Connecting a World of Healthcare Data using Machine Learning and Blockchain

A Project Report
Presented to
The Faculty of the College of
Engineering
San Jose State University
In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Computer Engineering
Master of Science in Software Engineering
By
[Kavya Dayananda, Jason Gonsalves, Priyanka Kumar, Deeksha Mahalingappa]
[May 2020]

APPROVED

Nima Karimian, Project Advisor

David Bruck, Director, MS Computer Engineering

Dan Harkey , Director, MS Software Engineering

ABSTRACT

Personalized medicines and emergency treatment of ailments have been simplified due to automated systems that are based on machine learning. It has been established that machine learning algorithms outperform experts in detecting heart ailments from electrocardiograms (ECGs). Secure data exchange of medical records is presently a critical area in healthcare. In this project, we present a roadmap for a decentralized private health information ecosystem enabled by blockchain and machine learning to predict various health conditions based on the symptoms gathered from the ECG data of the patient.

No matter how advanced an algorithm is, limited and low-quality data is an obstacle to the success of any application but it can potentially be life-threatening in healthcare applications. Another important challenge is the accuracy of the algorithm. Also, along with the security of the data exchanged between two authorized entities, there is also a need to ensure the asynchronous transfer of data.

The goal of this research is to use machine learning techniques along with blockchain protocols to predict diseases using a pervasive cyber-physical device and ECG data. It is found that various diseases affect the ECG patterns in a unique way and exhibit discernable patterns for various ailments. We, therefore, use this property of ECG to detect hyperglycemia of patients in a non-invasive way. Subsequently, we convey the data extracted to the care provider in a secured channel via block chain.

Acknowledgments

The authors are deeply indebted to Professor Nima Karimian for his invaluable comments and assistance in the preparation of this study.

Connecting a World of Healthcare Data using Machine Learning and Blockchain

Deeksha Mahalingappa, Jason Gonsalves, Kavya Dayananda, Priyanka Kumar
Computer Engineering Department
San José State University (SJSU)
San José, CA, USA
Email: {deeksha.mahalingappa, jason.gonsalves, kavya.dayananda, priyanka.kumar}@sjsu.edu

***Abstract*—Personalised medicines and emergency treatment of ailments have been simplified due to automated systems that are based on machine learning. It has been established that machine learning algorithms outperform experts in detecting heart ailments from electrocardiograms (ECGs). Secure data exchange of medical records is presently a critical area in healthcare. In this project, we present a road-map for a decentralized private health information ecosystem enabled by blockchain and machine learning to predict various health conditions based on the symptoms gathered from the ECG data of the patient.**

No matter how advanced an algorithm is, limited and low-quality data is an obstacle to the success of any application but it can potentially be life-threatening in healthcare applications. Another important challenge is the accuracy of the algorithm. Also, along with the security of the data exchanged between two authorised entities, there is also a need to ensure the asynchronous transfer of data.

The goal of this research is to use machine learning techniques along with blockchain protocols to predict diseases like hyperglycemia using ECG data of the patients. It is found that various diseases affect the ECG patterns in a unique way and exhibit discernable patterns for various ailments. We, therefore, use this property of ECG to detect a repertoire of ailments using sensors in devices such as smart watches. Subsequently , we convey the data extracted to care provider in a secured channel via block chain. We used ECGs of 1,119 patients to experiment with different machine learning models and achieved an accuracy of 86% with the Random Forest algorithm and 75% with 10 layer DNN model.

Keywords- Electrocardiogram, hyperglycemia, machine learning, blockchaining

I. Introduction

With the invention of smart devices like smartphones, smart watches, of wearable devices like Apple watch, Fitbit and Microsoft band, it has become effortless to monitor individual health vitals. The data collected from such devices through their embedded sensors can be used to improve healthcare management.

ECG has been used to diagnose several cardiac diseases in medical field over the years. Thousands of people die because of cardiac diseases in the world every year and many of them are unaware of their health conditions. Researchers have realised that ECG, being a non-invasive method, can be used to diagnose not only cardiac diseases,but other diseases like diabetes, arrhythmia or hyperglycemia.

In this paper, we develop a machine learning model to detect hypoglycemic patients based on their ECG signals. We investigate novel fiducial methods for feature extraction and filtering and develop a model for accurate classification. Another import concern has been the security of sensitive health data. To ensure asynchronous transfer of data, we plan to develop a blockchain architecture using open chain.

The remainder of this report is divided into following sections. Second section discussed the literature review. Third Section is about project justification. Section four is about identifying baseline approaches. Fifth section discussed the dependencies and deliverables of the project while the next one elaborated on project architecture. Next section talked about evaluation methodology used in this project. Section eight described System Methodology while the next one put a light on experimental results. Finally, we concluded the paper in section tenth.

II. Literature Search

The electrocardiogram(ECG) is a non-invasive method of recording heart activity [1]. According to World Health Organization(WHO), the prominent reason for major deaths are cardiovascular diseases like cardiac arrhythmia, myocardial infarction etc.[2]. ECG diagnosis play a significant role in the diagnosis of cardiovascular diseases. Predicting heart diseases symptoms using prediction techniques would bring vast change in health life style of mankind [3].

Mostly, ECG recordings contain various types of noises. According to Berkaya et al. [4], ECG noises can be grouped into 7 categories: "Powerline interference, baseline wander, electrode contact noise, electrode motion artifacts, muscle contractions, electrosurgical noise and instrumentation noise". These noises can degrade the accuracy of a physician as well a machine learning model. Researchers have used several denoising techniques to reduce the noise levels in the preprocessing stage. Baloglu et al. [4] digitized signal at a sampling rate of 1000 Hz. to remove noise and baseline wander.

Various feature extraction techniques are used to generate significant feature sets from the ECG records. Baloglu et al. [4] used waveform transform technique to generate the feature sets in the preprocessing stage. After pre-processing, the features are fed to different classifiers. Baloglu et al. [4] used 10-layered Deep CNN based approach with 650 samples to classify 10 different myocardial infarction types with an accuracy of more than 99 percent.

Neural networks have been extensively used to identify various patterns in ECG recordings. These patterns may indicate an ailment or perturbation in a healthy lifestyle. Such pattern analysis is complicated due to noise in the measured samples. Based on the fundamentals of neural networks, it is found that a deep neural network rejects noise to extract underlying patterns in samples data. The deep neural network itself needs to be ensured of no overfitting and other modeling issues with sufficient training, test and validation data. We, therefore, review the following two efforts – an effort that directly utilizes the neural network to detect patterns; the other effort augments the data and performs classification on the augmented data.

Intuitively, modeling ECG based models per person over different conditions provides the most accurate results for the person under consideration. The effort presented by researchers in [5] extends this intuition to build a neural network to identify various ailments for a person. Various ECG samples are collected over time of a person and a deep learning model is built. The known patterns corresponding to ailments are used to train, test and validate the neural network model. Additional data coverage is achieved by augmenting the patient's data with the MIT/BIH database. The database is found in various cases of ECG variations that represent different heart conditions that are practically observable. They initially built a convolutional neural network (CNN) using the augmented data. Then, this model was optimized for embedded computing platforms for real-time feedback to the patient. Their work presented an accuracy of up to 99 percent and was also able to detect rare cases of ventricular ectopic beats and supraventricular ectopic beats. However, deriving and deploying such networks can be infeasible due to large numbers of people subscribing to such service. So, we consider another solution that derives neural networks using augmented data.

Data augmentation is necessary as it is difficult to extract several use cases per person on ailments and their effect on ECG. Furthermore, the data augmentation is performed such that it naturally follows the input data and extends the data for various cases. The researchers in [6] split their effort into two parts – the data augmentation model and a classifier. The data augmentation model is built using MIT-BIH data. The model to generate data is built as a generative adversarial network model. A GAN is a class of neural network that generates data with the same statistics as the data it is trained on. The data generated by GAN is then compared with the test data set to ensure correctness in data generation. The GAN model for work presented by [6] is built as a combination of convolutional layers with sigmoid function and an output softmax layer to generate source and label data. The classifier is built as a combination of residual and longest short-term memory (LSTM) networks. Initially, the GAN model is trained and tested to ensure the quality of data augmentation. The classifier is trained along with the trained GAN model that generates data for incomplete samples. The overall system accuracy was found to be around 99 percent over various cases. Most of the researchers above found overall accuracy to be 99 percent. One possible reason for this can be a very small sample size and overfitting of data.

Secure Transmission and collection of Data is more important for patients and medical organization [7]. Blockchain helps in secure transactions. Blockchain is an immutable record of transactions shared by various stakeholders by a distributed ledger. Consensus among all the members in the blockchain is achieved in a decentralized manner which makes it suitable for transfer of highly sensitive and classified data[8].

When data needs to be transferred from the testing center to the hospital, distributed ledger technology will be implemented to secure data using Blockchain with an open-source software platform such as Openchain, Ethereum etc.[9].Blockchain comprises of network layer,data layer and application layer[10].

The data layer contains the basic unit of the blockchain including the hash and hash pointer,Merkel tree,digital signature. The network layer encompasses the decentralized IP protocol-based network ,peer-to-peer (P2P) network, locking and unlocking scripts, and the mechanism used for the distributed block validity contract [11]. The network layer also allows the block to be updated and distributed.The application layer integrates blockchain to achieve consensus among the various distrusted nodes[10].

Author	Number of Samples	De-Noising Technique	Feature Extraction Technique	Machine Learning Model	Performance(in percentage)
Baloglu et al.	650	Baseline Wander	Wavelet Transform	10-layered CNN	99
Wang et al.	99244 single heartbeat segments 0	No augmentation	ACGAN	LSTM and residual network	88.3 (F1-score)
S. Kiranyaz et al.	360	PCA	Translation invariant dyadic wavelet transform	GAN	Adaptive 1-D CNN

TABLE I: Summary Table of Literature Review for Machine Learning.

Author	Type of Blockchain	Protocol	Application
Ayaskanta Mishra, Biswarup Chakraborty	Ethereum	MQTT	Smart Healthcare System using Internet of Things
Ferraro, Pietro , King, C	Open Chain	HTTP	Distributed Ledger Technology for Smart Cities, the Sharing Economy, and Social Compliance
Emre Yavuz et al.	Ethereum	HTTP	Towards secure e-voting using ethereum blockchain

TABLE II: Summary of Literature Review for Blockchain Technology.

A. State of the Art Summary

In the preprocessing stage, we will use several techniques, for example, filtering, digitization and normalization. The preprocessed data will be used to generate feature sets using P-QRS-T complex features technique. Dimensionality reduction techniques may be used if required. The features sets will be given to various models like Support Vector Machines (SVM), logistic regression, and Deep Neural Networks techniques.

Blockchain is governed by 3 principal technologies.

They are stated below:
1. Key cryptography
2. A distributed ledger
3. Network service protocol

1. Private key cryptography

Each member in the blockchain has the public key and a private key ,the combination of which is called digital signature. [10].

2. System of record

When the cryptographic keys are combined in a network ,it forms a system of digital interactions. For example, when a member Alice wishes to send data to Bob, she can do so by using her private key and attaching it to Bob's public key and this transaction can be broadcast to all the other nodes in the network[10].

3. Network Service Protocol

The blockchain protocol can be designed in accordance with certain rules and verification can be done separately for each blockchain.

III. Project Justification

An electrocardiogram (ECG) wave is a tool to record the electrical activity of the heart [1]. The ECG wave analysis is used to diagnose several cardiovascular diseases like cardiac arrhythmia, Myocardial infarction, hypertension, and other cardiac abnormalities[2]. The ECG signal for an individual is unique. One pattern which is abnormal to one person can be normal for the other. Hence, there is a need to record the ECG signal of a patient continuously to be able to analyze the pattern accurately. This is a very time-consuming task for the practitioners to do this task. High-risk patients need to visit the hospital several times to get their ECG recorded. Now-a-days, many wearable devices, for example, Fitbit, Apple watches, Nymi band are available in the market to record the ECG signal continuously. The purpose of this project is to continuously monitor the ECG data of the patients and accurately predict and diagnose using various machine learning techniques. In case of any abnormality, the data is securely transmitted to the hospitals using block chaining techniques. The data is then analysed thoroughly by the medical practitioners and the patient is reported immediately, if required.

One of the major challenges for healthcare providers is to share patient data with other organizations as patients move about for their treatment.Blockchain ensures that the records flow seamlessly between organizations.Security is mission critical in the healthcare sector as patient might not share all his true personal data due to the fear of records being accessed by third party intruders.With blockchain, every participant would have access to the same records. Managing databases are expensive and hence shared decentralised databases created by blockchain architecture help in reducing the overhead in managing databases thereby reducing cost.Blockchain helps in unlocking new value as its architecture facilitates accurate data sharing which enables healthcare companies to create new products and services that improve patient outcomes.

Researchers have realized that machine learning and Block chaining can play a very significant role in the healthcare industry. A lot of research has been done in the field to automate the medical field. There are two major concerns of the researchers, one is the secure transmission of data and the second is the performance of the machine learning algorithms. The goal of the project is to improve the performance of previous machine learning models and to transmit the data securely from the wearable devices to the hospital and vice versa. This can, particularly, be helpful to predict the disease at its earliest stage and can give doctors and patients more time to prepare themselves.

IV. Identify Baseline Approaches

With remote-information gathering, healthcare professionals can evaluate, diagnose and treat patients in remote locations using telecommunications technology. This project aimed at developing a small-scale electrocardiogram (ECG) monitoring device that will measure heart rates and waveforms and send the data to the nearest healthcare provider for further analysis if there is an anomaly detected. Historically, traditional ECG analysis was based on manual measurement using noticeable points P, Q, R, S, T. The goal of the project is to improve upon the analytical methods that are already present using deep learning algorithms. We plan to focus on increasing the degree of precision in classifying a disease and number of diseases that can be categorized using ECG signal.

The first step in the process is the preprocessing step . The preprocessing step can be further divided into filtering and feature extraction. The filtering step involves de-noising of the ECG signal. Noise is a major issue for signal processing in the biomedical field . Noise manifests itself in the ECG data due to several reasons like poor sensor quality, noise introduced due to data transmission, encoding or other channel noise. Apart from these, there are physiological noise that comes from muscle movements and external noise due to environment. It is imperative to remove noise as noisy data can dramatically impact the accuracy. A smoothing filter can be used to remove the noise in the ECG data The next step is the feature extraction process. The feature extracted from ECG signal can be fiducial or non fiducial in nature. In this project, we focused on the fiducial points P, Q, R, S and T. The ECG waveform is denoted by 5 major points, namely, P, Q, R, S and T. Figure 1[12] shows the ECG waveform points and the interval of major points. P wave is the arch before the QRS complex. Q wave is the first negative deflection of the QRS complex followed by R wave which is the positive deflection. S wave is the negative deflection after the R wave and T wave is the last arch after the QRS complex[13]. To perform a waveform comparison, each parameter is recorded from the actual ECG device/ ECG Test and a comparison of the results needs to be conducted.

After the feature extraction, we transmitted the pre-processed data to a pre-trained offline machine learning model securely using blockchain. In this project, we explored different machine learning techniques like One class SVM, Logistic Regression, Linear SVM, Random Forest and deep learning techniques like LSTM and DNN to train the ECG data.

For the block-chain implementation , Geth or Go-Ethereum which is a command line interface that is used to run full Ethereum node. Geth is implemented in Go and allows to mine blocks to generate Ether to deploy and interacts with smart contracts. Then ,a truffle project is spun up that laid out the structure of the project. Once the smart contract is written locally, Compiler compiled our Solidity code to byte code (the code that the Ethereum Virtual Machine (EVM) understands), After the contract is deployed, a transaction id is created . It contained the following data: Transaction Hash , Contract address of sender ,contract address of receiver and the amount of gas that was transferred. In our case, Ganache emulates the EVM.Testing can be done by creating a fake network using Ganache framework. In the deployment file (a file that Truffle gives when the project is created), the project can be pointed to either use Ganache or to use the main network. Then truffle migrate is used (which automatically runs truffle compile), to deploy the contracts with the data provided in the migration files.

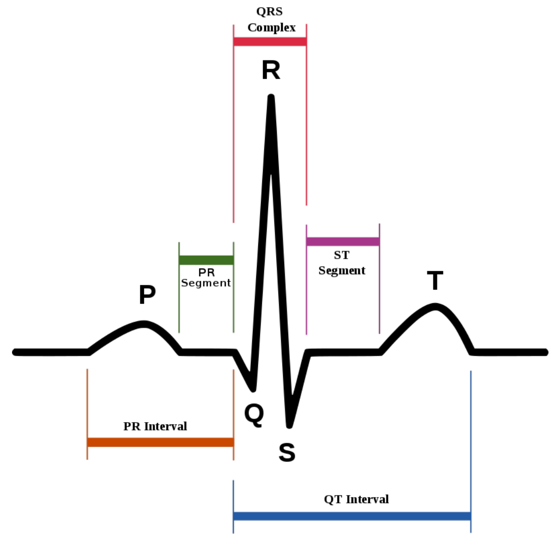

Fig. 1: Components of ECG Waveform

V. Dependencies and Deliverables

A. Dependencies

Our project is based on experimenting various machine learning and deep learning algorithms to detect abnormalities in the ECG signal of the patient received from the dataset and reports it back to the doctors. So, in order to correctly determine the abnormalities correctly ,there is a need to eliminate the noise in the ECG signal i.e, the quality of the ECG data is extremely important.This will greatly reduce the number of false positives and false negatives got after training the model.

Another important dependency in our project is the selection of the machine learning model to analyze the dataset. So, from the data ,it is essential to analyze the bias and variance in the data and decide what model should be used.

In the transaction between the patients and doctors a solidity contract is written to maintain the state for each of the patients and doctor state and that is chained in the block-chain to maintain the ledger.

B. Deliverables

This project has primarily 3 deliverables to be completed.

- The first deliverable is the creation of a decentralized healthcare information ecosystem application using Openchain which is an open source software to develop decentralized applications.

- The second deliverable is analysis of the dataset and creation of a machine learning model to predict the various health conditions from the ECG data of the patients.
- The third deliverable is to develop a blockchain network that ensures fast and secure transfer of patient data asynchronously with the doctor and maintain the ledger.

VI. Project Architecture

Hyperglycemia is a universal recognized disease that causes death around the world. Predicting the symptoms before it becomes fatal would benefit health and wealth of a patient.

Here we have designed a system architecture to predict whether the patient is suffering from hyperglycemia or not using ECG data to provide faster treatment for the disease as shown in Figure 1.

The ECG data will be tracked and collected from the activity tracker devices such as Fitbit, apple watch etc. The data collected is preprocessed in the device[14]. The preprocessing of the data involves Filtering, Feature Extraction and Segmentation. Then, the preprocessed data is obtained.

The preprocessed data which is in CSV format converted into JSON format before transmitting in a secure transmission protocol as HTTP protocol. The preprocessed data is peer-to-peer transmitted using distributed ledger technology securing data during transmission [15][16]. The data will be transmitted in a reliable protocol with a smaller packet ensuring every data is sent without drops in data packet. Here we are using Openchain distributed ledger technology which manages digital assets in a robust, secure and scalable way. Moreover, it is an open source which has many benefits.

The preprocessed data is transmitted to Machine Learning Laboratory. Here the data which collected from JSON format is converted into CSV format. Deep Learning (DL) as well as other classification models predict if the patient has hyperglycemia or not[17]. If the patient is predicted positive, a notification will be sent to hospital again using blockchain transmission with data conversion format where further precautions will be taken by the hospital authorities to protect patient's life as shown in Figure 2. If the prediction results return normal, the model will keep on monitoring in ML lab as shown in Figure 2.

VII. Evaluation Methodology

A. Machine Learning Metrics

System performance is assessed based on multiple metrics such as accuracy, recall, f1-score, and receiver operating curve (ROC). Any model's accuracy is calculated on the basis of accuracy and recall, which in turn is calculated on the basis of the true positive and false negative results. Precision tests how often a procedure correctly provides a positive outcome for people suffering from hyperglycemia. Recall tests the ability of a test to accurately produce a negative outcome for people without hyperglycemia.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{Numberofsamples} \tag{3}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

The area under the curve (AUC) of the receiver operating curve (ROC) is also used to compare the models. The ROC plots the performance of the model across different thresholds in terms of True Positive Rate (TPR) versus False positive rates (FPR). FPR is the percentage of healthy ECG which is falsely categorized as unhealthy whereas TPR is the percentage of unhealthy events which is correctly classified as unhealthy. Thus, the area of these curves provides a combined performance measurement of all these thresholds.

B. Blockchain Metrics

This section defines evaluation metrics for the blockchain network.

1) Read Latency: Read Latency = Time of receipt of response – submit time
Read latency is the time between submitting the request for reading and receiving the reply.

2) Read Throughput: Read Throughput = Total read operations / total time in seconds
Reading throughput is a measure of the number of read operations, expressed as reads per second (RPS), are performed in a defined time span. This metric may be insightful, but it is not the blockchain performance's primary indicator. In fact, systems adjacent to the blockchain will typically be deployed to facilitate significant reading and queries.

3) Transaction Latency: Transaction Latency = (Confirmation time @ network threshold) – submit time
Transaction Latency is a view of the amount of time taken to use the impact of a transaction. The measurement includes the time from the point that the result is generally available in the network. It includes the time of dissemination and any period of resolution due to the current consensus process.

4) Transaction Throughput: Transaction Throughput = Total transactions that are committed / total time in seconds @ number of committed nodes//

Transaction throughput is the rate at which the blockchain SUT conducts valid transactions over a period of time. Notice that this is not the price for a single node, but for the whole SUT, i.e. for all network nodes.

5) Considerations for Blockchain Performance Evaluation: This section addresses performance considerations for blockchain. Developers should carefully weigh these factors, document their decisions clearly, and accurately disclose their results in order to build a credible and reproducible model.

The test results should be reproducible independently. To achieve this objective, all parameters of the environment and test code should be recorded and made available, including any workload.

Here are some additional considerations when determining the setting in which quality evaluation can take place:

1) Protocol for Consensus. Description of consensus protocol used during the experiment, such as RAFT, Practical Byzantine Fault Tolerant (PBFT), etc.
2) Geographical node distribution. Describe the network's physical setup.
3) Hardware environment of all peers. Indicate the processor speed, number of cores, memory, and so on.
4) Network model. Any information about the complexity of the network employed between peers might reveal potential bottlenecks that could be highlighted.
5) Count of nodes involved in the transaction. Some blockchain platforms broadcast transactions while others restrict the transactions to a subset.
6) Software component dependencies. Does the system require additional components to work?
7) Test tools and framework. Describe techniques used to generate results in loading and capturing.
8) Type of data store used. The data store has an impact on performance, especially if it includes both reads and writes.
9) Workload. Code used to generate and validate the transaction

For a further discussion of workloads, see the article below.

Observation Points Measurements should be made, whenever possible, from the viewpoint of the test harness preferably situated outside the SUT. This will be closer to the end user's perspective. The measurement point should be clearly identified if this can not be done.

Transaction Characteristics

1) Complexity: How computational is the smart contract logic?
2) Data access patterns: Does the use of the production model the habits of reading and writing?
3) Dependencies: Do the transactions the production's transaction and data dependencies?
4) Size: How big are the transactions? This has implications for network-wide propagation times.
5) Workloads: A workload determines the task of the SUT. It is critical that the workload is representative of the actual use of production in any performance test. For example, if a workload simply creates new entries in the database, but the use of production primarily modifies existing entries, then the workload would tell us nothing about how to expect the process to function in production.
6) Faultloads:All blockchains should be built as core properties to provide ledger immutability, cryptographic integrity, and resistance to faults and attacks. By allocating additional resources, these features are allowed in the process.

6) Baseline approach for Blockchain: A central authority regulates the validity of different coupons that are produced is the baseline approach that we are considering . It means that this verification process must be done somewhere in the premises of either party or a trusted third party.

Based on these assumptions, it is highly necessary to have a decentralized, impartial and stable solution capable of validating and implementing transactions between two parties. Blockchain and smart contracts also tend to solve the problem as excellent solutions. LetPKX, SKX be the public and the secret key, respectively, in relation to a generic X object.The contract C(PA, PB) signed between platform PA and platform PB is formalized as follows: platform PA guarantees to platform PB that it will have access to resources available in platform PA when one of its applications AB comes with this coupon in the future.The contract C(PA, PB) signed between platform PA and platform PB is formalized as follows: platform PA guarantees to platform PB that it will have access to resources available in platform PA when one of its applications AB comes with this coupon in the future.The contract will be sent as a multi-signature transaction TC upon creation and deposited in the Blockchain, which will provide automatic confirmation of legitimacy.

$$T_C = [T_{ID}, D, CB, S, ts]\sigma \tag{5}$$

where TID is the transaction ID, D is the smart contract address and is empty to trigger the creation of smart contract, CB is the smart contract byte-code, S = H(PKX) is the sender address where H () is a generic hashing function,

$$\sigma = E(H(T_{ID}, D, CB, S, ts), SK_{P_A}, SK_{P_B}) \tag{6}$$

represents transaction signatures where E() is a generic digital signature algorithm and ts is a timeline introduced to resist replay attacks. However, a change of status will also be stored in the Blockchain when a contract is called. Then the issuer may view the history of the status. Since each modification is stored and checked, the validity of the transaction can be verified by checking how many times it has been used, what is the status and uses of the contract left, or whether it has expired.

VIII. System Design/Methodology

In this section, we provide an end-to-end overview of the system architecture. The entire detection process consists of periodically collecting the raw ECG signal, detecting the distinct features, identifying healthy and diseased ECG waveforms and securely transmitting the unhealthy ECG waveform to the appropriate care provider using block chain.

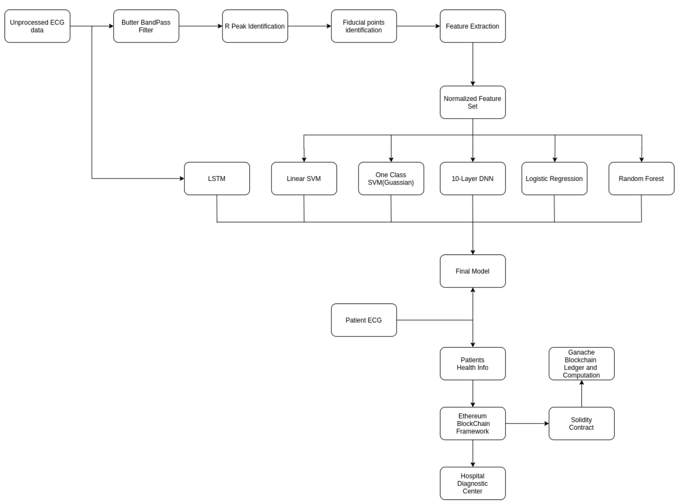

Fig. 2: System Architecture

The following are the detailed steps are as follows:

- Unprocessed ECG Data : We applied different preprocessing techniques on this data as discussed below.
- Filtering : This is a part of pre-processing step. The collected raw ECG signals are noisy due to various internal and external sources. These noise degrade the quality of ECG signal and affect the accuracy. Firstly, we removed first and last 1 second of ECG signal to remove noise from ECG signal. Therefore we use Butterworth Band pass filtering technique to filter out the noise using a frequency range of 1Hz to 40 Hz and order equals to 4.
- R- peak detection : This is the next part of filtering process. ECG feature extraction can be done by detecting fiducial points P,Q,R,S,T. The R-peak is necessary for segmenting heartbeat required to analyze the ECG signal. We use the wfdb library, Biosspy library and an adaptive filtering technique called Pan Tompkins to detect R-peaks.
- P,Q,S,T waves detection : The remaining waves named as P, Q, S, and T are then identified using Neuro Kit Library.
- Feature extraction : After filtering and fiducial points detection, we proceeded to feature extraction. There are different feature extraction techniques called fiducial and non fiducial techniques. Fiducial techniques are considered as finding P,Q,R,S,T points based on distances and amplitudes of the waves. Non fiducial techniques involves wavelet transformations like Discrete wavelet techniques. We used the fiducial technique and created 18 subset of features from every cardiac cycle using

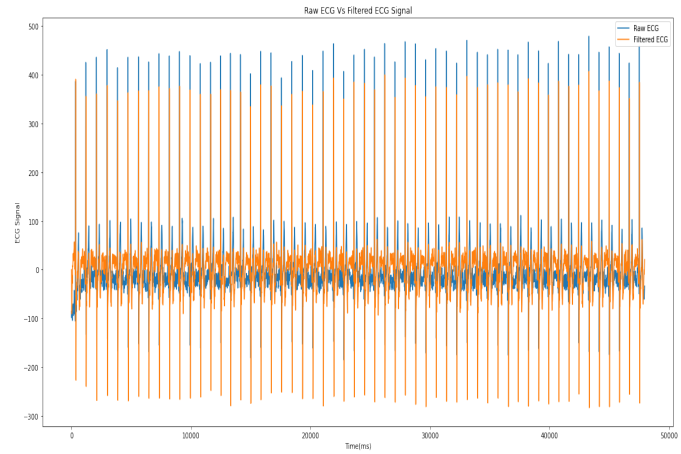

Fig. 3: Butterworth BandPass Filtering

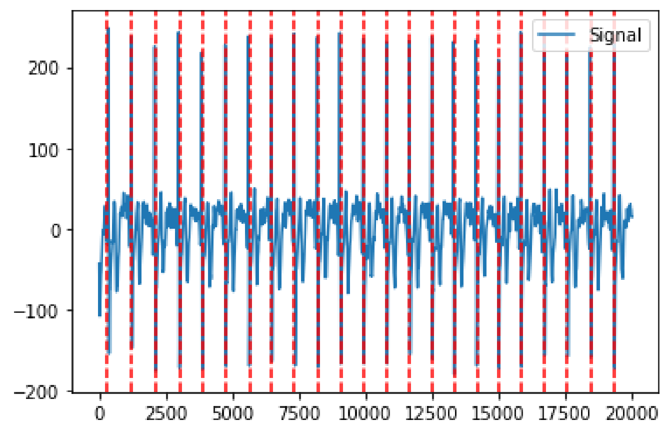

Fig. 4: R Peaks in ECG Signal

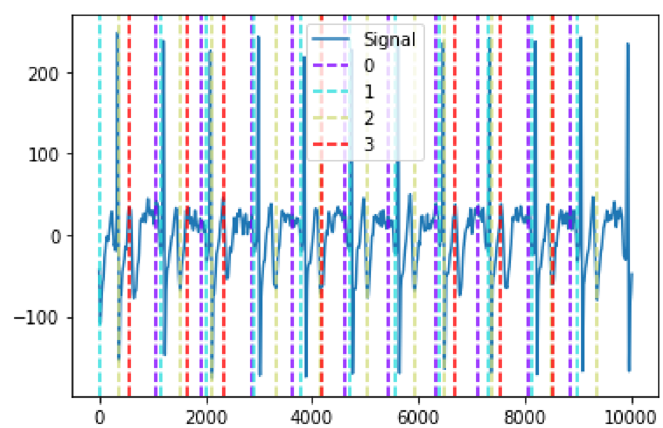

Fig. 5: Detected Peaks in ECG Signal

their respective euclidean distances and slopes.

- QT correction : Q wave is the first negative deflection of the QRS complex and T wave is the last arch after the QRS complex. The QT interval is the time is theme interval between these two waves. QT correction is required to reduce the interference caused by QT interval on heart rate.

- Normalization : We used standard scalar function(Scikit-learn library) to standardize using mean as well as scaled the data using standard deviation.

- Model Selection : After filtering, preprocessing and feature extraction from the ECG signal, we split data into 80-20 ratio and experimented with various machine learning as well neural network models. As part of pre-processing the feature set, we handled missing values, imbalanced data using SMOTE and dimensionality reduction using Principal Component Analysis. We experimented using LSTM, One Class SVM with different kernels, Linear SVM, logistic Regression, Random Forest and DNN. For hyper- tuning the parameters, we also used GridSearchCV(Scikit-learn library) and !0 fold cross validation.

We presented a predictive approach using the best of two models: Random Forest and DNN.

A. *Random Forest*

Random Forest classifier is an ensembled algorithm that uses a set of decision trees. Different decision trees are trained using randomly selected features set for training the model. The final decision of the predicted class is done by majority voting from all decision tree classifiers in case of classification problems. This models can efficiently handle large datasets like ours and is very accurate. To overcome the possibility of overfitting, we plotted a graph of F1 score with the maximum depth of trees for training and testing data which is shown in Figure 6.

B. *DNN Model*

As an alternative to Random Forest model, we approach this problem by hand selecting the features and training a customised DNN model. We evaluate manual feature selection and its efficacy as compared to automatic feature extraction using LSTM. We provide the peaks, the slopes with respect to each other and their distribution across the entire sample set. We then feed this data to train a fully connected neural network and softmax classifier at the output. The ReLU activation was used in fully connected network since it was the only activation layer that provided converged results for the given data.Also, we found empirically that Xavier initialisation for initial random weights converges faster than Gaussian initialisation.The model was trained and tested with non-overlapping data set to test for the efficacy of the model fit. We derive test and validation results from these models and pick the model with lowest fit error for online deployment.

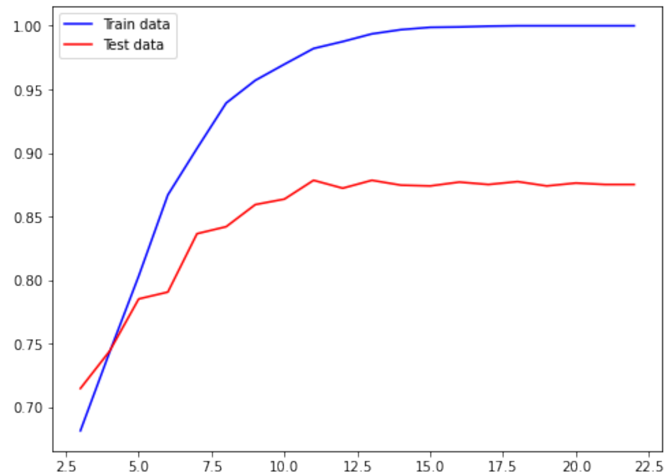

Fig. 6: Random Forest. F1 Score Vs Max Depth

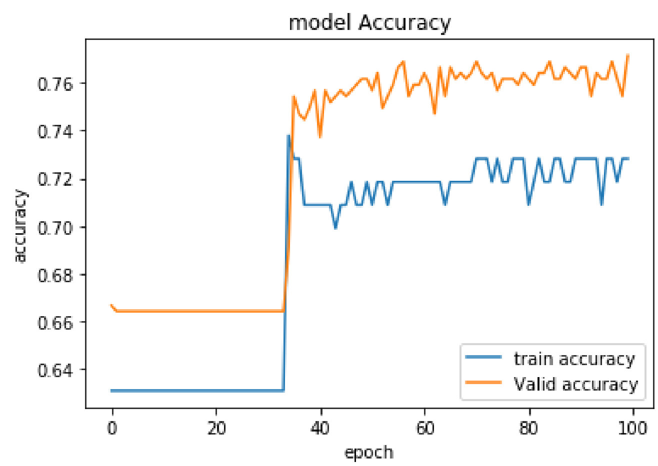

Fig. 7: DNN. Accuracy Vs Epoch

Deploying the machine learning model using Flask Our purpose of deploying the machine learning model using Flask is to make the model available to the stake holders of the blockchain network.

Project Structure As illustrated in the figure below, the deployment has four parts:
1. model.py - Contains the code to predict whether the patient is hypoglycemic or not.
2. app.py - Contains the REST API written in flask that computes the predicted value based on the model.
3. request.py - Contains requests to call APIs defined in app.py and displays the returned value.
4. HTML/CSS - Renders the predicted value once the model is selected,

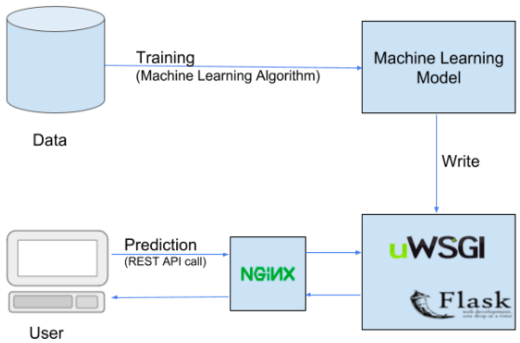

Fig. 8: Machine learning model Deployment

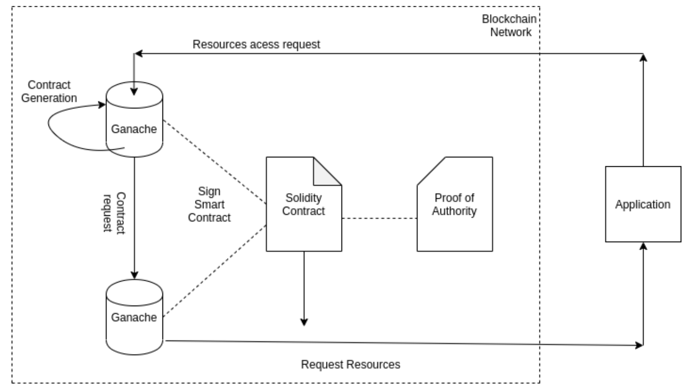

Fig. 10: Blockchain Interaction

We basically identify patients with an abnormal glucose level using the machine learning model and transmit the patient data via blockchain to the doctor . Each patient registered id is validated before entering into the blockchain network and he can view the results of the other patients but the identities of the other patients is anonymous. If a person with an id not belonging to this blockchain network attempts to enter, he will be denied entry access.

Once the registered user logins to view his results, a transaction id encrypted by hashing will be generated and a transaction will be recorded on the distributed ledger shared by all the stakeholders of the blockchain network. The transaction ledger is maintained by Ganache as shown Figure 7.

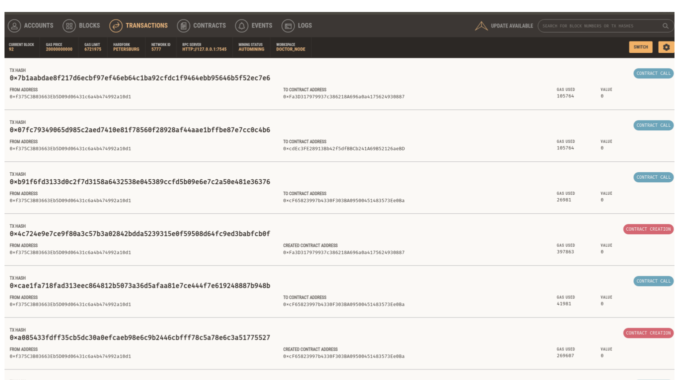

Fig. 9: Blockchain Transaction Ledger

- Ganache is the network protocol which runs on port 5777 that acts as a network identifier.
- Ganache provides dummy private keys for the users registered in the blockchain network. Whenever a transaction takes place, the ledger will contain the record of the private key of the sender and the receiver transaction information.
- Communication between Ganache and the Ethereum application is done through Solidity contract which signs the transaction.
- There are two types of modifier called access and view modifiers. Gas is charged for access modifiers and not for view modifiers.

C. Data Visualization using Dataset

D3.js

D3.js is a JavaScript library for the data-based processing of documents. D3 lets the data come alive using HTML, SVG and CSS. D3's focus on web standards allows you the full potential of modern browsers without binding yourself to a proprietary platform, merging strong visualization elements with a data-driven approach to manipulating DOM.

D3 lets you bind arbitrary data to a Database Object Model (DOM), then apply data-driven transformations to the content. You may for example use D3 to create an HTML table from a number list. Or, to construct an interactive SVG bar chart with smooth transitions and interaction using the same data.

Bar Chart

Bar graphs, also called column maps, use vertical or horizontal bars to visually reflect data along both an x-axis and a y-axis. Every bar is a one-value. The user will compare the different bars, or values, at a glance when the bars are stacked next to each other. In this bar chart we have plotted the Users on the x-axis and the corresponding glucose level on the Y-axis.

In the below graph we have plotted a bar graph displaying the glucose level of patients vs user ids.

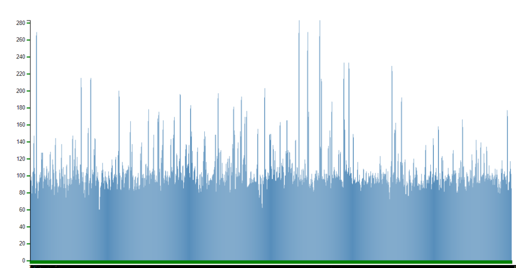

Fig. 11: Bar graph of Dataset

Line Chart A line graph is generally used as a set of data points connected by straight line segments on two axes to display change over time.

In the graph shown below, we have plotted a line chart displaying the glucose level of patients vs user ids.

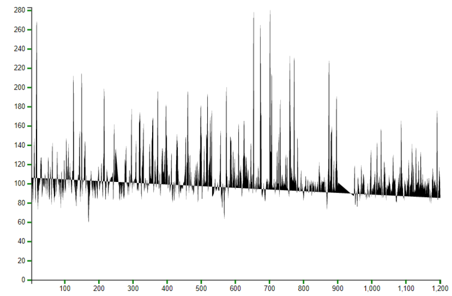

Fig. 12: Line chart of Dataset

Highcharts

Highcharts is a modern multi-platform charting library, based on SVG. This makes it possible for online and mobile initiatives to have interactive charts included.

In the Figure 13 we have plotted a column chart displaying the ECG level of patients vs Time.

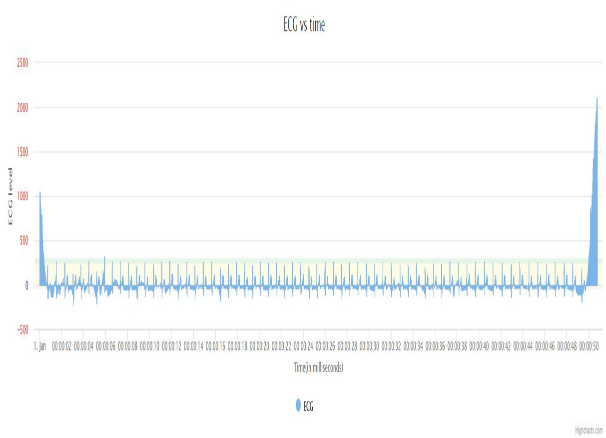

Fig. 13: Highcharts of Live data

IX. Experimental Result

Several machine learning models, like One Class SVM, Logistic Regression, Linear SVM, Random Forest and deep learning models, like Deep neural network and LSTM have been experimented. The performance metrics used to evaluate the performance of models are: True positive rate and False positive rate. The table shown below summarizes the experimental results.

X. Conclusion

ECG-based hyperglycemia detection, being a non-invasive approach provides easy accessible diagnosis for the patients who require continuous monitoring. Since various machine learning approaches were experimented along with hyper parameter tuning to obtain the best accuracy. Our proposed model using Random Forests achieved an accuracy of 86%.

Model Name	**Accuracy**
10-Layer DNN	76%
Logistic Regression(C = 3)	71%
One Class SVM Gaussian(nu = 0.081, gamma=0.00000001)	60%
Random Forest(number of estimators = 150, max depth = 10)	86%
SVM Linear(C = 3)	71%

TABLE III: Models Performance.

In future, the work can be improved by exploring other available libraries to analyse and filter the ECG signal. Also, we can predict glucose concentration of a patient using regression models instead of classification ones.

References

[1] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of biomedical informatics*, vol. 59, pp. 49–75, 2016.

[2] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ecg analysis," *Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[3] G. Shanmugasundaram, V. M. Selvam, R. Saravanan, and S. Balaji, "An investigation of heart disease prediction techniques," in *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*. IEEE, 2018, pp. 1–6.

[4] U. B. Baloglu, M. Talo, O. Yildirim, R. San Tan, and U. R. Acharya, "Classification of myocardial infarction with multi-lead ecg signals and deep cnn," *Pattern Recognition Letters*, vol. 122, pp. 23–30, 2019.

[5] P. Wang, B. Hou, S. Shao, and R. Yan, "Ecg arrhythmias detection using auxiliary classifier generative adversarial network and residual network," *IEEE Access*, vol. 7, pp. 100 910–100 922, 2019.

[6] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[7] M. Elhoseny, G. Ramírez-González, O. M. Abu-Elnasr, S. A. Shawkat, N. Arunkumar, and A. Farouk, "Secure medical data transmission model for iot-based healthcare systems," *IEEE Access*, vol. 6, pp. 20 596–20 608, 2018.

[8] W. Liu, S. Zhu, T. Mundie, and U. Krieger, "Advanced block-chain architecture for e-health systems," in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2017, pp. 1–6.

[9] E. Yavuz, A. K. Koc, U. C. Çabuk, and G. Dalkılıç, "Towards secure e-voting using ethereum blockchain," in *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*. IEEE, 2018, pp. 1–7.

[10] Y. Lu, "The blockchain: State-of-the-art and research challenges," *Journal of Industrial Information Integration*, 2019.

[11] C. Gao, Y. Ji, J. Wang, and X. Sai, "Application of blockchain technology in peer-to-peer transaction of photovoltaic power generation," in *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. IEEE, 2018, pp. 2289–2293.

[12] Wikipedia, "Ecg overview–Wikipedia, the free encyclopedia," 2008. [Online]. Available: \url{https://en.wikipedia.org/wiki/Electrocardiography#/media/File:SinusRhythmLabels.svg}

[13] A. Mishra and B. Chakraborty, "Ad8232 based smart healthcare system using internet of things (iot)."

[14] L. Zheng, W. Hu, and Y. Min, "Raw wind data preprocessing: a data-mining approach," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 11–19, 2014.

[15] P. Ferraro, C. King, and R. Shorten, "Distributed ledger technology for smart cities, the sharing economy, and social compliance," *IEEE Access*, vol. 6, pp. 62 728–62 746, 2018.

[16] R. Roverso, S. El-Ansary, and S. Haridi, "Peer2view: A peer-to-peer http-live streaming platform," in *2012 IEEE 12th International Conference on Peer-to-Peer Computing (P2P)*. IEEE, 2012, pp. 65–66.

[17] X. Li, H. Wang, H. He, J. Du, J. Chen, and J. Wu, “Intelligent diagnosis with chinese electronic medical records based on convolutional neural networks,” *BMC bioinformatics*, vol. 20, no. 1, p. 62, 2019.