

Novel Approach for Disease Prediction in Healthcare using Blockchain and Machine Learning

A M Siri¹, Kushala R², Meghana J P³, Prerana R Shetty⁴, Shruthi K⁵, A S Poornima⁶ and M B Nirmala⁷
¹⁻⁴UG Student, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

Email: kushalargowdaofficial@gmail.com

⁵Assistant Professor, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

Email: shruthik@sit.ac.in

⁶Professor, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

⁷Associate Professor, Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, India

Abstract— Today we have a tremendous measure of information accessible in each area, with the appearance of innovation accessible, giving answers for some problems is conceivable. In this venture we will give answers for the issues connected with medical care information and the executives utilising Machine Learning and Blockchain. Extricating just the vital data from the information is conceivable with the utilisation of Machine Learning. This is finished utilising prepared calculations. When this information is put away, the following issue is Data sharing and its dependability. This is where Blockchain comes into picture. The agreement in Blockchain innovation ensures that information is authentic and exchanges are secure. Blockchain innovation might possibly improve medical services by setting patients at the focal point of the medical care framework and expanding the protection and interoperability of wellbeing information. This paper centres fundamentally around taking care of medical services information the board issues by utilising Blockchain innovation and including a few key elements utilising Machine Learning.

Motivation : The main motive of this project is to bring the characteristics of Machine learning and Blockchain together. Machine learning plays a crucial role in detecting deadly diseases. On the other hand, blockchain technology can revolutionise medical database interoperability and restrict unauthorised access to records. This would ensure that the sensitive patient data is tightly secured.

Summary of the outcomes : This project aims to build a secure, ML driven healthcare management system which would ensure that the diseases are accurately predicted and categorised in the early stage. Further, it ensures that the trained model filters the data and discards all the personal details of the patient and protects it from data leaks and breaches. It powers the system with blockchain to secure the transactions between patients and the authorised doctor and give patients the flexibility to choose which doctor should view his health record and who should not.

Index Terms— Blockchain, Machine learning, Flask, Ethereum, Moralis, Ganache, Metamask.

I. INTRODUCTION

Today we have an enormous amount of data available in every sector, with the advent of technology available, it is possible to provide solutions to many problems. In the current world we come across new inventions and implementations every day. Extensive amounts of data is collected in healthcare organisations which is used to power Machine learning models and AI driven applications. While this is a benefit, there are certain certifiable computerised assault concerns that have emerged in clinical consideration organisations during the past couple of years. Issues such as data leak or revelation of sensitive information of patients are a major concern. This problem could be solved by building efficient systems using Machine Learning backed by Blockchain.

A. Machine Learning in Healthcare

Machine learning is a computer-based method for analysing free-form text or voice using a preset set of theories and technologies, such as linguistic and statistical methodologies, to extract rules and patterns from the data. Knowledge and experience are two very significant variables for physicians in terms of patient care; yet, humans are limited in terms of gaining knowledge through accumulating data, whereas machine learning excels in this area.

Once a Machine learning model is well trained with a large set of training data with high accuracy, it reduces the human dependency and human effort by automating the tasks. The efficiency of the model mainly depends on the algorithm used and the amount and type of data used for training. It is prefer to use as large a training dataset as possible. The variety of examples in the training set is important in maintaining the balance between different target classes. If the examples of different target classes are imbalances the model will tend to get inclined to the class with the majority of examples. The variety of training examples is important as it can help in cases where different testing data is observed.

There are two types of machine learning: supervised learning and unsupervised learning. Extracting only the relevant information from the data is possible with the use of trained algorithms. This is done using trained algorithms. Data mining has a very important role in extracting hidden insights from large clinical datasets that doctors collect regularly which has the details of diagnosis of patients symptoms, which helps them to decide the treatment to be given. Important set of symptoms specific to diseases are considered and a model is built on that. Using machine learning to predict disease in healthcare is quite challenging but would be of great help to doctors.

B. Blockchain in Healthcare

Earlier in Healthcare the patients data was stored in a centralised database where it was solely accessible. There are many disadvantages in the centralised system including centralisation of power, central point failure, unethical operations, scalability issues, trust issues, privacy issues. Patients' data should be accessible in a decentralised way. The advantage of the distributed database is that each participant has a copy of the database. In most cases, it's easier to control access and integrate your systems and processes in your own copy of a database. However, synchronisation of changes to each database is required. handling failures and conflicts can add complexity and data integrity issues.

Blockchain is a decentralised, peer-to-peer, append-only, distributed ledger where 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 utilises a hash. The hash function generates a hash by encrypting the information fed in Blockchain. It shapes the foundation of a decentralised medicinal service stage shared by the patients and suppliers, acting as an interface to the patient's record. It consists of blocks and these blocks are linked together using an unmodifiable key referencing mechanism.

Every patient and doctor acts as a node, each and every activity done by any node will be recorded as a transaction. These transactions are bundled together to form a block and these blocks will be mined by miners and blocks will be a part of blockchain. In blockchain each block stores Keccak 256-bit parent hash. Each block is linked by the next following block through previousHash. These blocks store the information in a secured way in an append-only database.

C. Consensus in Blockchain

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 patients at the epicentre of the healthcare system and increasing the privacy and interoperability of health data. The primary focus lies on solving healthcare data management problems by using Blockchain technology and including some

indispensable features using Machine Learning. Since we are using Ethereum blockchain in our project consensus mechanism we are using is Proof of Stake(PoS).

Proof of Stake(PoS) is defined as a cryptocurrency consensus mechanism for processing transactions and creating new blocks in a blockchain as well as it is a method for validating entries into a distributed decentralised blockchain and keeping the blockchain secure.

How Proof of Stake(PoS) is different from Proof of Work(PoW):

- In Proof of Stake (PoS), miners validate block transactions based on the number of coins a validator invests as stakes.
- Proof of Stake(PoS) helps to validate a blockchain and add new blocks in a more secure way than Proof of Work(PoW).
- Proof-of-stake (PoS) is seen as less risky in terms of an attack on the network.
- Proof of Stake(PoS) uses Merkle Patricia Tree which is used to store data and facilitates efficient insert/delete operations and key-value lookups are very efficient.

How does Proof of Stake Work :

When a transaction needs to be successfully done, that complete transaction needs to be shared by the validator who will be responsible for adding the transaction into the blockchain. Validators once got blocks to be proposed they need to broadcast to all nearby miners in one's network and later they all use one algorithm to propose new blocks. Here at least 128 members should validate the transaction, the group is called "committee" and the Byzantine Problem is solved. There will be a timestamp of 30sec to propose and validate the transaction within that timestamp target. In Proof-of-work, the miner who solves the cryptic puzzles the fastest gets to mine the new block and get rewarded for it, while the other competitors are left with computational loss and nothing in hand. Proof-of-stake eliminates this demerit and hence opens the competition to only those miners who can afford to put the currency in stake. This ensures survival of the fittest and doesn't let anyone else lose.

So the validators don't need to use significant amounts of computational power because they're selected at random and aren't competing. They don't need to mine blocks; they just need to create blocks when chosen and validate proposed blocks when they're not. This is known as attesting. If a malicious block is attested, the validator loses his stake.

It rules over Proof-of-Work as the latter had caught all eyes and has had a reputation for the massive amounts of computational power and electricity it consumes.

D. Integrating Machine Learning and Blockchain

Blockchain and machine learning technologies can mitigate healthcare issues such as slow access to medical data, poor system interoperability, lack of patient agency, and data quality and quantity for medical research. Blockchain technology facilitates and secures the storage of information in such a way that doctors can see a patient's entire medical history, but researchers see only statistical data instead of any personal information. Machine learning can make use of this data to notice patterns and give accurate predictions, providing more support for the patients and also in research related fields where there is a need for accurate data to predict credible results. The medical history and patient's sensitive data is stored in blockchain and using machine learning doctors and researchers can use the medical data of the patient to predict the severity of the disease and treat it at an early stage. The machine learning models built for a particular disease takes the symptoms and other parameters as input and outputs the severity of the disease in terms of classification.

II. METHODOLOGY

A. Machine Learning

Machine Learning Libraries

- Pandas - Python library for data manipulation and analysis.
- Numpy - Library that supports large dimensional array, matrix operations.
- Pickle - Module used for serialising and deserializing python object structure.
- Seaborn - Data visualisation library used for statistical graphs.
- Scikit-Learn - It features various classification, clustering and regression algorithms.

The datasets for various diseases were taken from kaggle. Initially preprocessing of data is done in order to replace missing values, handle outliers, and find correlation between the attributes. Then different Machine learning algorithms are applied to each dataset in order to decide the best suitable algorithm for each disease prediction based on the accuracy and confusion matrix obtained for each. Python files containing code of best predicting algorithm along with the dataset is used for the further step of the project.

Flask

- Flask is a micro web framework written in Python.
- It is used to develop web applications easily

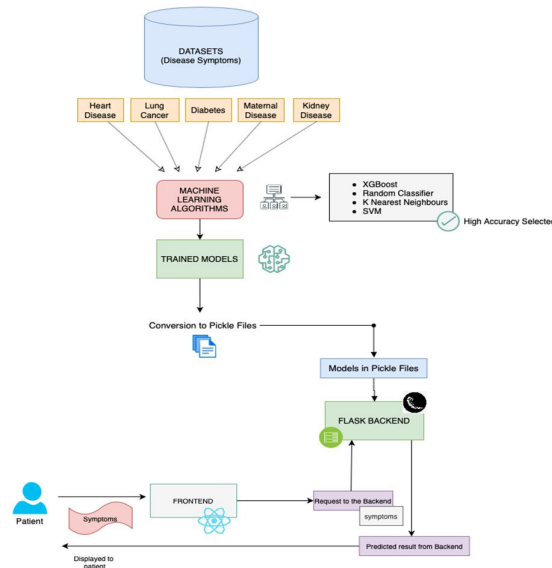


Fig. 1 Data Flow diagram of Machine learning

The frontend of the application is backed by React. The symptoms of a patient is taken through the user interface and is sent as POST request to the backend. In flask application, the respective routes are defined where the data from the patient is accepted and further operated to predict the disease. Each trained model associated with prediction of a particular disease is converted to a pkl file using the “pickle” module. The test data sent from the frontend is applied on these trained models and results are shown.

Algorithms used:

- Support Vector Machine:** It is a supervised machine learning algorithm used for both classification and regression. The main advantage of this algorithm is to find the best possible decision boundary that can separate the maximum points correctly into different classes. Among various possible decision boundaries, the best boundary is called Hyperplane.
 Consider data point :X
 Target label:Y
 if $Y_i (W \cdot X_i + b) - 1 = 0$:
 then (X_i, Y_i) is support vectors
 then save parameters W,b
 else if $Y_i (W \cdot X_i + b) - 1 > 0$:
 then save parameters W,b
 else if $Y_i (W \cdot X_i + b) - 1 < 0$:
 then update parameters W,b
- K Nearest Neighbor:** It is a simple classification algorithm. A new data point is classified on the basis of similarity with the nearest already classified points. Here, K is the number of nearest neighbours considered. Majority voting is usually done on the K points. It is an important factor to determine the most accurate value for K. It considers majority voting for classification instances and calculates average in case of regression
 - Consider a data point “x” to be classified.
 - Calculate $d(x, x_i)$ for $i=1,2,...,n$ where d indicates the Euclidean distance between the points considered.
 - Represent the n distances in ascending order.
 - Consider first k distances from the list where k is a positive integer.

4. Determine the k-points corresponding to these k-distances.
 5. Let k_j denote the number of points in the j th class among the classes the k points belong to i.e $k \geq 0$.
 6. If $k_i > k_j$ and $i \neq j$, then put the query data point, x in class i .
- **Random Forest Classifier:** It is an improved version of decision tree classifier. It contains many decision trees on various samples of available dataset and calculates the average to improve the prediction. It can be used for classification as well as regression. It considers majority voting for classification instances and calculates average in case of regression.
 1. Store the prediction of each decision trees

$P += 1$ if $X(i) = P$
 $N += 1$ if $X(i) = N$
 2. Compute the total votes for individual class.
 3. Declare majority class as final outcome.
 - **XGBoost Classifier:** Extreme Gradient Boosting classifier helps in fast and accurate classification. Boosting is a technique to enhance the accuracy of a model using multiple weak models. Sequential decision trees are created. Weights are assigned to each input variable, these variables are then fed to the tree.
 1. Load feature vector
 2. Save the individual scores in S
 3. Instantiate classifier as clf
 4. fit clf
 5. Generate FIs
 6. Determine FI threshold FIp
 - for $i=1$ to n do

if $(FI(X_i) \geq FIp)$ then
 append $FI(X_i)$ into S
 7. Use the scores in S to generate the feature vector

TABLE I. MACHINE LEARNING DATASET INFORMATION

Disease	Attributes	Target class
Heart	age, anaemia, creatinine_phosphokinase, diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time	Risk of heart : Low/High
Lung Cancer	gender, age, smoking, yellow_fingers, anxiety, peer_pressure, chronic_disease, fatigue, allergy, wheezing, alcohol_consuming, coughing, shortness_of_breath, swallowing_difficulty, chest_pain	Risk of Lung cancer: Low/High
Maternal Health	Age, SystolicBp, DiastolicBp, BS, BodyTemp, HeartRate	Risk of Maternal health: Low/Mid/High
Kidney	id, age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcv, wc, rc, htn, dm, cad, appet, pe, ane	Chronic Kidney Disease/ Not
Diabetes	Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age	Diabetic/Not

B. Blockchain

Blockchain Technology : A blockchain is a decentralised, distributed, and oftentimes public, digital ledger consisting of records called blocks that is used to record transactions across many computers so that any involved block cannot be altered retroactively, without the alteration of all subsequent blocks.

Ethereum : Ethereum is a platform powered by blockchain technology that is best known for its native cryptocurrency, called ether, or ETH, or simply ethereum. The distributed nature of blockchain technology is what makes the Ethereum platform secure, and that security enables ETH to accrue value.

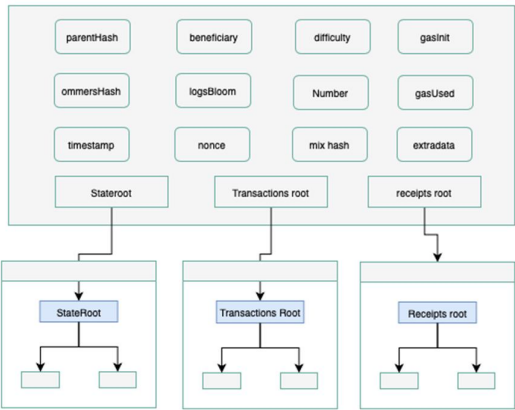
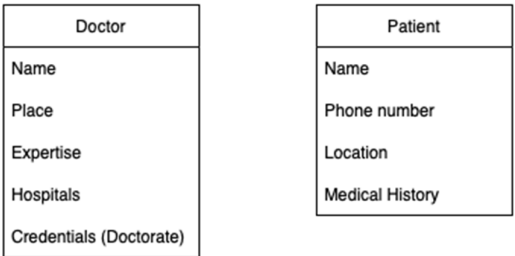


Fig. 2 Ethereum Block Structure

Block is the core primary unit of blockchain. Blocks contain transactions and some important information such as previous hash which ensures immutability and security in the blockchain network. Each block contains a previous hash sequentially so it is almost infeasible to tamper the data. Block will be created when a miner finds nonce, which is a one-time generated value after a lot of computation. It is necessary to transmit password securely and prevent replay attacks.



Hash of the Record stored

Fig. 3 Models that get stored in Blockchain

As shown in Fig. 3 the necessary information is collected. Smart contracts are written in a file with .sol extension. It consists of all the fields mentioned for both Doctor and Patient in doctor.sol and Patient.sol files respectively. Smart contracts are compiled and deployed. The deployed contracts are displayed in ganache by allotting a block for each contract along with transaction hash, contract address, block number, account, gas used, gas price, value sent and total ethers deducted from ganache.

Ganache : Ganache is a personal blockchain for rapid Ethereum and Corda distributed application development. You can use Ganache across the entire development cycle; enabling you to develop, deploy, and test your dApps in a safe and deterministic environment. Ganache comes in two flavours: a GUI and CLI.

Metamask : MetaMask is a free crypto wallet software that can be connected to virtually any Ethereum-based platform. This will allow you to store any assets you create or buy, as well as connect to any platform built on top of the Ethereum blockchain.

Moralis : Moralis is a leading web development platform that offers everything that the user needs to create, host, and grow great dApps in one place. It allows you to simply interface with an infinite number of external projects, chains, and technologies.

TABLE II. BLOCKCHAIN PROPERTY

Property	Value/Description
Consensus	Handled by set of nodes
Transaction Validation	Set of Authorised nodes
Transaction Reading	Any node or set of predefined node
Data Immutability	Could be tampered
Transaction throughput	high
Network Scalability	Low to Medium
Infrastructure	Decentralised
Features	• Applicable to tightly controlled business • Fee-free transaction • Laws on services are easier to manage • Effective defence from outside perturbations
Example	HyperLedger, Ethermint, Tendermint

Moralis is the fastest way to build and deploy dApps on Ethereum, BSC, Polygon, Solana, and Elrond. All Moralis dApps are cross-chain by default. Building on Moralis ensures that your dApp is future-proof. Even if new blockchains are invented, dApp will instantly work on any chain. Moralis provides a serverless database where the raw data is stored and the data is queried in a representable format using indexing.

The smart contracts are written using the solidity programming language for patient doctors to run on each node of the ethereum blockchain. Each smart contract will be compiled and deployed in the local blockchain. The local blockchain used is Ganache which consists of 10 different nodes with 100 Ethers each. The application is linked to the local blockchain through the web3 javascript connector and the chrome extension metamask. Whenever a node(patient or doctor) logs in, the transaction will be recorded and the activity done by that node will be stored as a transaction inside a block by each and every node of the blockchain.

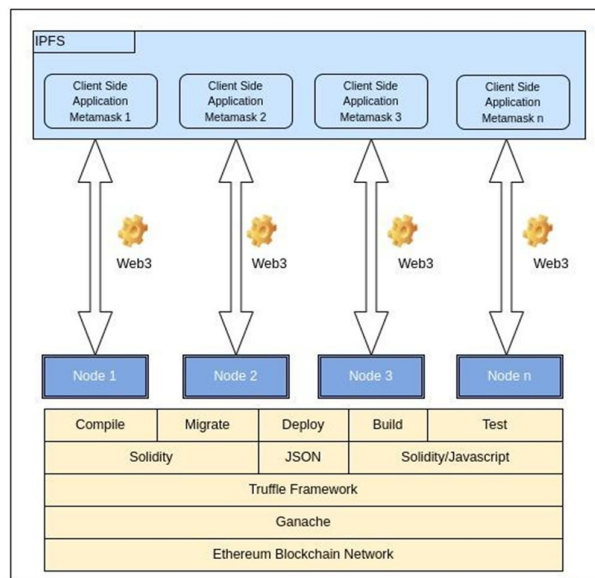


Fig. 4 Use case diagram of Blockchain

C. Blockchain and Machine Learning

Hospital registration is made based on the data of the hospital like hospital name, mail id, phone number, address, year of establishment and password. Entire data is stored in a decentralised database called Moralis. Authorised registered hospitals register patients and doctors considering necessary credentials. Blockchain a decentralised and distributed ledger permits the registered hospitals to access the needed patients data.. Credentials for patients include patient name, phone number, location, aadhar number, email id, medical history, symptoms and age. Credentials for doctors include name, phone number, location, email id, registration number, registration council, specialisation, experience. The entire data is stored in the Blockchain in a decentralised and secured way. The doctors registered by every hospital are visible across the network and the patient data is accessible to the authorised doctor. This proves the decentralised property of blockchain.

The data entered in the frontend part will be stored using smart contracts written separately for doctors and patients. Metamask confirmation is mandatory for the registration and a certain number of ethers would be deducted for every transaction. Soon after displaying doctors to all hospitals then any patient who wants to consult a doctor needs to consult hospital admin such that he/she will request doctor for consultation. Once a doctor accepts the request he/she can look into the medical history of patients and treat them accordingly. After treatment, the doctor can send a prescription to the patient who consulted him. Later patients can view the prescription provided.

Simultaneously the patient can upload his previous medical history report. Every doctor and patient acts as a node. Here decentralisation is achieved such that every hospital can access the registered doctors and consult them in no time. If any serious issues happens with a patient he can go to any of nearby hospitals and get his treatment done since patient medical history is made easily available through decentralisation.

In machine learning, separate models are created for major diseases like Heart, Diabetes, Kidney, Lung cancer, Maternal health. Doctors will login to this page and according to the tests performed, the particular model will be accessed. Based on the set of symptoms trained models will predict severity of the disease. These Machine learning models with high accuracy will reduce the human dependency, effort and time required. Reports will be generated once the prediction is done. Doctors can generate the reports for further processes.

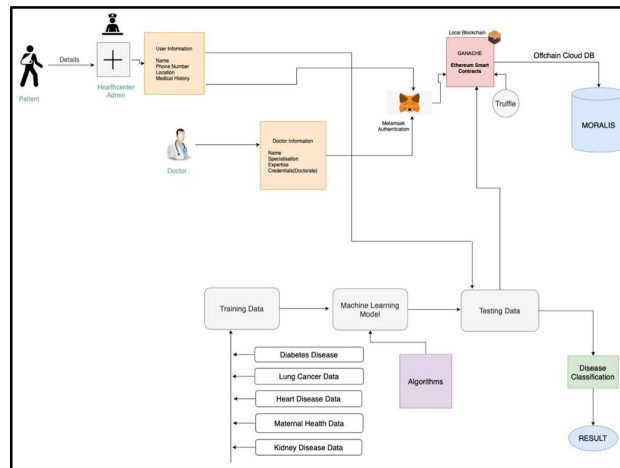


Fig. 5 Workflow

III. RESULTS

- Various diseases according to the set of symptoms are accurately predicted using trained Machine learning models in the early stage and categorised to help researchers keep track of new variants.
- Ensure that the trained model filters the data and discards all the personal details and information of the patient and protects it from data leaks and breaches.
- Power the system with blockchain to secure the transactions between patients and the authorised doctor and give patients the flexibility to choose which doctor should view his health record and who should not.

Machine Learning predictors for Lung cancer, Diabetes, Heart Disease, Kidney Disease and Maternal Health. Blockchain smart contracts are compiled and deployed in local blockchain ganache.

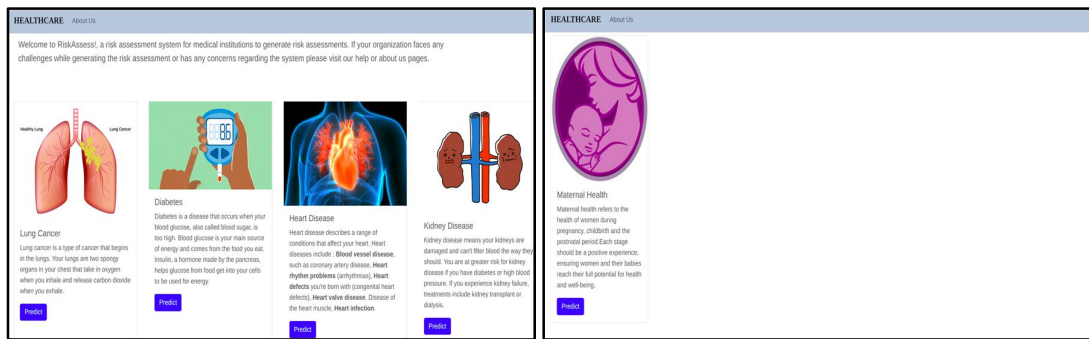


Fig.6 Home page

Fig.7 Home page

gender	0/1
Age in Years	eg. 34
smoking	1-no, 2-yes
yellow_fingers	1/2
Activity	1/2
peer_pressure	1/2
chronic_disease	1/2
fatigue	1/2
allergy	1/2
wheezing	1/2
alcohol_consuming	1/2
coughing	1/2
shortness_of_breath	1/2
swallowing_difficulty	1/2
chest_pain	1/2
Predict	

Fig. 8 Lung cancer Predictor

Age in Years	eg. 34
bp	1-no, 2-yes
sg	
al	
su	
bgr	
bu	
sc	
sod	
pot	
hemo	
pcv	
wc	
rc	
Predict	

Fig. 9 Kidney Disease Predictor

Age in Years	eg. 34
Anaemia	0/1
creatinine phosphokinase	3/4 digit
Diabetes	0/1
ejection_fraction	10-40
high_blood_pressure	0/1
platelets	eg. 2.27, 000.00
serum_creatinine	1.9-.2.5
serum_sodium	110-310
sex	0/1
smoking	0-no, 1-yes
time	4.5.6..
Predict	

Fig. 10 Heart Disease Predictor

Age in Years	34
systolicbp	0
diastolicbp	0
bs	1
body_temperature	1
heart_rate	1
Predict	

Fig. 11 Maternal Health Predictor

pregnancies	male-0
glucose	85..197
blood_pressure	50..98
skin_thickness	29..45
insulin	2/3 digit
bmi	2 digit
diabetes_pedigree_function	eg. 0.627
age	
Predict	

Fig. 12 Diabetes Predictor



Fig. 13 Predictor Result

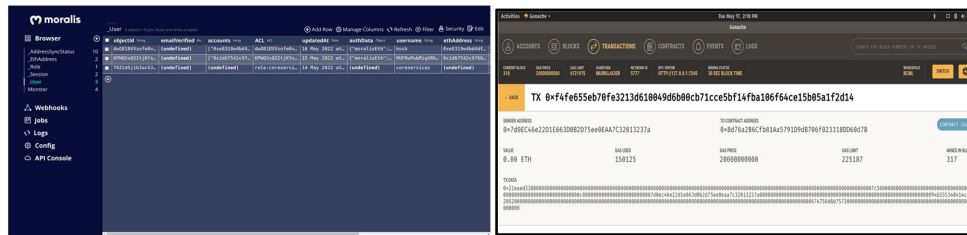


Fig. 14 Storing Data in Offchain Moralis

Fig. 15 Transaction in Ganache

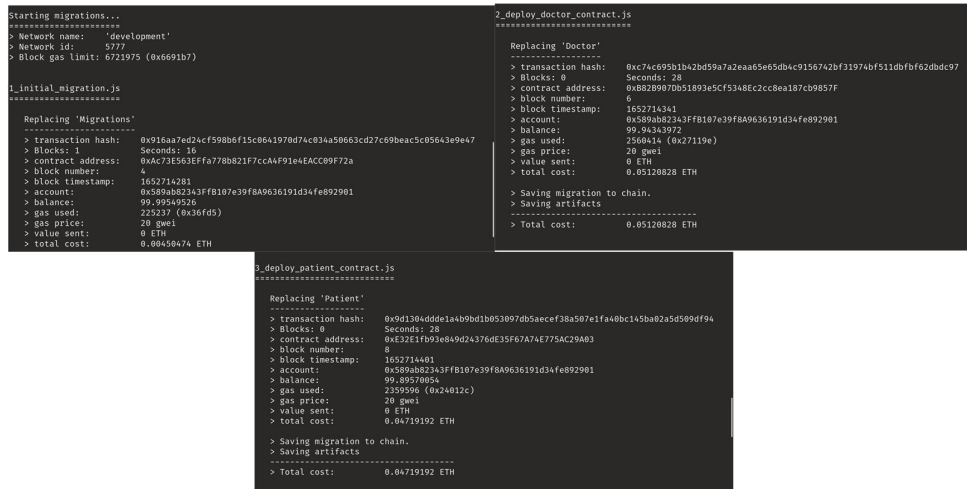


Fig. 16 Deploying contracts in blockchain

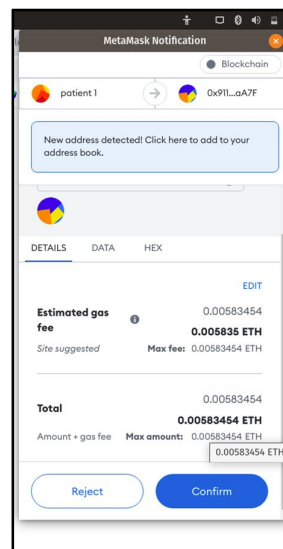


Fig. 17 Metamask Confirmation for a Transaction

IV. CONCLUSION

Blockchain Technology has been developing with time, banking and monetary areas are now utilising Blockchain maintaining as a primary concern the unrivalled benefits it offers. With the critical expansion in wellbeing information break through hacking, and use of Blockchain for security becomes significant and imperative. Machine learning then again has turned into a required innovation of great importance because of the

