, HI, USA, Jun. 2017, pp. 9–16.

[7] T. Chen and C. Guestrin, ''XGBoost: A scalable tree boosting system,'' in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining, (SIGKDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[8] G. Kowshalya and M. Nandhini, ''Predicting fraudulent claims in automobile insurance,'' in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Coimbatore, India, Apr. 2018, pp. 1338–1343.

[9] K. Supraja and S. J. Saritha, ''Robust fuzzy rule based technique to detect frauds in vehicle insurance,'' in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS)*, Chennai, India, Aug. 2017, pp. 3734–3739.

[10] T. Badriyah, L. Rahmaniah, and I. Syarif, ''Nearest neighbour and statistics method based for detecting fraud in auto insurance,'' in *Proc. Int. Conf. Appl. Eng. (ICAE)*, Batam, Indonesia, Oct. 2018, pp. 1–5.

[11] J.-M. Long, Z.-F. Yan, Y.-L. Shen, W.-J. Liu, and Q.-Y. Wei, ''Detection of epilepsy using MFCC-based feature and XGBoost,'' in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Beijing, China, Oct. 2018, pp. 1–4.

[12] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu, and J. Peng, ''XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud,'' in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Shanghai, China, Jan. 2018, pp. 251–256.

[13] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang, ''Untangling blockchain: A data processing view of blockchain systems,'' *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1366–1385, Jul. 2018.

[14] K. Kim, Y. You, M. Park, and K. Lee, ''DDoS mitigation: Decentralized CDN using private blockchain,'' in *Proc. 10th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Prague, Czech, Jul. 2018, pp. 693–696.

[15] S. Chakraborty, S. Aich, and H.-C. Kim, ''A secure healthcare system design framework using blockchain technology,'' in *Proc. 21st Int. Conf. Adv. Commun. Technol. (ICACT)*, PyeongChang Kwangwoon_Do, South Korea, Feb. 2019, pp. 260–264.

[16] V. Gatteschi, F. Lamberti, C. Demartini, C. Pranteda, and V. Santamaria, ''To blockchain or not to blockchain: That is the question,'' *IT Prof.*, vol. 20, no. 2, pp. 62–74, Apr. 2018.

[17] F. Lamberti, V. Gatteschi, C. Demartini, C. Pranteda, and V. Santamaria, ''Blockchain or not blockchain, that is the question of the insurance and other sectors,'' *IT Prof.*, early access, Jun. 16, 2017, doi: 10.1109/MITP.2017.265110355.

[18] X. Liang, J. Zhao, S. Shetty, J. Liu, and D. Li, ''Integrating blockchain for data sharing and collaboration in mobile healthcare applications,'' in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[19] F. Lamberti, V. Gatteschi, C. Demartini, M. Pelissier, A. Gomez, and V. Santamaria, ''Blockchains can work for car insurance: Using smart contracts and sensors to provide on-demand coverage,'' *IEEE Consum. Electron. Mag.*, vol. 7, no. 4, pp. 72–81, Jul. 2018.

[20] M. Raikwar, S. Mazumdar, S. Ruj, S. S. Gupta, A. Chattopadhyay, and K.-Y. Lam, ''A blockchain framework for insurance processes,'' in *Proc. 9th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS)*, Paris, France, Feb. 2018, pp. 1–4.

[21] C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang, and W. Guo, ''Patient cluster divergence based healthcare insurance fraudster detection,'' *IEEE Access*, vol. 7, pp. 14162–14170, Dec. 2019.

[22] K. Vassiljeva, A. Tepljakov, E. Petlenkov, and E. Netsajev, ''Computational intelligence approach for estimation of vehicle insurance risk level,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 4073–4078.

[23] D. Muller and Y.-F. Te, ''Insurance premium optimization using motor insurance policies—A business growth classification approach,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4154–4158.

[24] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, ''A very deep transfer learning model for vehicle damage detection and localization,'' in *Proc. 31st Int. Conf. Microelectron. (ICM)*, Cairo, Egypt, Dec. 2019, pp. 158–161.

[25] M. Castro and B. Liskov, ''Practical Byzantine fault tolerance,'' in *Proc. OSDI*, vol. 99, 1999, pp. 173–186.

[26] K. Lei, Q. Zhang, L. Xu, and Z. Qi, ''Reputation-based Byzantine fault-tolerance for consortium blockchain,'' in *Proc. IEEE 24th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Singapore, Dec. 2018, pp. 604–611.

[27] J. Sousa, A. Bessani, and M. Vukolic, ''A Byzantine fault-tolerant ordering service for the hyperledger fabric blockchain platform,'' in *Proc. IEEE Int. Conf. Dependable Syst. Netw. (DSN)*, Luxembourg City, Luxembourg, Jun. 2018, pp. 51–58.

[28] F. Tang, S. Ma, Y. Xiang, and C. Lin, ''An efficient authentication scheme for blockchain-based electronic health records,'' *IEEE Access*, vol. 7, pp. 41678–41689, Mar. 2019.

[29] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and R. Liu, ''XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system,'' *IEEE Access*, vol. 7, pp. 13149–13158, Jan. 2019.

[30] J. West and M. Bhattacharya, ''Intelligent financial fraud detection: A comprehensive review,'' *Comput. Secur.*, vol. 57, pp. 47–66, Mar. 2016.

[31] V. G. T. da Costa, A. C. P. de Leon Ferreira de Carvalho Carvalho, and S. Barbon, Jr., ''Strict very fast decision tree: A memory conservative algorithm for data stream mining,'' *Pattern Recognit. Lett.*, vol. 116, pp. 22–28, Dec. 2018.

[32] V. Losing, B. Hammer, and H. Wersing, ''Incremental on-line learning: A review and comparison of state of the art algorithms,'' *Neurocomputing*, vol. 275, pp. 1261–1274, Jan. 2018.

[33] S. Desai, S. Roy, B. Patel, S. Purandare, and M. Kucheria, ''Very fast decision tree (VFDT) algorithm on Hadoop,'' in *Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Pune, India, Aug. 2016, pp. 1–7.

[34] B. Raahemi, W. Zhong, and J. Liu, ''Peer-to-Peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree,'' in *Proc. 20th IEEE Int. Conf. Tools with Artif. Intell.*, Dayton, OH, USA, Nov. 2008, pp. 525–532.

[35] H. Zou and T. Hastie, ''Regularization and variable selection via the elastic net,'' *J. Roy. Stat. Soc., Ser. B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[36] J. H. Friedman, ''Stochastic gradient boosting,'' *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[37] A. E. Hoerl and R. W. Kennard, ''Ridge regression: Biased estimation for nonorthogonal problems,'' *Technometrics*, vol. 12, no. 1, pp. 55–67, Apr. 1970.

[38] R. Roy and K. T. George, ''Detecting insurance claims fraud using machine learning techniques,'' in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Kollam, India, Apr. 2017, pp. 1–6.

[39] Y. A. Alshehri, K. Goseva-Popstojanova, D. G. Dzielski, and T. Devine, ''Applying machine learning to predict software fault proneness using change metrics, static code metrics, and a combination of them,'' in *Proc. SoutheastCon*, St. Petersburg, FL, USA, Apr. 2018, pp. 1–7.

[40] H. Sukhwani, J. M. Martinez, X. Chang, K. S. Trivedi, and A. Rindos, ''Performance modeling of PBFT consensus process for permissioned blockchain network (hyperledger fabric),'' in *Proc. IEEE 36th Symp. Reliable Distrib. Syst. (SRDS)*, Hong Kong, Sep. 2017, pp. 253–255.

NAJMEDDINE DHIEB (Student Member, IEEE) received the Diplôme d'Ingénieur degree (Hons.) in telecommunication engineering from the École Supérieure des Communications de Tunis (SUP'COM), Tunis, Tunisia, in 2019. He is currently working as a Research Assistant with the Smart City Lab, School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, USA. His general research interests are at the intersection of blockchain, data mining, artificial intelligence, image processing, and the Internet-of-Things.

HAKIM GHAZZAI (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from KAUST, Saudi Arabia, in 2015, and the Diplome d'Ingenieur degree (Hons.) in telecommunication engineering and the master's degree in high-rate transmission systems from the Ecole Superieure des Communications de Tunis (SUP'COM), Tunis, Tunisia, in 2010 and 2011, respectively. He was a Visiting Researcher with Karlstad University, Sweden, and a Research Scientist with the Qatar Mobility Innovations Center (QMIC), Doha, Qatar, from 2015 to 2018. He is currently a Research Scientist with the Stevens Institute of Technology, Hoboken, NJ, USA. His general research interests include the intersection of wireless networks, UAVs, the Internet of Things, and intelligent transportation systems. Since 2019, he has been on the Editorial Board of the IEEE Communications Letters and the IEEE Open Journal of the Communications Society.

HICHEM BESBES received the Ph.D. degree in electrical engineering from ENIT, University EL MANAR, Tunisia, in 1999, the Diplome d'Etudes Appronfondies (DEA) (master's diploma) degree systems theory from ENIT, in 1991, an Engineering Diploma degree in electrical engineering from ENIT, in 1991, and the Habilitation à Diriger des Recherches (HDR) degree in telecommunications from the Ecole Superieure des Communications de Tunis (SUP'COM), in June 2005. He is currently a Professor in wireless communication with the University of Carthage, the former President, the Chairman, and a member of the Board of the National Telecommunication Regulatory Authority of Tunisia (INTT), for more than five years. He was the former Head of the Department of Applied Mathematics and Signal Processing, Higher School of Engineering in Communication in Tunisia (Sup'Com). He has more than 25 years of experience in higher education and ICT. He was a Postdoctoral Fellow of Concordia University, from 1999 to 2000, and a Visiting Scholar with Colorado State University, in 2011. He served as a Senior Engineer and a Member of Technical Staff at Legerity Inc., and Celite Systems Inc., Austin, TX, USA, working on xDSL technologies. He published more than 120 scientific articles in international journals and conferences dealing with wireless communications. He holds one U.S. patent, and he supervised nine Ph.D. thesis. Under his leadership, the INTT was elected as the Best African Regulator, in 2015, and he was nominated for the GSMA Government Leadership Award, in 2018.

YEHIA MASSOUD (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently the Dean of the School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, USA. He has authored over 280 articles in peer-reviewed journals and conferences. He has served as a Distinguished Lecturer of the IEEE Circuits and Systems Society and an Elected Member of the IEEE Nanotechnology Council. He was selected as one of ten MIT Alumni Featured by the MIT's Electrical Engineering and Computer Science Department, in 2012. He was a recipient of the Synopsys Special Recognition Engineering Award, in 2000, the DAC Fellowship, in 2005, the National Science Foundation CAREER Award, in 2005, the Rising Star of Texas Medal, in 2007, several best paper award nominations, and two best paper awards at the IEEE International Symposium on Quality Electronic Design, in 2007, and the IEEE International Conference on Nanotechnology, in 2011. He has held several academic and industrial positions, including a Member of Technical Staff with the Advanced Technology Group, Synopsys, Inc., CA, USA, a tenured faculty with the Departments of Electrical and Computer Engineering and Computer Science, Rice University, Houston, USA, the W. R. Bunn Head of the Department of Electrical and Computer Engineering, The University of Alabama, Birmingham, USA, and the Head of the Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, USA. He has served as the Editor for *Mixed-Signal Letters—The Americas* and also as an Associate Editor of the IEEE Transactions on V bioscery Large Scale Integration Systems and the IEEE Transactions on Circuits and Systems I.