Journal of Universal Computer Science, vol. 29, no. 5 (2023), 446-469
submitted: 17/8/2022, accepted: 19/1/2023, appeared: 28/5/2023 CC BY-ND 4.0

Big Data Provenance Using Blockchain for Qualitative Analytics via Machine Learning

Kashif Mehboob Khan
(Software Engineering Department, NED University of Engineering and Technology, Karachi, Pakistan
https://orcid.org/0000-0002-7208-6072, kashifmehboob@neduet.edu.pk)

Warda Haider
(Software Engineering Department, NED University of Engineering and Technology, Karachi, Pakistan
https://orcid.org/0000-0002-5054-2313, haider4202397@cloud.neduet.edu.pk)

Najeed Ahmed Khan
(Computer Science Department, NED University of Engineering and Technology, Karachi, Pakistan
https://orcid.org/0000-0003-1986-7192, najeed@neduet.edu.pk)

Darakhshan Saleem
(Bio-Medical Engineering Department, SSUET, Karachi, Pakistan
https://orcid.org/0000-0001-8712-3617, darakhshansaleem@yahoo.com)

Abstract: The amount of data is increasing rapidly as more and more devices are being linked to the Internet. Big data has a variety of uses and benefits, but it also has numerous challenges associated with it that are required to be resolved to raise the caliber of available services, including data integrity and security, analytics, acumen, and organization of Big data. While actively seeking the best way to manage, systemize, integrate, and affix Big data, we concluded that blockchain methodology contributes significantly. Its presented approaches for decentralized data management, digital property reconciliation, and internet of things data interchange have a massive impact on how Big data will advance. Unauthorized access to the data is very challenging due to the ciphered and decentralized data preservation in the blockchain network. This paper proposes insights related to specific Big data applications that can be analyzed by machine learning algorithms, driven by data provenance, and coupled with blockchain technology to increase data trustworthiness by giving interference-resistant information associated with the lineage and chronology of data records. The scenario of record tampering and big data provenance has been illustrated here using a diabetes prediction. The study carries out an empirical analysis on hundreds of patient records to perform the evaluation and to observe the impact of tampered records on big data analysis i.e diabetes model prediction. Through our experimentation, we may infer that under our blockchain-based system the unchangeable and tamper-proof metadata connected to the source and evolution of records produced verifiability to acquired data and thus high accuracy to our diabetes prediction model.

Keywords: Blockchain, Data provenance, Big data, Machine learning
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.93533

1 Introduction

Blockchain is a decentralized, peer-to-peer database that appends an ever-increasing volume of transactions [Zheng et al., 2018]. Each transaction is hashed into a block which is also timestamped into the main consensus chain. To develop and implement blockchain systems that run within or across organizations, Multichain is one of the available blockchain development platforms. The platform offers a straightforward command-line interface and supporting APIs that are appropriate for handling transactions. Accounts' rights management, handling data streams, and native assets creation are just a few of the many aspects that make up Multichain a good choice to adopt. A range of applications benefit from these top-notch features in terms of integrity, security, integration, scalability, and adherence [Ismailisufi et al., 2020].

Data provenance is a process of logging data from its origin through transformation and transmission. It describes a documented chronology of an artifact. This refers to how the object was made, changed, spread, and dispersed to reach its current state. We may determine how reliable an object is by looking at its provenance [Buneman et al., 2006].

Over the last decade, there has been an extraordinary increase in the worldwide data flow, which has gained a particular interest in "Big data". According to an estimate, In 2025, the market for Big data will be worth $227.4 billion and will drastically cut costs across a range of vertical industries, including medical, commerce, logistics, manufacturing, entertainment, and media. Big data is being researched in many areas of science and engineering, including organizational structure, computer vision, and smart cities, despite the lack of a clear definition for it. In addition to the structural manifestation, attributive, correlative, and structural considerations for large data were made [Sagiroglu and Sinanc 2013].

Companies can gather and handle enormous amounts of data due to the advancement of cloud storage. The Internet of things, corporate systems, and unorganized resources like internet forums all provide data. Firms can shed light on data with the aid of new analytics technologies like Hadoop [Bhosale and Gadekar 2014]. However, possessing, gathering data, and analyzing tools alone does not guarantee that the findings of a research are significant. The accuracy of the data is essential for gaining meaningful insights from it [Dai et al.,2008]. There are several opportunities for inaccuracies to be introduced intentionally or accidentally due to the numerous streams that input into data storage and the various transformations that Big data undergoes during processing. The ability for businesses to monetize their data via distributing it to others is constrained by a lack of confidence in the data, which also restricts its usage within the firm that obtained it.

Large-scale real-world problems have been thought of as having potential solutions, including the merging of Big Data and blockchain technology. The exponential expansion in data generation poses its own data protection obstacles, as well as problems with the dependability of data sources and data sharing [Syed et al., 2013]. Blockchain technology's distinctive characteristics, such as decentralized storage, visibility, immutability, and consensus mechanisms, are the solution to resolve the problems faced by the Big Data ecosystem. By combining it with Big data advantages, one can improve Big Data security and privacy while also facilitating the data sharing of continuously growing data, real-time data analytics, improving data quality, preventing fraud, and streamlining data access. Thus blockchain makes data management more effective in a variety of applications, including managing high-quality web data and scientific data [Karafiloski and Mishev 2017].

A crucial data provenance component is reorganizing the history of the data connected to each data processing or scientific finding. It must store content in a tamper-proof and reproducible manner for the information it provides to be trusted. The unchangeable and tamper-proof metadata related to the content source and evolution of documented records provided by Big data provenance offers verifiability of gathered data [Appelbaum 2016].

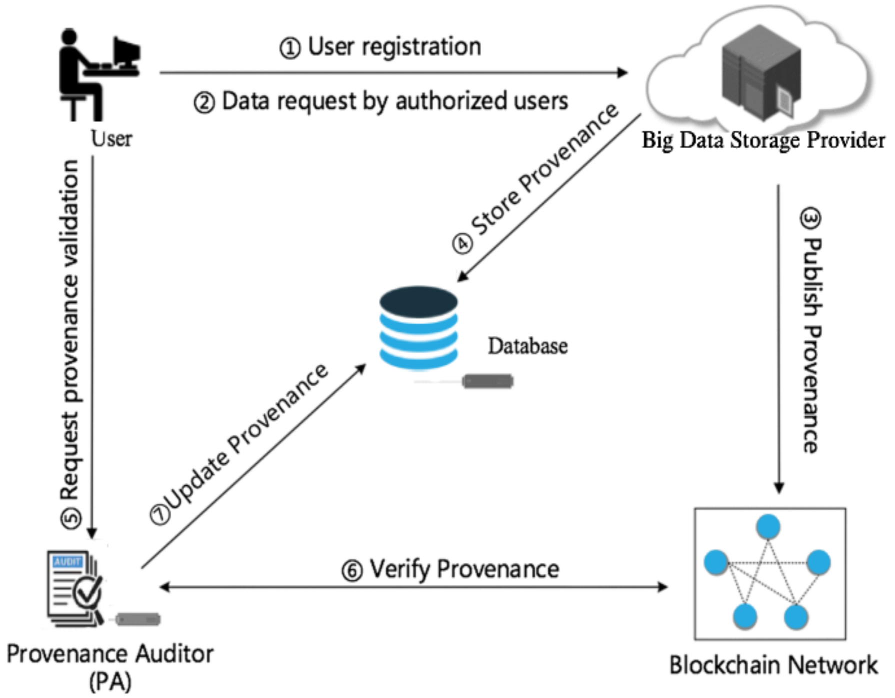


Figure 1: An authorized access can only make changes in the Big data once its request is approved and validated. The metadata and lineage information is collected throughout the lifecycle of an object.

Figure 1 (inspired by a paper by Hu et al. (2020) on "A survey on data provenance in IoT" containing ProvChain interaction framework) shows the cycle of producing, publishing, and verifying provenance data without incorporating machine learning techniques. This paper particularly takes into account the incorporation of big data provenance methodology inside blockchain transactions to overcome the challenges with data collecting and verification along with a machine learning prediction model. Following are the primary contributions of this proposed study:

I. Extensive research and empirical analysis allowed us to concretely show how Big data provenance via blockchain is a potential strategy to guarantee the integrity of stored data. According to our knowledge, no previous empirical research has been done that considers the malleability of data collection, manipulation, creation, and recording specifically in this context.

II. A blockchain-based diabetes predictor model has been proposed, implemented, and empirically evaluated to demonstrate the record tampering and big data provenance paradigm

III. Evaluation of how Big data provenance systems via blockchain can offer details on the beginnings and development of data, including the many phases of data acquisition and data manipulation, who started them, when and how they happened, and who initiated them.

The paper provides novelty in the implementation of a blockchain-based provenance approach to ensure that the data remains tamper-proof and its history of genesis may be retained intact to produce the best out of a machine learning algorithm. Empirical analysis has also been conducted to observe the impact on tampered (without blockchain-based provenance) and without tampered (with blockchain-based provenance) to show the significance of our approach. In order to provide security and prevent data from any means of fraudulent transaction, we distributed the contents of our data into on-chain and off-chain transactions using a blockchain platform where the data is validated & compared between on-chain & off-chain. Miners also require enhanced difficulty levels to propose blocks.

The paper has been organized as follows: Section II describes the existing research on Big data provenance to examine the importance of blockchain-based solutions. Section III discusses the required and optional components of a general big data provenance system along with the latest tools and technologies. Section IV examines the importance of adopting blockchain for Big data provenance and its applications. Section V discusses the proposed blockchain-based diabetes predictor model to demonstrate the record tampering and big data provenance paradigm. In section VI, the architecture and functioning of our system are provided, together with information about our blockchain setup, including the machine learning model we used, the kind of blockchain we built, and the approaches and procedures we employed to implement our prediction model. The experimentation work has been illustrated in Section VII, followed by Section VIII to analyze the experimentation results. Section IX of the report contains summaries of the investigation.

2 Related Work

This section demonstrates the scientific research associated with our investigation of Big data provenance using blockchain. As discussed in [Buneman et al., 2001], S. Khanna et al. provide a method for query provenance in which they distinguish between "where" and "why" provenance for databases. For the last ten years, the amount of data traffic has risen at an unprecedented rate, giving "Big data" a unique level of attention. Big data is being studied in many branches of science and engineering, including operation management, computer vision, and smart cities. There are several techniques to use Big data efficiently. For instance, Big data gives mobile networks more opportunities to improve the quality of their services. With an emphasis on examining the Big data features from the mobile network operator's and users' viewpoints, the research work of [Zheng et al., 2016] investigated the incorporation of mobile network optimization with Big data. Data from the core network, Internet service providers, and radio access networks are all included in network operator data. Profile information and geographic data of users are included in user data. Network operators' ability to analyze this data and make informed judgments is what determines how well mobile network services perform. For better network optimization, effective data analysis mechanisms are necessary. Despite its numerous advantages and applications, Big data presents many challenges that need to be solved to improve service quality.

Recently, blockchain technology, a decentralized ledger, has become one of the most alluring options for improving the quality of services of different systems. For example, the paper [Kosba et al., 2016] suggests a system Hawk extracts transactional privacy from open blockchains.

Blockchain technology also offers appealing resolutions for ensuring security and privacy, a critical issue in Big data applications. It is demonstrated by [Kiayias et al., 2017] that the provision of high-quality and secure data has the potential to be greatly aided by blockchain technology. PoS blockchain protocol, published as Ouroboros, was suggested to offer strict security assurances in blockchains and greater reliability than proof-of-work blockchains. Also, Xueping [Liang et al., 2017] presented the architecture and execution of a system named ProvChain. It is a data-provenance solution for cloud monitoring built on the blockchain with better availability, reliability, and user privacy protection. However, it does not provide data provenance in a federated cloud computing environment. Furthermore, Esmeralda and Nazri [Abdullah et al., 2017] proposed blockchain as a requirement to enhance Big data security in distributed environments as earlier authentication protocols like Kerberos have many security issues. The implementation involves an authentication framework including passwordless authentication, encryption of data, and a decentralized database that is supported by blockchain.

A smart toy prototype based on IoT devices, edge computation, and blockchain is presented by [Yang et al., 2018] for secure data exchange. IDs are stored and verified by smart contracts for each data transmission. Chaincode, a smart contract built on the blockchain, manages the intricate bookkeeping in the market for smart toys. Low-latency response is made possible for local client computations by edge computing. The framework ensures that data flow between members of the smart toy industry is private, adaptable, scalable, and secure.

Some studies by [Al-Mamun et al., 2018; Bandara et al., 2018; Uchibeke et al., 2018] have been conducted on data provenance using blockchain. In [Al-Mamun et al., 2018] work, a novel in-memory blockchain-based system design is demonstrated, where ledger collection is primarily kept in memory, along with a recent unison mechanism designed for the latest design on HPC systems. It was constructed and tested with more than a million transactions, and the results showed a 32-fold speedup over the provenance service based on file systems and a four-order-of-magnitude speedup over the provenance service based on databases. Mystiko [Bandara et al., 2018] is a blockchain database that easily incorporates Big data with blockchain. It supports features like high transaction throughput, searching based on keywords using Elasticsearch, and scalability of storage. Financial and banking areas integrated it to develop large-scale applications. In the paper [Uchibeke et al., 2018], Ralph Deters and Kevin A. presented a structure for managing access controls using a decentralized method of security hinged on the individual and authorized blockchain with a hyperledger. The blockchain's underlying technology presents a response to the problems with conventional It ensures data openness and centralized access control and data auditability, traceability, and safe data sharing.

Later blockchain studies concentrate on several system viewpoints. In [Dai et al., 2018], a novel architecture is put forth to use network coded distributed storage to resolve the problem of retention bloating in blockchains. Another piece of research [Gao et al., 2018] investigates how blockchain attacks powered by quantum computing can be thwarted. [Xiao et al., 2018] make a detailed recommendation for improving the hardware-level dependability of blockchain topologies.

Although some work [Hogan and Helfert 2019; Ruan et al., 2019; Tosh et al., 2019] have been conducted on data provenance using blockchain, these are on a limited amount of data. For Example, Gabriel and Markus in their paper examined the relationship between the PROV requirements for data provenance and blockchain distributed ledger technologies. DLT can be used as a medium for data provenance of Big data in the cloud. The experimentation reveals that not all PROV data models have a correspondent relationship with blockchain because the single linked list property of DLT does not make it possible. Also, Pingcheng Ruan and Gang Chen [Ruan et al., 2019] have presented LineageChain, a provenance system for blockchains. Provenance data is effectively captured by the system while it is running and is kept in safe storage. It makes smart contracts accessible through basic APIs. LineageChain's efficient provenance inquiries are made possible by a unique skip list index. On the crest of Hyper Ledger, they implemented LineageChain, as well as compared it with various benchmarks. The outcomes highlight LineageChain's advantages in reinforcing robustness in provenance-dependent applications. They show how effective provenance inquiries are, as well as how little storage overhead the system has. Furthermore, BlockCloud, a blockchain-hinged data provenance system for Big data generated via the cloud, is proposed by [Tosh et al., 2019] to offer this security feature by tamper-resistant auditing of each cloud user's transaction on multiple data objects. They looked at the idea of employing a PoS-featured consensus to maintain continuity in the distributed network due to several performance and security challenges associated with a PoW-based blockchain. Furthermore, the framework for cyber-based physical social systems by [Tan et al., 2020] was proposed for real-time Big data applications. Blockchain is used to control access. At the edge nodes, the architecture takes advantage of fog computing to dynamically process local data. For data transactions that require privacy protection, encryption is carried out using a compact symmetric technique. In the blockchain, information about access control is kept and managed, including authentication and authorization. Experiments reveal that maintaining anonymity takes more time because all authorizations are processed within the blockchain. Strengthening retrieval mechanisms is necessary to increase performance.

According to our knowledge, only surveys and theoretical studies on blockchain for Big data have been conducted. Only a few and limited published empirical investigations are present in the given context. This situation demands a strong empirical analysis of Big data using blockchain based provenance over an efficient machine learning based model.

3 Requirements and Tools

This section outlines the essential and optional components which are required for a Big data provenance system. We have kept managed the provenance data information along with individual ID on a separate file.

I. Abstraction of provenance: To enable provenance use cases to map their own requirements, the data should offer generic data origin collection, saving, and inquiring functionality.

II. Provenance with high and low levels: These records include high levels accompanied by low-level data items. Lower-level components include sensors whereas components of high level reflect more abstract notions, such as a physical thing inside a distribution chain or an insight outcome hinged on several inputs, rather than acquiring a singular origin (like sensor reading).

III. Entirety: The provenance logs of a piece of data are said to be full when all pertinent activities that have ever been made on it have been collected. Therefore, relevance suggests that certain acts may be disregarded if pedigree information is not added.

IV. Lineage Creation: For instance, by generating a new lineage trace depending on the previous at every vital stage of the chain, it is possible to trace the ancestry of data reflecting a physical object moving throughout its lifecycle.

V. Deduction: A derivation or deduction of data relates specifically to the origin information of the data points which were used to create it. For example, a lineage reference for an analytical finding related to asset readings from many sensors must view provenance information for readings like the position and time of recording, in relation to talking about the sensor values themselves.

VI. Provenance for data point alterations: The framework enables the recording of a data point's history of modifications. The chronology of these computations can be followed, for example, if an analytical result goes through several phases and several calculations.

VII. Provenance in Parallel: For a given data point, multiple provenance records may exist concurrently. For instance, one provenance record may track a data point's ownership (for example, the present owner of a physical object), while a second record may track its location.

VIII. Integrity: Provenance records must be free from any manipulation or alteration on account of integrity. For the data to be trusted, this is essential. Clients may potentially reject provenance records if there is no guarantee of integrity.

IX. Accessibility: This enables the client to accurately reconstruct the chronology of data production and modification with the help of the provenance model. To do this, even if some system components malfunction, the provenance data must be accessible to clients upon request.

X. Privacy: In general, maintaining privacy entails preventing unauthorized gathering, storing, and access to sensitive data. Confidentiality and anonymity are both included in privacy. The provenance logs of IoT devices also include private information, such as in a system for monitoring health. Therefore, it is essential to maintain the data's confidentiality and to stop unauthorized parties from accessing it. In order to protect secrecy, tracking of provenance data must be avoided. Both concepts are referred to as privacy from here on out.

XI. Extensibility: A big data provenance system should be of a rational cost. Both retrieving provenance data and storing it must have little overhead. In particular, Internet of Things devices that are resource-curbed must not be barred from taking part in traceability. A provenance solution also needs to take into account the possibility for applications in the Internet of Things to deal with enormous volumes of data and frequent data upgrades.

Building blockchain-based apps that adhere to corporate standards will become simpler for organizations to do if features like scalability, query ability, and audit trails are added. The adoption of blockchain by businesses will be prompted by the growth of the tools as well as the ongoing push for digital transformation, which necessitates that they make better use of data generated and collected digitally. The following are the tools available for companies to adopt to accomplish the target of the provenance of big data.

i. BigChainDb

With BigChainDb [McConaghy et al., 2016], the target of decentralized building blocks with blockchain databases for storage and processing can be fulfilled. The key obstacles of the big data can be straightened out via bigChainDB:

ii. HBasechainDB

In the Hadoop ecosystem, HBasechainDB incorporates the immutability and decentralization of blockchain technology into the HBase database [Sahoo et al., 2018]. To achieve linear scaling, computation is pushed to the data nodes. Since a distributed, decentralized, and impenetrable Big Data store can be created using HBasechainDB.

iii. Mystiko

A blockchain database called Mystiko makes it simple to combine big data and blockchain. It is built on top of Apache Cassandra. It offers functions including high transaction throughput, Elasticsearch keyword searching, and storage scalability.

iv. Neo4j

For securing data, a decentralized data storage approach is preferred which makes the data less compromisable in comparison to centralized data storage. Neo4j is a graph database that supports big data storage for many purposes like analytics.

4 Significance And Application

Blockchain technology has exploded in popularity due to its numerous uses in a wide range of fields. It is still not in its maturity stage and is being utilized in several use cases to address a variety of challenges in various domains such as data governance, fraud protection, decentralization, etc. The current age is living in a time when man-made and automated machines are producing an excessive amount of digital data. As a result, there is a continuously high demand to store, arrange, filter, and investigate this large data. We believe blockchain has the potential to achieve the desired expected results. These contributions may benefit as follows;

i. It may be recommended for real time data analytics due to the blockchain's ability for decentralized storage of each and every transaction.

ii. Enhancement in data integrity to restrict tampering of records and thereby provide a perfect baseline for the analytics' predictions against increased accuracy of data.

iii. Since big data is not housed within an organization's network perimeter, conventional security methods like firewalls are unable to handle this problem. This is because enterprises do not have control over the data. Blockchain storage of massive data can solve this problem.

iv. The combination of big data and blockchain enables service organizations to offer data to other stakeholders while reducing the danger of data loss. Additionally, since every experiment performed is documented on the blockchain, analyzing the huge data gathered from many sources does not need to be repeated.

v. The data quality can possibly be raised by hoarding it on a blockchain as it is structured and comprehensive. As a consequence, data scientists can use improved standard data to produce more precise forecasts that are made in real-time.

Managing access control and upholding data ownership and transparency has always been extremely difficult. By keeping identity management to private information in the blockchain framework, blockchain technology solves this problem. By establishing a protocol that allows individuals to have and check their data, a decentralized personal data management system is established utilizing blockchain technology. Organizations may now concentrate more on data consumption than system security and encapsulation because the reliance on third parties has been fully eliminated. The potential domains where a big data provenance system based on blockchain technology may be implemented are listed below.

i. Finance

It is the first industry to consider when using blockchain with big data [Treleaven et al., 2017]. The blockchain for Bitcoin stores every transaction. Bitcoin is pseudo-anonymous; it provides some privacy, but the financial information is not private. It can determine trends in Bitcoin exchanges and eventually connect those to specific individuals with enough data. Many businesses are developing these solutions such as Chainalysis, which provides analytics to stop, detect, and look into bitcoin fraud, money laundering, and compliance problems.

ii. Supply Chain

Supply-chain is one of the most promising areas dominated by blockchain technology nowadays. Any industry can make smart decisions by preserving detailed transaction data. There are several opportunities for institutions to make the most of the provenance of data. In order to gather information on usage throughout a product's lifecycle, sensors may be attached to the product. Immutable, verifiable, and transparent product data on a blockchain may be very helpful in providing customers knowledge about the source of their food and insurance of data reliability.

iii. Smart City

Rapid urbanization has resulted in the formation of "smart cities," which need effective and thoughtful elucidation for governance, ecology, freightage, and energy optimization[Li 2018]. There are a lot of issues with poor security, dependability, maintenance, flexibility, and expenses. Blockchain technology satisfies these requirements for IoT device maintenance, space, energy efficiency, and transparency. A tamper-proof transaction can be ensured using asymmetric encryption.

iv. Smart Health Care

Medical data collection has dramatically increased as a result of recent developments in the healthcare industry. For purposes of diagnosis, prognosis, and treatment, these facts are crucial. One typical tool for treating older patients, in particular, is the telemonitoring system. Despite the fact that these technologies offer many advantages, the aforementioned security concerns when transmitting and recording data transaction information exist. However, these problems have the potential to seriously violate both data privacy and security.

Similarly, blockchain is being extensively used in various application areas such as health monitoring, vaccination management activities, digital forensics, and agriculture as an entirely decentralized blockchain-based platform called AgriBlockIoT keeps data provenance for the agro-based supply chain.

5 Proposed Blockchain-Based Diabetes Predictor

In our suggested big data analysis approach, illustrated in Fig. 2, several entities interact with one another to facilitate the prediction model. To ensure that the entries are valid and not prone to error, we have constrained our model to only collect the valid records thus securing information provenance records, and boosting data trust. To fully understand the impact of outliers on predictions that are introduced in records due to record tampering, we have provided a thorough implementation of our predicting model.

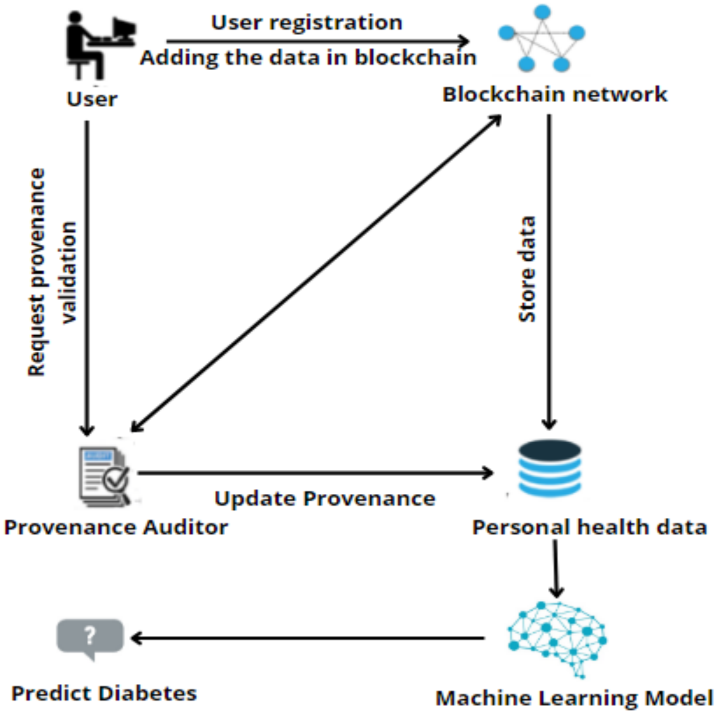


Figure 2: A Blockchain-based machine learning model architecture for Big data

Our proposed diabetes prediction model, which has been integrated into the blockchain platform to study the effects of mutated records, is shown in Figure 2. By first creating unique addresses via blockchain, we mapped the users using blockchain wallet addresses (those who store patients' data in databases). Throughout the entire patient's data recording, these addresses will serve as an ID. Hashes representing the registration data for users (including some provenance metadata) will be added to the blockchain. These IDs (user addresses) were later utilized by us to create user lists (CSV files). Thus only a valid user can add or update the patient's record. A record origination is now verifiable, meaning it can be traced back to the specific user who created it. Otherwise, it was impossible to rule out the potential that the alteration in the record was the result of a hostile attack that tampered with the system if indeed the events that led to it were unclear. In the context of data provenance, we will look into how our developed model behaves when employing various active users and how their transactions affect our model diabetes prediction. We'll be able to keep an eye on and manage different users' transactions owing to this. Based on specific diagnostic parameters included in the dataset, the goal of the data provenance-oriented dataset is to forecast whether or not a patient has diabetes. The datasets consist of one target variable, Outcome, and a number of medical predictor parameters. The patient's BMI, the frequency of pregnancies she has, insulin and glucose levels, age, skin thickness, diabetic pedigree function, and blood pressure are among the predictor parameters.

In order to avoid the threat of any possible manipulation within the consensus blockchain by taking over more than 50% of the mining power, we categorically made two changes;

1. Increasing the required PoW bits (difficulty level) for the miners so that the cost of attack may be raised to inject fraudulent transactions as it would require a considerable amount of energy and resource consumption by the potential attacking mining node(s) to propose a malicious block for the consensus blockchain.

2. Although the chain is private but in order to maximize the level of decentralization, we created our own scalable mining pool. This approach also raises the cost to inject fraudulent transactions by increasing the difficulty level for miners to propose a block (which requires a considerable amount of energy and resource consumption)

3. The second security measure includes the creation of a mining pool containing a sufficient number of miners. Whenever any transaction arrives in the pool of unconfirmed transactions, 60% of the miners are selected at random to compete for the PoW. Since the difficulty level has already been increased to 16 bits and miners are now being selected randomly (to restrict monopoly and selfish mining) through mining diversity [Ismailisufi et al., 2020], the system now safeguards against malicious threats and attempts.

Table 1 shows the attributes of blockchain with extended security configurations of consensus algorithm (PoW) difficulty level along with mining diversity

Platform	Blockchain Parameters				
	Mining Diversity	No. of miners	Block generation Rate	Maximum Allowable Block size(MB)	No. of Bits Required for Proof of work
Windows	0.6	15	15	1	16

Table 1: Blockchain Attributes

6 Implementation and Setup

We built and established a procedure where, when the record is created, a corresponding digital token is generated, acting to authenticate its point of origin, to run the prediction model as shown in Figure 2. The digital token is then transmitted concurrently with each time the record is transferred, precisely mirroring the chain of custody in the real world through a series of blockchain transactions. We evaluated our model using a bulk of records from several remote registered users to see how well it could anticipate diabetes in real-time, based on the provenance of the record's current state. The open-source blockchain technology Multichain (Version 2.0 Alpha 4), which was made available by the Multichain community in 2018, was used to build the decentralized blockchain network. We have used the Python platform to build our model for diabetes prediction inculcating big data provenance-oriented dataset. The model has been implemented via Python libraries including seaborn (version 0.11.1), numpy (version 1.23.1), matplotlib (version 3.1.0), pandas (version 1.4.3) and scikit_learn (version 1.1.1).

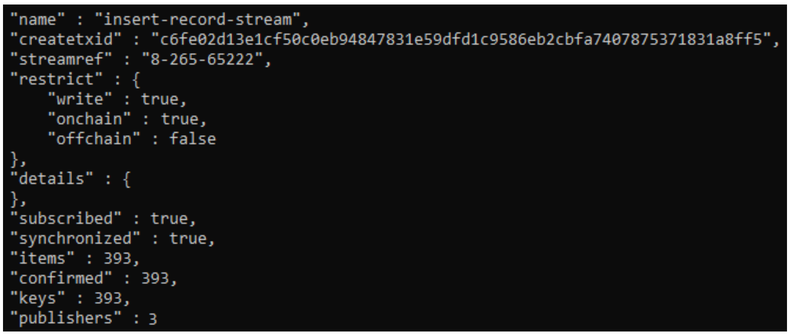
Algorithm 1 Diabetes Prediction

```
procedure Predict Diabetes (PatientChain, UserList)
    UniqueAddress ← UsertList[arrayindex]
    if  user in UserList
    PatientData ← Health Record
    prediction = model.predict(sc.transform(Patient Data))
    if prediction == 1 then
        pred.message( "You have Diabetes")
    else if prediction == 0 then
        pred.message( "You don't have Diabetes.")
    return;
End procedure
```

7 Experimentation

The scenario of record tampering and big data provenance has been illustrated here using a diabetes prediction scenario, previously indicated in Section VI, as an implementation use case. Over three hundred and fifty patient records were used to perform the evaluation overall and to observe the effect of tampered records on big data analysis and interpretation. According to a diabetes prediction model used in the system's implementation, only patient information can be recorded using registered addresses found in users' lists. In order to make observations, this aids in creating a realistic data recording scenario. One target variable, Outcome, and a variety of medical predictor characteristics make up the datasets. The patient's BMI, insulin and glucose levels, age, skin thickness, the function of hereditary diabetes, blood pressure, and the number of pregnancies the patient has had are some of the predictive parameters [Joshi and Dhakal 2021]. A blockchain does aid in our model's management of this patient data access by establishing an immutable record of digitally signed authorizations. The chain of records cannot be tampered with by a single entity or limited group of entities, giving end users more assurance about the predictions made by our prediction model. Moreover, various tokens can be immediately and safely traded, and the simplest blockchain level guarantees a two-way swap.

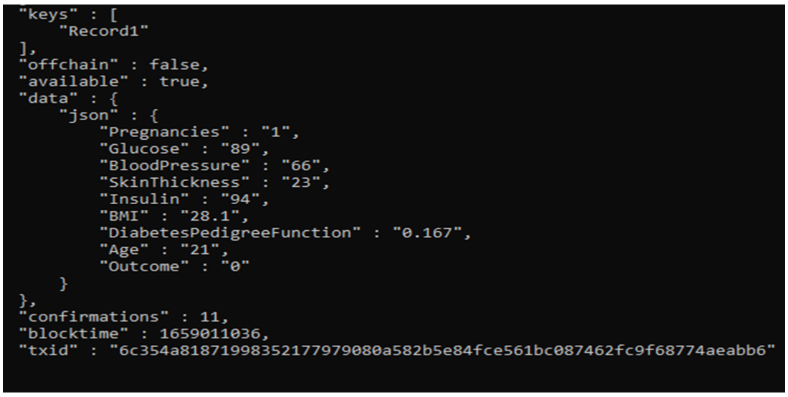
A multichain, "diabetesRecord," was created to allow an authenticated user to store only accurate patient data and prevent any tampering or erroneous mutations in the dataset as shown in Figure 3. The users were subscribed to the “insert-record-stream”, to permit them to publish the records along with associated metadata via a unique address. This will enable us to trace any induced modifications in the dataset back to its origin and throughout Big data evolution.



```
"name" : "insert-record-stream",
"createtxid" : "c6fe02d13e1cf50c0eb94847831e59dfd1c9586eb2cbfa7407875371831a8ff5",
"streamref" : "8-265-65222",
"restrict" : {
    "write" : true,
    "onchain" : true,
    "offchain" : false
},
"details" : {
},
"subscribed" : true,
"synchronized" : true,
"items" : 393,
"confirmed" : 393,
"keys" : 393,
"publishers" : 3
```

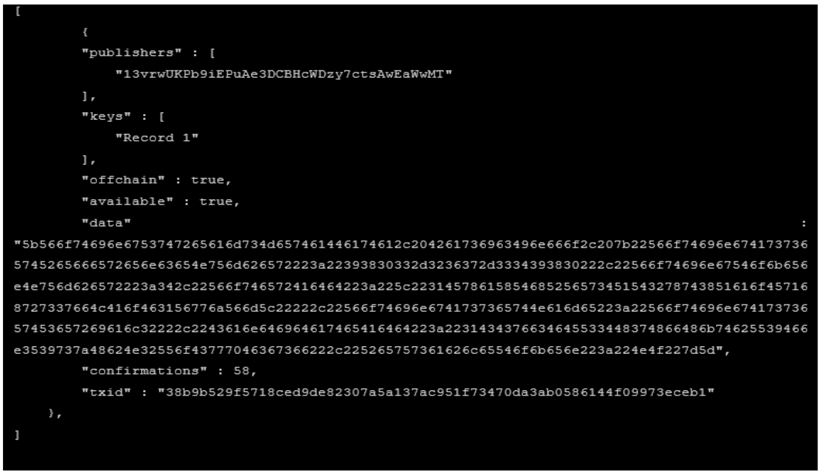
Figure 3: Illustration of “insert-record-stream” with 3 publishers

Hundreds of patients’ records have been stored in the stream including the patient's BMI, insulin and glucose levels, age, skin thickness, the function of hereditary diabetes, blood pressure, and the number of pregnancies the patient has had. A sample of the published records is represented in Figure 4.



```
"keys" : [
    "Record1"
],
"offchain" : false,
"available" : true,
"data" : {
    "json" : {
        "Pregnancies" : "1",
        "Glucose" : "89",
        "BloodPressure" : "66",
        "SkinThickness" : "23",
        "Insulin" : "94",
        "BMI" : "28.1",
        "DiabetesPedigreeFunction" : "0.167",
        "Age" : "21",
        "Outcome" : "0"
    }
},
"confirmations" : 11,
"blocktime" : 1659011036,
"txid" : "6c354a81871998352177979080a582b5e84fce561bc087462fc9f68774aeabb6"
```

Figure 4: Sample of published Record1

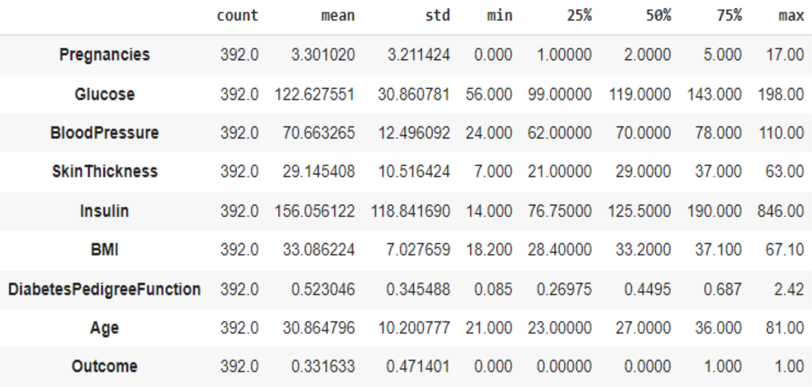


```
[
        {
        "publishers" : [
            "13vrwUKPb9iEPuAe3DCBHcWDzy7ctsAwEaWwMT"
        ],
        "keys" : [
            "Record 1"
        ],
        "offchain" : true,
        "available" : true,
        "data"                                                                                  :
"5b566f74696e6753747265616d734d657461446174612c204261736963496e666f2c207b22566f74696e674173736
574526566572656e63654e756d626572223a22393830332d3236372d3334393830222c22566f74696e67546f6b656
e4e756d626572223a342c22566f746572416464223a225c2231457861585468525657345154327874385161 6f45716
8727337664c416f463156776a566d5c22222c22566f74696e6741737365744e616d65223a22566f74696e674173736
57453657269616c32222c2243616e646964617465416464223a2231434376634645533448374866486b74625539466
e3539737a48624e32556f43777046367366222c225265757361626c65546f6b656e223a224e4f227d5d",
        "confirmations" : 58,
        "txid" : "38b9b529f5718ced9de82307a5a137ac951f73470da3ab0586144f09973eceb1"
    },
]
```

Figure 5: Sample of off-chain published Record1

We extended our security check to two layers, on-chain and off-chain for the transaction data. This is achieved by introducing an off-chain provenance layer which is accountable to capture provenance data in the form of hashes at the node level (off-chain). In order to validate the state of data in the main consensus chain at the time of its genesis and at the time when it is committed at the chain, it may easily be verified through its respective off-chain data counterpart. Figure 5 shows the demonstration of one of the off-chain published transaction records.

The summary of our big data that was recorded via registered addresses through blockchain is presented in Figure 6.



	count	mean	std	min	25%	50%	75%	max
Pregnancies	392.0	3.301020	3.211424	0.000	1.00000	2.0000	5.000	17.00
Glucose	392.0	122.627551	30.860781	56.000	99.00000	119.0000	143.000	198.00
BloodPressure	392.0	70.663265	12.496092	24.000	62.00000	70.0000	78.000	110.00
SkinThickness	392.0	29.145408	10.516424	7.000	21.00000	29.0000	37.000	63.00
Insulin	392.0	156.056122	118.841690	14.000	76.75000	125.5000	190.000	846.00
BMI	392.0	33.086224	7.027659	18.200	28.40000	33.2000	37.100	67.10
DiabetesPedigreeFunction	392.0	0.523046	0.345488	0.085	0.26975	0.4495	0.687	2.42
Age	392.0	30.864796	10.200777	21.000	23.00000	27.0000	36.000	81.00
Outcome	392.0	0.331633	0.471401	0.000	0.00000	0.0000	1.000	1.00

Figure 6: Summary of patient's record dataset

Following observations can be made from Figure 5. The dataset consists of 392 entries and 9 features altogether.

1. Both integer and float data types are available for each feature.
2. The mean values for some parameters, such as BMI, insulin, glucose, and blood pressure are 122.62, 70.66, 156.0, and 33, respectively.
3. The dataset contains no NaN values.
4. 1 in the outcome column denotes a positive diabetes result, while 0 denotes a negative diabetes result.

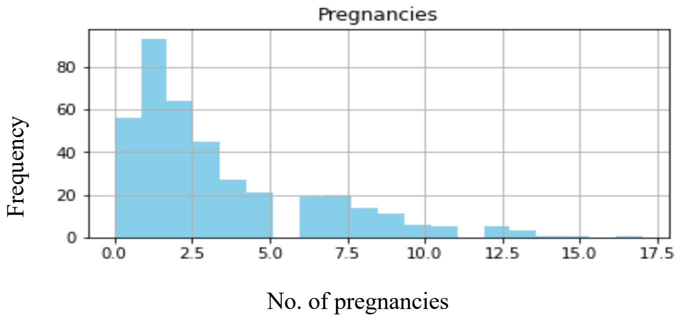


Figure 7 a: Histogram representing the trend of pregnancy feature

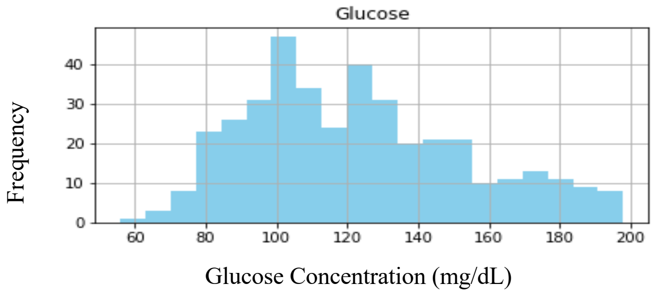


Figure 7 b: Histogram representing the trend of glucose feature

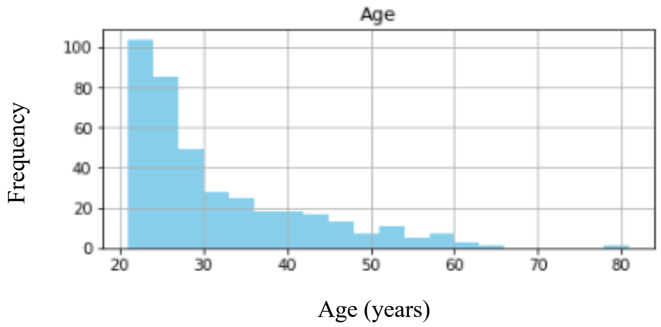


Figure 7 c: Histogram representing the trend of Age feature

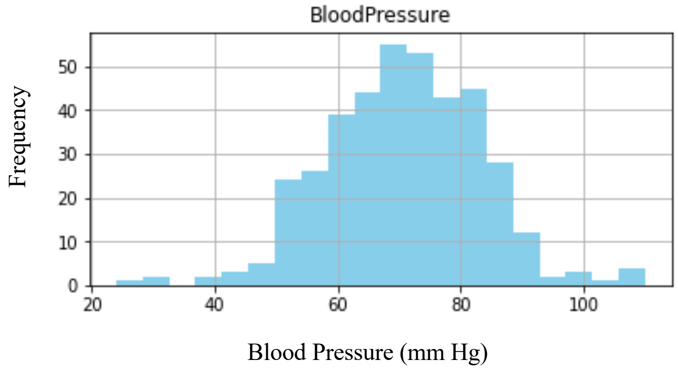


Figure 7 d: Histogram representing the trend of blood pressure feature

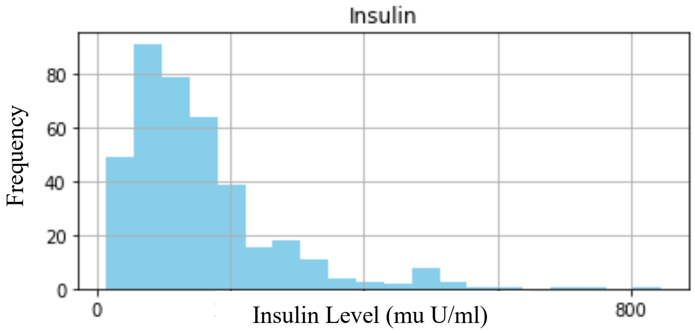


Figure 7 e: Histogram representing the trend of Insulin feature

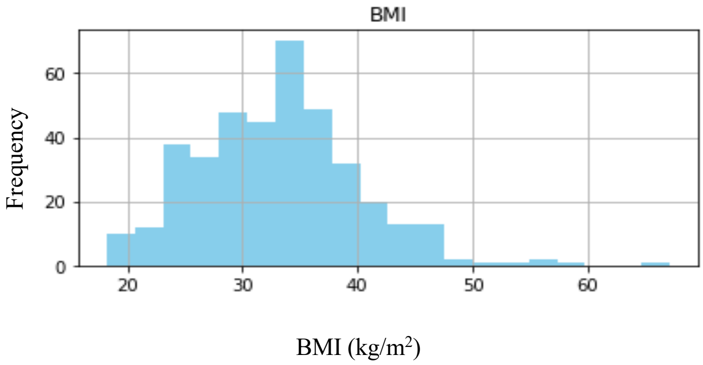


Figure 7 f: Histogram representing the trend of BMI feature

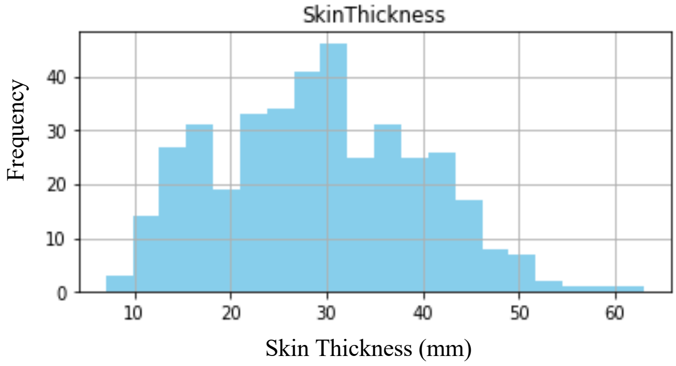


Figure 7 g: Histogram representing the trend of skin thickness feature

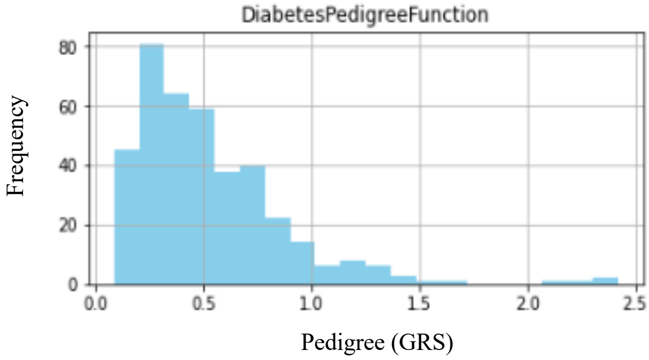


Figure 7 h: Histogram representing the trend of the Diabetes Pedigree Function feature

Figure 7 shows the graphical representation of all the parameter data involved in predicting the outcome. The majority of the individuals in our dataset had two to three pregnancies, as seen in figure 7a. Figure 7b illustrates that the glucose level has a substantial influence, which is why the likelihood of being exposed as a diabetic patient grows as the glucose level in the blood cell rises. The blood sugar level rises because insulin is not created properly or is not utilized effectively by our bodies. Additionally, figure 7c shows that the majority of the individuals in our sample are in the 20 to 35 age range. At this age, the normal blood pressure level is between 60-80 as shown in figure 7d. Only a few have high blood pressure which is caused due to diabetes. The insulin level in (mu U/mL) is represented in figure 7e. Figure 7f shows the body mass index in kg/m^2. Despite having the same BMI, diabetic people showed significantly thicker triceps and biceps skinfolds than healthy ones. Men's average thickness is 12 mm, compared to women's 23 mm. The thickness of the triceps skin fold is shown in mm in figure 7g. While figure 7h shows how many people have diabetes in their family history. These Bar graphs depict how each feature and label is spread throughout multiple ranges, emphasizing the further necessity of scaling. Each discrete bar indicates that this is a categorical variable.

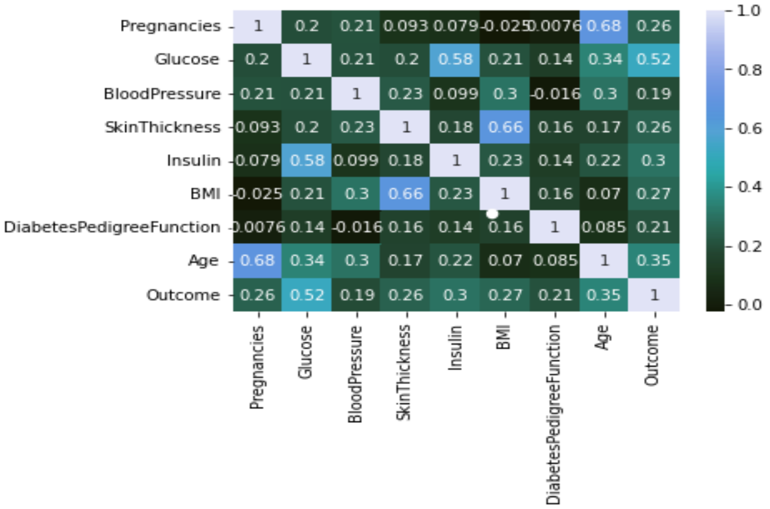
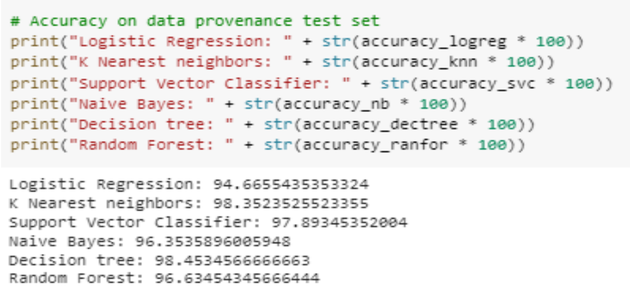


Figure 8: Co-relation Heat map of all 8 parameters

Figure 8 illustrates the correlation between all the features and shows which feature has the most impact on the outcome. The risk of being identified as a diabetic patient rises with an increase in the blood glucose level since there is a notable positive correlation between the glucose level and the outcome as shown in Figure 8. It is clear that no particular property has a strong correlation with our result value. Some of the attributes have a negative correlation, while others have a positive correlation with the outcome. By analyzing this correlation matrix, we came to the conclusion [Tang et al., 2006] that the four features listed below have a strong correlation with the outcome: Insulin level, blood sugar or glucose levels, and the patient's age Body Mass Index (BMI). These features can be chosen to accept user input and forecast the result.

Based on accuracy score metrics, we compare different machine learning methods [Uddin et al., 2019] during the model evaluation process and determine the mean accuracy for our model on data-provenance-oriented datasets. We have implemented

Naive Bayes, K Nearest Neighbors, Support Vector Classifier, Logistic Regression, Decision Tree, and Random Forest machine learning algorithm.

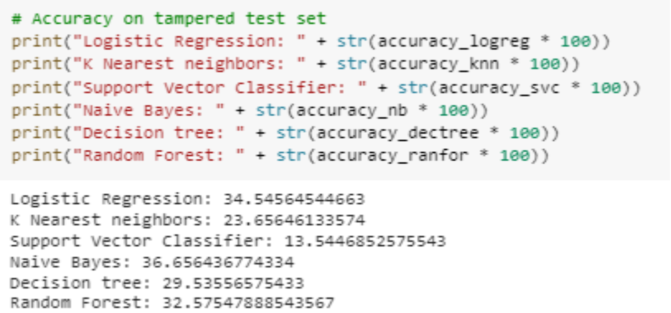


```
# Accuracy on data provenance test set
print("Logistic Regression: " + str(accuracy_logreg * 100))
print("K Nearest neighbors: " + str(accuracy_knn * 100))
print("Support Vector Classifier: " + str(accuracy_svc * 100))
print("Naive Bayes: " + str(accuracy_nb * 100))
print("Decision tree: " + str(accuracy_dectree * 100))
print("Random Forest: " + str(accuracy_ranfor * 100))
```

```
Logistic Regression: 94.6655435353324
K Nearest neighbors: 98.3523525523355
Support Vector Classifier: 97.89345352004
Naive Bayes: 96.3535896005948
Decision tree: 98.4534566666663
Random Forest: 96.63454345666444
```

Figure 9: Accuracy of diabetes prediction model on immutable data

Figure 9 highlights our model accuracy using the above-mentioned algorithms and it is clear that our model accuracy mean is more than 96%. In the next step, we considered big data without provenance and metadata information as it was not based on the blockchain and was open for mutation. The records do not have associated lineage information and the artificially introduced modifications can not be traced back to their source. Thus the records have been tampered and we do not have confidence in the new big dataset. Our new accuracy metric also reflects the same idea.



```
# Accuracy on tampered test set
print("Logistic Regression: " + str(accuracy_logreg * 100))
print("K Nearest neighbors: " + str(accuracy_knn * 100))
print("Support Vector Classifier: " + str(accuracy_svc * 100))
print("Naive Bayes: " + str(accuracy_nb * 100))
print("Decision tree: " + str(accuracy_dectree * 100))
print("Random Forest: " + str(accuracy_ranfor * 100))
```

```
Logistic Regression: 34.54564544663
K Nearest neighbors: 23.65646133574
Support Vector Classifier: 13.5446852575543
Naive Bayes: 36.656436774334
Decision tree: 29.53556575433
Random Forest: 32.57547888543567
```

Figure 10: Accuracy of diabetes prediction model on tampered data

Figure 10 shows the sudden drop in accuracy from 96% to 26%. Thus because of tampering and modifications, big data is of little use and no accurate prediction or analysis can be done via it.

8 Result and Analysis

In the first instance, we sought to research how the potential for analysis and prediction of big data might be impacted by tampering and manipulation. Reliable data is necessary for drawing useful conclusions from Big data. Due to the different streams of data that are input into data storage and the numerous transformations that Big data goes through during processing, there are several possibilities for errors to be introduced, either purposefully or unintentionally. Through our experimentation, we may infer that under our blockchain-based system, the unchangeable and tamper-proof metadata connected to the source and evolution of records gave verifiability to acquired data.

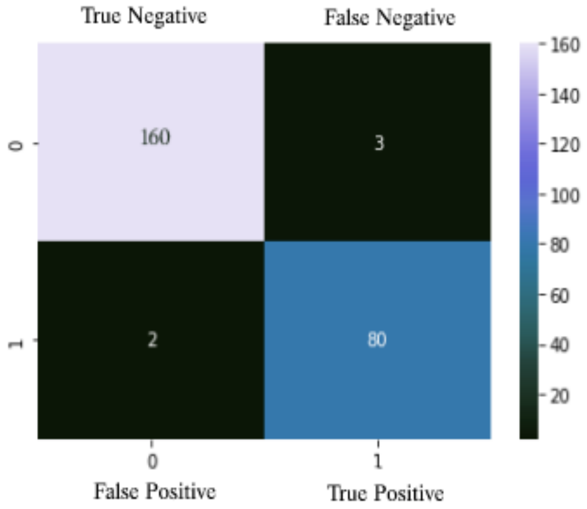


Figure 11: Classification report of immutable data

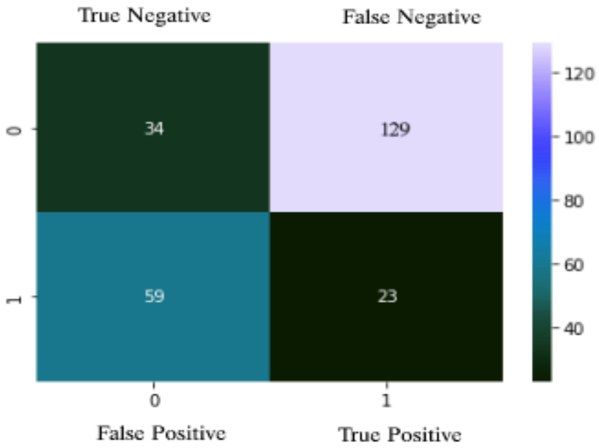


Figure 12: Classification report of tampered data

Figure 11 shows that just 5 out of 245 predictions were erroneous. 3 diabetic patients were identified as healthy and 2 healthy patients were predicted to have diabetes. Thus overall accuracy of the blockchain-based big data provenance-oriented model has more than 96% accuracy. About 179 out of 245 predictions were completely inaccurate in tampered data (recorded without blockchain), according to Figure 12. 59 healthy patients were predicted to have diabetes, and 129 diabetic patients were diagnosed as healthy. The tampered big data's accuracy has decreased from 96 to 26 percent as a consequence.

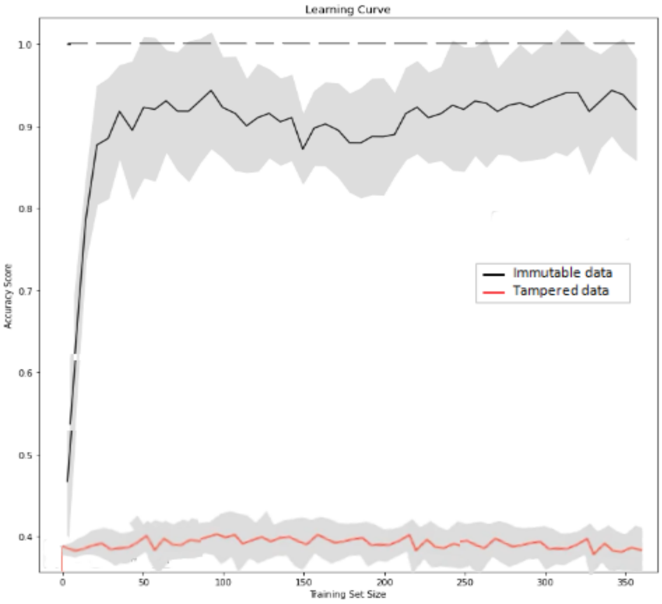


Figure 13: The learning curve of immutable and tampered data

In Figure 13 the learning curve of both immutable and tampered records is represented. We can monitor the evolution of our model on both datasets (i.e tampered and immutable) using the learning curve as it is a metric of predictive performance. There is a wide difference in the curve of both experiments. Over the three hundred and fifty records the immutable dataset model (recorded via blockchain) has progressed and approached an accuracy of more than 0.9 however the tampered data accuracy was even below 0.4.

9 Conclusion

Blockchain technology can fundamentally alter how Big Data is handled, evaluated, and analyzed. The methodology for deploying data of patients consists of more than three hundred records and access control on a blockchain platform is presented in this paper. We have also examined how Big data without blockchain can be a big challenge and a big buck. Blockchain-based data is more organized, vast, and provides useful analytics using immutable and tampered records. Blockchain is helping people to build their trust in the data they see. Blockchain in Big data will solve problems like who is in charge of the infrastructure when numerous actors are involved, and how reliable the data is leading to an increment in the accuracy of prediction models.

References

[Abdullah et al., 2017] Abdullah, N., Hakansson, A., Moradian, E.: Blockchain-based approach to enhance big data authentication in a distributed environment. In 2017 Ninth international conference on ubiquitous and future networks (ICUFN) (2017), pp. 887-892. IEEE.

[Al-Mamun et al., 2018] Al-Mamun, A., Li, T., Sadoghi, M., Zhao, D.: In-memory blockchain: Toward efficient and trustworthy data provenance for HPC systems. In: 2018 IEEE International Conference on Big Data (Big Data) (2018), pp. 3808-3813, IEEE.

[Appelbaum 2016] Appelbaum, D.: Securing big data provenance for auditors: the big data provenance black box as reliable evidence. Journal of emerging technologies in accounting.Vol. 13, No. 1 (2016), pp.17-36.

[Bandara et al., 2018] Bandara, E., Ng, W. K., De Zoysa, K., Fernando, N., Tharaka, S., Maurakirinathan, P., Jayasuriya, N.: Mystiko—blockchain meets big data. In: 2018 IEEE international conference on big data (big data) (2018), pp. 3024-3032. IEEE.

[Bhosale and Gadekar 2014] Bhosale, H. S., Gadekar, D., P.: A review paper on big data and hadoop. International Journal of Scientific and Research Publications, Vol. 10, No. 10 (2014), pp:1-7.

[Buneman et al., 2001] Buneman, P., Khanna, S., Wang-Chiew, T.: Why and where: A characterization of data provenance. In: International conference on database theory (2001), pp. 316-330, Springer, Berlin, Heidelberg.

[Buneman et al., 2006] Buneman, P., Chapman, A., Cheney, J.: Provenance management in curated databases. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data (2006), pp. 539-550.

[Chattu 2021] Chattu, V. K.: A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health. Big Data and Cognitive Computing, Vol. 5 No.3 (2021), p.41.

[Dai et al., 2008] Dai, C., Lin, D., Bertino, E., Kantarcioglu, M.: An approach to evaluate data trustworthiness based on data provenance. In: Workshop on Secure Data Management (2008), pp. 82-98. Springer, Berlin, Heidelberg.

[Dai et al., 2018] Dai, M., Zhang, S., Wang, H., Jin, S.: A low storage room requirement framework for distributed ledger in blockchain. IEEE Access. Vol. 6 (2018), pp. 970-975.

[Gao et al., 2018] Gao, Y. L., Chen, X. B., Chen, Y. L., Sun, Y., Niu, X. X., Yang, Y. X.: A secure cryptocurrency scheme based on post-quantum blockchain. IEEE Access. vol. 6 (2018), pp. 205–27

[Hogan and Helfert 2019] Hogan, G., Helfert, M.: Transparent Cloud Privacy: Data Provenance Expression in Blockchain. In: CLOSER (2019), pp. 430-436.

[Ismailisufi et al., 2020] Ismailisufi, A., Popović, T., Gligorić, N., Radonjic, S., Šandi, S.: A private blockchain implementation using multichain open source platform. In: 2020 24th International Conference on Information Technology (IT) (2020), pp. 1-4. IEEE

[Joshi and Dhakal 2021] Joshi, R. D., Dhakal, C. K.: Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health, Vol. 18 No. 14 (2021), p. 7346.

[Karafiloski and Mishev 2017] Karafiloski, E., Mishev, A.: Blockchain solutions for big data challenges: A literature review. In: IEEE EUROCON 2017-17th International Conference on Smart Technologies (2017), pp. 763-768, IEEE.

[Kiayias et al., 2017] Kiayias, A., Russell, A., David, B., Oliynykov, R.: Ouroboros: A provably secure proof-of-stake blockchain protocol. In Annual international cryptology conference (2017), pp. 357-388, Springer, Cham.

[Kosba et al., 2016] Kosba, A., Miller, A., Shi, E., Wen, Z., Papamanthou, C.: Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In: 2016 IEEE symposium on security and privacy (SP) (2016), pp. 839-858. IEEE.

[Li 2018] Li, S.: Application of blockchain technology in smart city infrastructure. In: 2018 IEEE international conference on smart internet of things (SmartIoT) (2018), pp. 276-2766, IEEE.

[Liang et al., 2017] Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., Njilla, L.: Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (2017), pp. 468-477). IEEE.

[McConaghy et al., 2016] McConaghy, T., Marques, R., Müller, A., De Jonghe, D., McConaghy, T., McMullen, G., Henderson, R., Bellemare, S., Granzotto, A.: Bigchaindb: a scalable blockchain database. white paper, BigChainDB. (2016).

[Ruan et al., 2019] Ruan, P., Chen, G., Dinh, T. T., Lin, Q., Ooi, B. C., Zhang, M.: Fine-grained, secure, and efficient data provenance on blockchain systems. Proceedings of the VLDB Endowment. Vol.12 No. 9 (2019), pp. 975-88.

[Sagiroglu and Sinanc 2013] Sagiroglu, S., Sinanc, D.: Big data: A review. In: 2013 International conference on collaboration technologies and systems (CTS), (2013), pp. 42-47. IEEE.

[Sahoo et al., 2018] Sahoo, M. S., Baruah, P. K.: HBasechainDB–a scalable blockchain framework on Hadoop ecosystem. In: Asian Conference on Supercomputing Frontiers (2018), pp. 18-29. Springer, Cham.

[Syed et al., 2013] Syed, A., Gillela, K., Venugopal, C.: The future revolution on big data. Future.Vol. 2, No. 6 (2013), pp. 2446-51.

[Tang et al., 2006] Tang, J., Alelyani, S., Liu, H.: Feature selection for classification: A review, Data Classif AlgorAppl, vol. 97, no. 7 (2006), pp. 1660-1674.

[Tosh et al., 2019] Tosh, D., Shetty, S., Liang, X., Kamhoua, C., Njilla, L. L.: Data provenance in the cloud: A blockchain-based approach. IEEE consumer electronics magazine Vol. 8 No. 4 (2019), pp. 38-44.

[Tan et al., 2020] Tan, L., Shi, N., Yang, C., Yu, K.: A blockchain-based access control framework for cyber-physical-social system big data. IEEE Access. Vol. 8 (2020), pp. 77 215–77 226.

[Treleaven et al., 2017] Treleaven, P., Brown, R. G., Yang, D.: Blockchain technology in finance. Computer Vol. 50 No. 9 (2017), pp:14-17.

[Uchibeke et al., 2018] Uchibeke, U. U., Schneider, K., A., Kassani, S. H., Deters, R: Blockchain access control ecosystem for big data security. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPS Com), and IEEE Smart Data (SmartData) (2018), pp. 1373-1378. IEEE.

[Uddin et al., 2019] Uddin, S., Khan, A., Hossain, M. E., Moni, M.A.: Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, Vol. 19 No. 1, (2019), pp.1-16.

[Xiao et al., 2018] Xiao, J., Lou, J., Jiang, J., Li, X., Yang, X., Huang, Y.: Blockchain architecture reliability-based measurement for circuit unit importance. IEEE Access. vol. 6 (2018), pp. 326-334.

[Yang et al., 2018] Yang, J., Lu, Z., Wu, J.: Smart-toy-edge-computing-oriented data exchange based on blockchain. Journal of Systems Architecture. vol. 87 (2018), pp. 36–48.

[Zheng et al., 2016] Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., Xiang, W.: Big data-driven optimization for mobile networks toward 5G. IEEE network. Vol. 30, No. 1 (2016), pp. 44-51.

[Zheng et al., 2018] Zheng, Z., Xie, S., Dai, H. N., Chen, X., Wang, H.: Blockchain challenges and opportunities: A survey. International journal of web and grid services.Vol. 14, No. 4 (2018), pp. 352-75.