Blockchain and Machine Learning in Health Care and Management

Sankarr Jain
Department of Electrical and Electronics Engineering
Ramaiah Institute of Technology
Bangalore,India
sanskar329@gmail.com

Aditya Anand
Department of Electronics and Instrumentation Engineering
Ramaiah Institute of Technology
Bangalore,India
aditya.anand556@gmail.com

Aman Gupta
Department of Information Science and Engineering
Ramaiah Institute of Technology
Bangalore,India
gupta.aman1602@gmail.com

Kushagra Awasthi
Department of Electronics and Communication Engineering
Ramaiah Institute of Technology
Bangalore,India
kushagra.awasthi33@gmail.com

Sarvesh Gujrati
Department of Mechanical Engineering
Ramaiah Institute of Technology
Bangalore,India
srvshgujarati3@gmail.com

Janamejaya Channegowda
Department of Electronics and Electrical Engineering
Ramaiah Institute of Technology
Bangalore,India
jc@msrit.edu

***Abstract*— Today we have enormous amount of data available in every sector, with the advent of technology available, it is possible to provide solutions to many problems. In this paper we are going to provide solutions to the problems related to healthcare data management using Machine Learning and Blockchain. Extracting only the relevant information from the data is possible with the use of Machine Learning. This is done using trained algorithms. Once this data is stored, the next problem is Data sharing and its reliability. This is where Blockchain comes into picture. The consensus in Blockchain technology makes sure that data is legitimate and transactions are secure. Blockchain technology can potentially change health care management for the better by placing patient at the epicentre of the healthcare system and increasing the privacy and interoperability of health data. This paper focuses primarily on solving healthcare data management problems by using Blockchain technology and including some indispensable features using Machine Learning.**

Keywords*— *Bag of words, blockchain, Electronic Health Records (EHR), Machine Learning, Social Security Numbers (SSNs).

I. INTRODUCTION

In Blockchain like a distributed ledger, singular transactions are encrypted into blocks by the applicable encryption, added to the ledger and never deleted. The information in Blockchain is verified fundamentally by a linked list of encoded exchanges that utilizes a hash. The hash function generates a hash by encrypting the information fed in Blockchain. It shapes the foundation of a decentralized medicinal service stage shared by the patients and suppliers, acting as an interface to the patient's record. [1]

Blockchain is a cryptographically secured, immutable, write once, read anywhere type data structure. It consists of blocks and theses blocks are linked together using an unmodifiable key referencing mechanism.

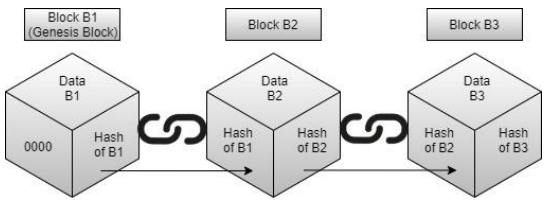

Fig.1 Blockchain Networking model

The Blockchain data structure consists of the following components:

- The Blockchain network has secured list of blocks which contains the useful information.
- A peer-to-peer network which contains identical examples of the Blockchain data structure
- A consensus mechanism which secures the harmonized growth of Blockchain.
- A security mechanism that ensures that the data sto red in the Blockchain network is immutable.

II. MEDICAL DATA BREACH

Reportedly, In April 2019, healthcare data breaches reached a record high and for the past few years, the healthcare field has had the second highest number of breaches compared to other sectors. In total, almost 10 million records were exposed in US healthcare sector in 2018 alone. The frequency of medical data breaches has been highly concerning. In particular, armed with someone's medical information, thieves can easily commit medical identity theft to get drug prescriptions, or make false insurance claims under the victim's name. Medical data mostly comes with personal and private information which includes Social **Security Numbers (SSNs),** as well as financial information. These pieces of data can be easily sold on the black market or used for malicious crimes, including credit card fraud and identity thefts. Criminal with such sensitive information can also file a fraudulent tax returns using the Social Security Number. [3]

The healthcare data breach statistics that was issued by the Department of Health and Human Services' Office for Civil Rights, U.S., clearly indicates the threat of data breach in the present scenario. This data was compiled for the decade starting from 2009 to 2018. The healthcare data breach statistics as shown in the above figure shows that the year 2018 has seen the highest amount of data breach. Also, in the last decade there has been about **2,546** healthcare data breaches involving more than **500** records, which was the cause of exposure of not less than **189,945,874** healthcare records equating to more than **59%** of the population of the U.S. [3]. The healthcare data breach statistics as shown in the above figure shows that the year 2018 has seen the highest amount of data breach

Proceedings of the International Conference on Mainstreaming Block Chain Implementation (ICOMBI) 2020
978-93-5406-901-7 © 2020 IEEE

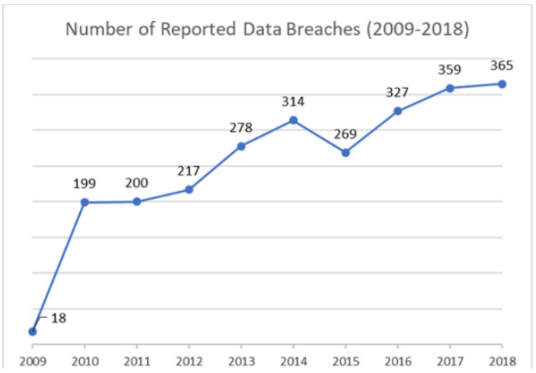

Fig.2 Number data breaches over the years [4]

Also, in the last decade there has been about **2,546** healthcare data breaches involving more than **500** records, which was the cause of exposure of not less than **189,945,874** healthcare records equating to more than **59%** of the population of the U.S. [3]. The next statistics shown in fig 3 indicates that Hacking and IT incidents have caused majority of these data breaches.

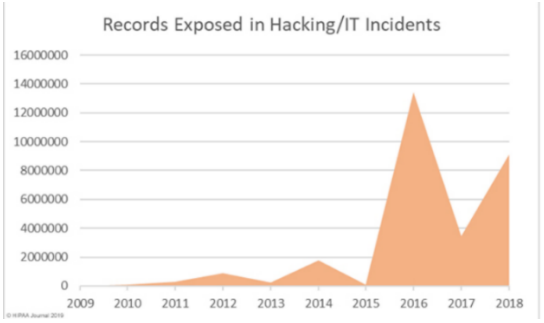

Fig.3 Number of hacking incidents in the past decade [4]

In many cases, leakage of data can prove to be life threatening and may cause serious safety problems to many innocents. In India specifically, there have been many incidents of medical report leakage. Some of them are listed here:

- In April 2018, Andhra Pradesh government websites were leaking Aadhaar number and sensitive information like reproductive history from pregnancy to delivery and details about abortions and so on, The Times of India reported.
- In 2018, an unsecured Andhra Pradesh government website exposed the names and contact numbers of every person who bought medicines which also includes the customers who bought medicine for erectile dysfunction from government-run stores. The dashboard on the Anna Sanjivini website unintentionally grated access to details which includes the names and phone numbers of every individual who bought medicines from every store. [5]

When sensitive information is leaked online, it can lead to malevolent consequences, also having the potential of increasing the crime rate significantly.

III.DECISION MAKING ERROR

Another problem is the transfer of data. clinicians heavily rely upon suitable clean data to make decisions regarding possible treatments that can resolve the health issue in the least possible time. It is important to have a secure means to transfer patient's historical data to hasten the treatment process.

In literature, studies have been carried out to estimate the number of deaths related to medical error. The **American Johns Hopkins Hospital by Makary et al** have concluded that deaths due to Medical Error have significantly increased over the years in USA. These statistics are shown in Fig.4

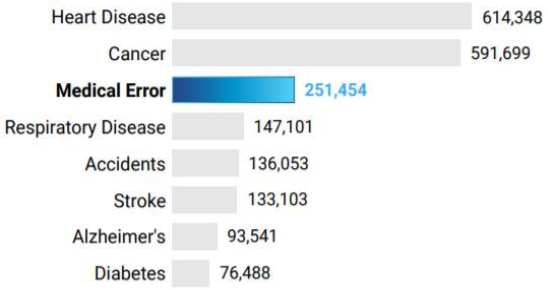

Fig.4 Leading cause of deaths in USA [6]

IV. STORING MEDICAL INFORMATION

Though there has been some advancement in the way medicals reports are stored, their transactions and security are still not up to the mark. Health data breaches are frequent and with the subsequent increase of hackers all around the globe, hacking of these data has become predominant, overshadowing the other breaching methods.

Meditab, a software company, claims itself as one of the foremost electronic medical records software creators for medical institutions. The company includes electronic faxes and this is method is still used significantly to share patient data to other data seekers. But this method of sharing data has proved to be insecure and unreliable making it very less preferable.

Spider Silk**,** a Dubai-based cybersecurity firm, had a fax server, which was running an Elastic search database which included not less than 6 million health care records .The server did not even had passwords for security, which indirectly granted access to everyone and thus anyone could read the transmitted faxes in real-time .The fax that were sent consisted of many recognizable information regarding patients such as their medical history, treatment undertaken in the past, their Social Security Numbers and other records. [7]

V.ELECTRONIC HEALTH RECORD [EHR] AND ITS LIMITATIONS

Keeping in mind the issues with Health record management that was used earlier, **Electronic Health Record (EHRs)** were introduced. EHR's are electronic version of a patient's medical history, which is maintained by the provider over time, and includes all important organisational clinical data which concerns the patient.

EHRs consist of several information management strategies for various purposes. There have been numerous data breaches cases reported which thereby confirms that the security of these reports has to be increased further.

There have been many cases where EHRs have failed in the area of security. Listed below are some incidents which proves this point. [8]

- Patient records were leaked in University of Michigan Medical Center patient records were disclosed on the internet the password.
- In Florida, a public health employee was able to access thousands of names of HIV infected patients and he sent it to many local news stations
- The University of Washington Medical network facility was accessed by unauthorized personnel and patient data was compromised
- Medical centres in New York and Holland and their weakness in protecting patient data was exposed
- A group of researchers from University of Minnesota inadvertently released names of kidney donors to the public.

Exposing a patient's private information to the general public can have disastrous effects on the reputation and mental health of the patient. This sort of information leak can also lead to financial losses to the individual involved.

The cases mentioned clearly indicates the unintended problem that comes up with the usage of EHRs. The Health care model proposed in this paper deals with many of these problems and provides meaningful solutions. [9]

In our proposed model we provide solution to almost all of the problems mentioned above.

VI. OUR PROPOSED MODEL

In our proposed model we have used both Blockchain technology and Machine Learning Algorithms to provide a better solution in terms of security. By using Machine Learning, we provide additional features which can be the base of ideas for further implementations on this subject.

Machine Learning is based on the concept of centralization of data, while Blockchain technology uses decentralization of data to provide high security. In this paper we have tried to project our model which showcases how we can use both these for this particular application.

A. Implementation of Blockchain

In this model, dual Blockchain structure is used, the first part grants access to health data and is built using the Hyperledger Fabric. The second part of the structure works on Ethereum and performs all application and services.

Medical information is very sensitive and personal so a closed Blockchain such as Hyperledger Fabric helps in retaining necessary privacy required.

Majorly blockchains are classified as public Blockchains and permissioned Blockchains.

This can be explained by considering the example of a user wants who to sell a book to person with some rebate and does not intend to tell about this to general public, the seller then can employ permissioned Blockchain to hide the information about the offer from the public. This model uses a double encryption mechanism on a permission-based Blockchain. The security that is provided by this model which uses Blockchain is beyond and far more advanced than any other centralized security system being used.

Furthermore, the patient's data is made inaccessible and unalterable. The Blockchain acts as a pointer, and provides the direction to the location of the stored data, meaning that anyone attempting to access patient data will be denied.

The health data is secured between the patient and the authorized doctor. When the authorised doctor adds additional information to the patients record history, the system will automatically update it. Only those clinicians who have authorized access can view the updates. None of the doctors are given permanent authorization, the access for the doctor ends when the patient wants so that the doctor can no longer update the record or access it.

This is vital in scenarios where there is a need to change the doctor in charge, therefore, with the help of Blockchain the information transfer will be easy and secure. Issues associated with the transfer of information by medical institute employees is completely eliminated and there will be no more data leaks probable in this transaction process and also overseas transactions of information can be cost effective as compared to the conventional techniques to do the same.

In emergency situations when the patient is unconscious and unable to provide any sort of input on his health, it would be vital to have access to the patients' health records. This information while performing lifesaving surgeries as history of past medications and illnesses are crucial before performing any sort of major surgery.

B. Implementation of Machine Learning

There are two steps in building a new Machine learning model. The first step is to take in the dataset and adjust the model weights to increase the accuracy of the model. The second step is testing the Machine learning model on independent or new datasets for the accuracy of the model and thus validate the model and prevent overfitting of the model. An over fitted model is very good at a given dataset but is bad at hypothesizing for the given problem. The procedure of Machine learning in our proposed model has been depicted in Fig.5.

Once a Machine learning model has been trained using Supervised Learning it can be used to do various tasks such as prediction, classification on the untrained dataset. In the proposed model, we use the "Bag of Words" algorithm which will extract only the required dataset and ignore the various other things like the Name, Age, Address and other personal details of the patient to maintain the privacy.

Supervised Learning builds a mathematical model of the given dataset that contains both input and output data. Supervised Learning can be used for classification and regression.

Steps for Supervised Learning:

- First, the category of training set data is selected
- The set of input data and its complimentary outputs are collected.
- The accuracy of the trained function depends on how the input dataset is represented. Typically, the input dataset is

converted into a feature vector, which contains a number of features that describe the dataset. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

- We need to determine the corresponding algorithms according to the given dataset.
- Run the algorithm being trained on the collected training set.
- The output of the test dataset determines the accuracy of the algorithm.

There are four major issues to be considered in Supervised learning such as Bias- Variance, Trade Off, Function complexity and amount of training data, Dimensionality of the input space.

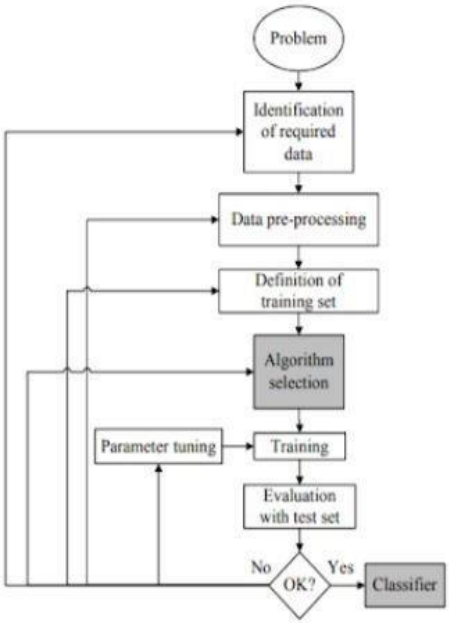

Fig.5 Supervised learning flowchart [10]

To get rid of some of the above issues we use the **"Bag of Words"** algorithm which is a method of extracting required features from text for use in modelling. The trained data contains terms in this algorithm and the correct features are selected from these words.
Upon defining the features, the feature is given a specific va lue, this feature representation is mapped to the training and test dataset. For the reflection of the BOW the two most common features value representation are:

1) Binary feature value:
The value of this representation feature can be either 0 or 1 where 1 indicates that the feature is in the dataset and 0 indi cates that the feature is not in the dataset.

2) Frequency Feature Value:
The value of a particular feature in this representation is determined by the number of times it appears in the datase t and if it did not appear, it is 0.

In this paper, we are using frequency feature value representation. A feature is considered important in this representation if it appear more than once in a sentence. Identifying and deleting phrases, that do not include illness or treatment information makes the data set more effective. The proposed algorithm extract from the HTML page only disease related information containing medical information and neglects other irrelevant materials such as forms, users, commentaries, navigation, menus, feedback and advertising. The benefit of using this method is that by using a weighted Bag of Words algorithm it improves precision with an accuracy of 79% to 82%. [11]

This model will take the dataset from reliable sources like Medline which provides the Machine learning model with the required information, which filters it into specific categories. Using health data of patients with diseases that are very rare is also of great importance to the medical scientists and researchers. Consider, for example, a case where there is an outbreak of a new disease in a particular area or country, about which there isn't much information available. Machine learning plays a vital role in such circumstances where the information produced by the health reports of the patient is fed in the form of data to the cloud without allowing access to the personal and private information of the patient which in turn will help the researchers and the doctors with some reliable and accurate information with better analysis of the symptoms and effects of the disease using the Machine learning technology.

When the health data of a lot of patients is fed to the machine, it provides enough statistical information about a particular disease or disorder. Moreover, categorization of diseases will make it more convenient for a patient to access and understand their respective health records and will also facilitate researchers in searching for the information.

It is important for the patients to have control over their health record and also benefit from the data that they share. Therefore, it is possible to find users willing to have their health data used for research. These patients will be given clear information about the exact purpose for which the information will be used and also what kind of data will be required. As the data is anonymized, there will be no issues of privacy. In return, participants can be compensated.

C. Blockchain and Machine Learning

Blockchain used along with Machine learning opens up a host of opportunities for securing health data. It is relatively strenuous task to combine both of them. There has not been significant work done in this area.

In this paper, we have used Supervised learning to train the Machine learning algorithm on the datasets obtained through various sources like MedLine, and to reduce the dimensionality of the data we will use the "bag of words" algorithm which will take only the necessary parts of the dataset to train the Machine learning model.

The Patient healthcare data is secured using Blockchain network through which transactions between patient and the authorized doctor is made. The functionality of Blockchain ends here.

A new set of patient healthcare data is then fed to the trained model which first filters the data and discards all the personal details and information, then this data is further categorized by diseases. The role of the Machine Learning terminates here.

This dataset from a hospital can be transferred to other hospitals by using e-mails or by uploading the data set from each hospital on a cloud-based storage system which can be accessed by hospitals and researchers.

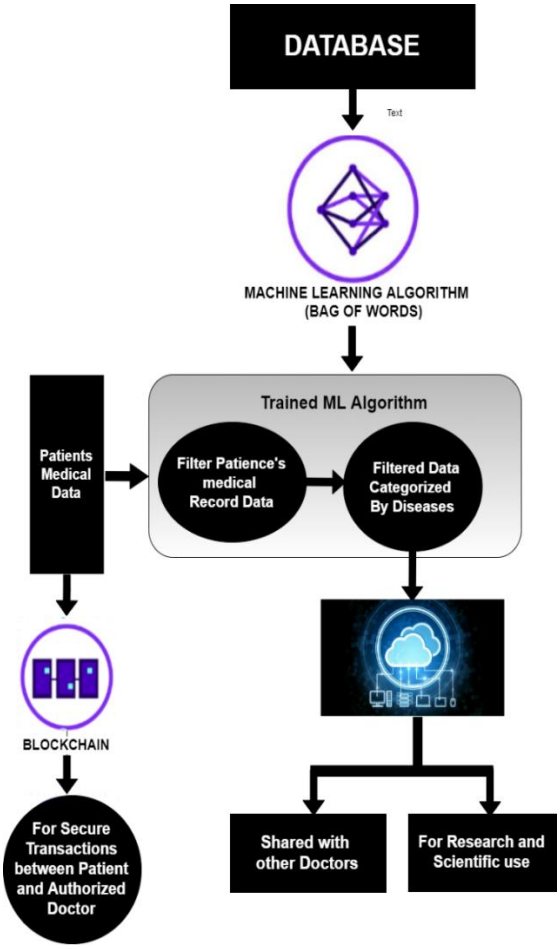

Fig.6 Implementation of the proposed model

This sharing of data is very important because the data for each hospital would be completely different. The biggest advantage of using Machine Learning algorithms is to make the health care data accessible to researchers and doctors for study purpose.

Thus, using this model can provide dual advantages with Blockchain providing the necessary security

of the healthcare record by allowing secure transactions between the doctor and patient, while the filtered data can be used by researchers and doctors because this data need not be secured as it is deprived of all personal information. This data is entirely built to be used for reference purpose.

The above application is performed by Machine Learning algorithm as discussed above.

VII. CONCLUSION

Blockchain Technology has been evolving with time, banking and financial sectors are already using Blockchain keeping in mind the unparalleled advantages it offers, with the significant increase in health data breach through hacking, and application of Blockchain for security becomes important and imperative. It will not be wrong to say that Blockchain based Health care model is the future in healthcare sector and has the potential to change the way health care records are managed and secured.

With the emergence of 5-G networks and faster than ever data transfer facilities it will encourage advancement of Machine Learning, Blockchain and other data based techniques in various sectors including Healthcare. As this new technology ecosystem emerges, Blockchain promises significant improvements in managing patient health records.

Continuous efforts are being made to increase the accuracy of wearable health tracking devices and if these data could provide more accurate and reliable results there will be brighter chances of integrating these devices with the health records to provide more information and also share some of these medical data securely with authorized doctor without actually visiting.

The ideas based on implementing Blockchain and Machine Learning is not much explored. Our paper proposes a unique healthcare model which is still in its infant stage but can surely form base to many more health care models to come.

There is further scope of improvising this idea by implementing Artificial Intelligence (AI), Internet of things (IoT) and much more available technology to develop a more comprehensive health care model in future.

REFERENCE

[1] Chen, Y.; Ding, S.; Xu, Z.; Zheng, H.; Yang, S. "Blockchain-Based Medical Records Secure Storage and Medical Service Framework", *Journal of Medical Systems*, vol.43, no. 5, 2018.

[2] G.Magyar,*Blockchain: Solving the privacy and research availability tradeoff for EHR data: A new disruptive technology in health data management,* Budapest, Hungary,24-25 Nov. 2017.

[3] Aimee O'Driscoll (2019, July), "The biggest medical data breaches in history",

a. [Online].Available: https://www.comparitech.com/blog/vpn-privacy/biggest-medical-data-breaches/

[4] HIPAA JOURNAL (2018,March) , "Healthcare Data Breach Statistics.",

a. [Online]. Available:
b. "https://www.hipaajournal.com/healthcare-data-breach-statistics/

[5] Zaheer Merchant (2019, April) , "Health department of northern state exposed data of 12.5 million pregnant women.",

a. [Online].Available: https://www.medianama.com/2019/04/223-health-department-indian-state-pregnant-women-data-leak/

[6] "Medicalchain Whitepaper - 2.1"

a. [Online].Available: https://medicalchain.com/en/whitepaper/

[7] Zack Whittaker (2019, March), "A huge trove of medical records and prescriptions found exposed.",

a. [Online]. Available:
b. https://techcrunch.com/2019/03/17/medical-health-data-leak/

[8] Centres for Medicare & Medicaid Services," Electronic Health Records",

a. [Online]. Available:
b. https://www.cms.gov › Medicare › E-Health › EHealthRecords

[9] K.T. Win. A review of security of electronic health records. Electronic Health Records: security, safety and archiving, 34, 2005

[10] .S. B. Kotsiantis,*Supervised Machine Learning: A Review of Classification Techniques,* Department of Computer Science and Technology University of Peloponnese, Greece.

[11] P. Shinde,S.Madhav,*Health Analysis System using Machine Learning*