The role of blockchain and IoT in recruiting participants for digital clinical trials

Fabio Angeletti
DIAG
University of Rome “La Sapienza”
Email: angeletti@dis.uniroma1.it

Ioannis Chatzigiannakis
DIAG
University of Rome “La Sapienza”
Email: ichatz@dis.uniroma1.it

Andrea Vitaletti
DIAG
University of Rome “La Sapienza”
Email: vitaletti@dis.uniroma1.it

***Abstract*—Our personal data is now more valuable than ever. The uncontrolled growth of internet-centered services has led us to accept many compromises about how we share it. In the era of Internet of Things, personal data is collected continuously. Now, more than ever, we are in need of privacy-preserving applications where users always retain control of their personal data. In this paper, we present a secure way to control the flow of personal data in the specific case of the recruitment of participants for clinical trials. We take special care to protect the interests of both parties: the individual can keep its data private until an agreement is reached, and the Clinical Research Institute can be assured that it is acquiring useful and authentic data. We provide a proof-of-concept implementation and study its performance based on a real-world evaluation.**

***Index Terms*—Blockchain; IoT; Clinical Trial; mHealth; Privacy.**

I. Introduction

The last decade we witnessed a tremendous progress towards the interconnection of the digital and physical domains, giving rise to the “Internet of Things”. A particular domain where the coexistence and cooperation of embedded systems with our social life is unveiling a brand new era of exciting possibilities is that of digital health. With the ever-increasing amount of data that is inherent in an IoT world, drug developers can potentially access a wealth of real-world, participant-generated data that is enabling better insights and streamlined clinical trial processes.

Developing drugs is a challenging process. Only around one in 10 drugs in Phase 1 actually make it through to the market [1]. This low rate to enter the market is one factor contributing to the high costs of drug development. A recent study indicates that developing a drug from bench to market costs an estimated $2.6 billion [2]. A large portion of those costs is related (a) to the stage of recruiting an adequate number of patients and (b) retaining the patients throughout the trials. According to [3] there are currently more than $244,000$ studies registered in the world and more than $42,000$ are recruiting. About 300 clinical trials in 2015 were reported to involve a wearable technology [7].

Some of these studies require thousands of participants, each of whom must meet precise criteria to join. So its not surprising that 80% of these important studies are delayed due to recruitment problems, according to the Center for Information and Study on Clinical Research Participation (CISCRP) [4]. Long recruitment phases prolong the execution of trials thus taking longer for innovative new medicines to be studied and approved, leaving patients to wait years for new treatment options.

The need to access an appropriate pool of patients in order to execute clinical trials is well known to the broader public. It is observed on multiple occasions: a growing pool of people willing to participate. In a 2015 study of CISCRP [4] 81% of respondents agree that clinical research studies are “very important” to the discovery and development of new medicines and 80% of them would be willing to participate in a research study.

According to a 2012 on-line survey [5], 85% of the respondents perceive privacy concerns as a barrier to sharing health information. If data were irreversibly anonymized, 71% of respondents were willing to share data with researchers and about half of respondents were either concerned or very concerned about the re-identification of their anonymized health and medical information. During a clinical trial recruiting phase, when the benefits of a possible future enrollment have not been fully clarified, patients expect information about their medical condition will be kept confidential.

Given the qualification criteria imposed by the researchers, about 5% of patients eventually constitute the group from which participation in trials is selected. It is, therefore, imperative to introduce new methods for facilitating recruitment that respects the privacy and confidentiality of the patients in order to maximize the participation of people - particularly in rare diseases where the communities of patients are small.

During the execution of the trials, the collection of high-quality data is absolutely vital. For this reason, trial centers require regular tests and observations to be conducted at their premises in order guarantee the accuracy of data collection. As reported in [3], 70% of potential participants live more than two hours away from the nearest study center. It is, therefore, common for patients to travel to those centers for regular tests and observations, sometimes several times each week for the duration of the trial. Such complexities sometimes overcome the perceived benefits of participating in a trial, inevitably increasing the attrition rate of patients. Clearly, redoing patient recruitment further delays the execution of the trials.

Understanding the above issues and addressing them adequately is critical in developing successful digital health solutions.

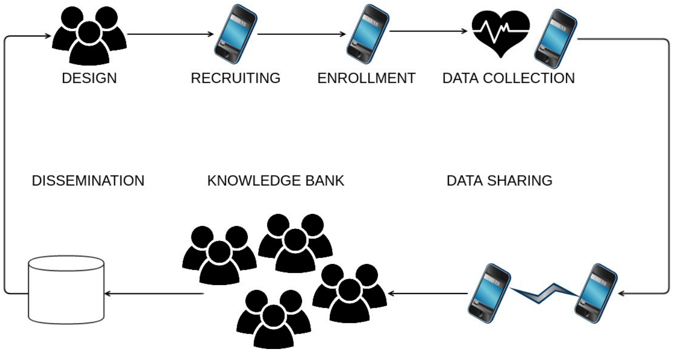

Fig. 1. A model of conducting digital clinical trials

In this work, we assume a generic model for the digitization of clinical trials as the one depicted in Fig. 1. This model is based on the ability of modern technologies to communicate over the Internet in order to reach nearly an unlimited number of potential participants and leverages the IoT to collect relevant data for a specific study employing suitable devices deployed outside the "study centers", possibly at participants' homes. In this model, a digital clinical trial is initially designed, possibly taking into account insights from previous trials. In the recruiting phase, patients matching the needs for the designed trial are recruited. Recruited patient are required to enroll upon accepting the informed consent. Enrolled patients remotely collect useful data and share those data to build a statistical significant knowledge bank. The outcomes of the trial are disseminated and can contribute to the design of a new trial.

Our work focuses on the recruiting phase. We rely on IoT technologies in order to provide a solution that suitably addresses the needs of both the clinical research institutes and the participants to clinical trials. The proposed approach moves beyond existing solutions that just connect patients to trials that they might be interested. It enables the ability to bring specific patient profiles into the recruitment process allowing clinical research institutes to digitally prescreen patients for the inclusion-exclusion criteria on a trial and learning early on if a patient is a candidate, or perhaps in a particular population seeing whether there are even enough patients that meet the criteria.

Facilitating and improving the effectiveness and efficiency of the recruitment phase by exploiting modern IoT health devices in order to get an unprecedented amount of data is indeed a tremendous opportunity. Still a major problem arises: who are the owners of such data? We agree with the principles discussed in [8]: the ownership of data should be of the participants. As observed in [9]: "data sharing could put clinical trial participants at increased risk of invasions of privacy or breaches of confidentiality. As a result, participants could suffer social or economic harms". The proposed solutions use suitably selected cryptographic tools and the blockchain in combination with IoT technologies to provide a holistic environment that respects the privacy of personal data and guarantees its confidentiality.

II. Previous work

In [10] a comparison of mhealth and traditional clinical research in different study stages is presented. Limiting our attention to the recruiting phase, the main pros of mhealth are the ability to immediately recruit nearly unlimited participants with no geographic limitations and the cost effectiveness of the whole procedure while cons include the quality of data used to recruit participants, both self-reported and device generated.

The need for security and privacy in mhealth solutions has been investigated in a number of papers, such as [11], [12] just to mention the most recent ones. The idea that the ownership of data should be retained by the participants is a key element as also reported in [8]. However, the exploitation of the blockchain to address some of those issues is relatively novel. The usage of blockchains and smart contracts within the Internet of Things sector is examined in [13] while, similarly to our approach, in [14] the authors propose a purpose-centric access model leveraging the blockchain to ensure patient own, control and share their healthcare data. They also pointed out the secure multi-party computing is a promising solution to enable untrusted third-party to conduct computation over patient data without violating privacy.

III. Use-case

We here present a generic use case scenario that helps us capture the key aspects of the recruitment phase from a digital health perspective. We encode the needs of each participating party and identify a set of minimum requirements.

We assume a *clinical research institute* that starts the recruitment phase by carefully specifying the desired patient profiles and the digital screening process. The screening relies on the evaluation of specific biometric attributes (e.g., body composition, heart operation, daily activity, etc.) collected by the patient using wearable technologies over a given period of time (e.g., blood pressure for past week, etc.). The clinical research institute has conducted certain accuracy evaluation tests over various off-the-shelf wearable devices and smart devices and has compiled a list of "trusted" devices that it considers accurate enough so that data collected from these devices can be used through the digital screening phase. Based on an adequate set of genuine historic values collected from one or more of these "trusted" devices, the clinical research institute analyses the data (e.g., using a combination of logistic regression models, principal component analysis, etc.) and decides if the patient will be included or excluded. We translate these steps into the following requirements:

R1 The quality of data provided by participants is guaranteed (only consider data collected from approved devices);

R2 The inclusion/exclusion of a candidate is based on the execution of a well-defined *recruiting test* automatically executed over a provided data set of historic values;

R3 The candidate must provide proof that the historic data set is real and collected over the stated period of time.

An individual that wishes to be considered for a specific clinical trial expects that the privacy of her personal data will be respected and the confidentiality of the private data will be guaranteed. If during the digital screening phase the individual is excluded, then the digital health system should guarantee that no personal data have been retained by the clinical research institute. The system is secure enough to ensure that only devices installed by the individual can participate and that no fake data can be injected into the system.

R4 Data are collected by certified and trusted devices;

R5 Fake data cannot be introduced into the private space;

R6 Candidates' privacy is preserved during the recruiting phase. Recruiting test is privacy preserving, namely, it does not disclose users' data to the institute. Data never leave the private space unless the candidate voluntarily enrolls to the trial because he/she is eligible according to the outcome of the recruiting test (see R2);

R7 Periodic certificates are provided to prove the authenticity, integrity, and conformance of collected data (see R3).

Once in the enrollment phase, the data are eventually delivered to the clinical research institute, the participant has to trust that institute for the subsequent management of his/her data. Huge and important markets, such as one of the digital media, have fundamentally failed to design data protection mechanisms capable of avoiding the duplication of data or their illegitimate sharing to the wider audience. The only possibility is to trust the receiver of data to behave correctly. While this assumption is critical in the context of digital media markets in which the receivers are potential all the Internet users, this becomes more realistic when the intended receiver is a clinical research institute with a well-known reputation.

IV. Reference Architecture

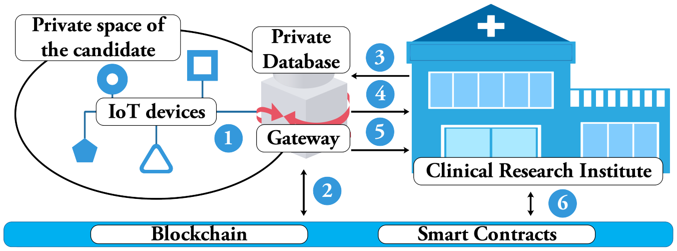

Fig. 2. A reference architecture for digital healthcare applications for recruiting in clinical trials

We now introduce a reference architecture (see fig. 2) for conducting recruitment in digital clinical trials designed to fully support the requirements outlined in Section III. We consider individual users that are using a collection of (a) wearable devices attached on them and (b) smart devices positioned within their home environment. Within the home environment there is also a so-called *gateway* that periodically communicates with all available devices to retrieve collected data (step 1, Fig. 2) and to store them on a local storage space.

Currently, all wearable devices (and smart devices) are equipped with an identifier that encodes the manufacturer, model number and production series. The proposed approach requires that they are also equipped with a temper proof memory space where the manufacturer places a private key. The corresponding public key is stored on the blockchain to make it publicly accessible by the *gateway*s and the research institutes (step 2, Fig. 2). This allows a research institute to identify the devices available to the individual and to decide if the data collected are within the accuracy requirements (R1). Furthermore, whenever the wearable devices communicate the collected data to the *gateway*, all packets are digitally signed and therefore the *gateway* can check their integrity and authenticity (R4), and discard those that are arriving from untrusted devices (R5). Periodically, the *gateway* applies a hash function over the collected (trusted) data that is stored on the blockchain (step 2, Fig. 2, R7). The hashing allow the *Clinical Research Institute* to check that the dataset provided are indeed the ones collected over the period of time claimed (R3). The recruiting test analyses a historic dataset provided by the individual users and decides if the candidate should be included or excluded. The analysis is encoded into code that generates only constant-size output (i.e., O(1)) and thus cannot include any personal data that can violate the confidentiality of the candidates. The *Clinical Research Institute* sends the code to all interested parties (step 3, Fig. 2). The *gateway*s locally execute it on the locally stored historic data (R2). The output is transmitted to the *Clinical Research Institute* without revealing any of the private data (step 4, Fig. 2, R6). If the participant is eventually enrolled for the actual trial, she provides her actual historic data (step 5, Fig. 2) and the *Clinical Research Institute* can finally check the data integrity employing the hash function published in the chain (step 6, Fig. 2, R3).

As already observed, in the subsequent enrollment phase, participants have to trust the data management of the research institute. In view of this, instead of the blockchain, a more traditional trusted and certified database could be used to store all relevant information. However, we decided to employ the blockchain in our solution, because its P2P (and trustless) nature is in line with the principles of participation, personal data ownership and accountability that inspired our design.

V. Proof of Concept

We implemented a Proof-of-Concept (PoC) of the reference architecture using RiotOS [15] and ethereum [16]. We evaluate the performance using the real-world experimentation facility IoT-Lab [17]. *M3 Open Node*s provided by IoTLab are used as IoT devices for the generation of biometric data. A Raspberry Pi3 (CortexA53 1.2GHz, 1Gb DDR2) is used as a *gateway* to interact with the IoT devices (residing at the IoTLab) and an ethereum instance of the testing blockchain. A standard PC (Intel i7-6500U 2.50GHz, 8Gb DDR3) is used for the *Clinical Research Institute* to communicates with the *gateway* and with the testing blockchain.

A. Signing in the private space

Requirements R1, R4 and R5 are addressed by signing the generated data using the uEcc library available on RiotOS and in particular the NISTP-256 elliptic curve [1]. Fig. 3 shows the time necessary to sign 64000 bytes of data on the resource constrained *M3 Open Node*s dividing them in chunks of different size (1000×64 bytes, 100×640 bytes, 10×6400 bytes, 1×64000 bytes).

The evaluation indicates that the signature function is by far the most time consuming, while the hashing is always relatively fast. The total time necessary to sign 1000 chunks of 64 bytes were more than 200sec, namely more than 400 times bigger than the time necessary to sign the chunk of 64 Kbytes (about half a second). Furthermore, when we consider chunks of 64 bytes, the signature occupies about one third of the payload, while in the 64000 bytes case it is less than 0.1%. This result suggests that some form of aggregation is always necessary to implement a practical solution.

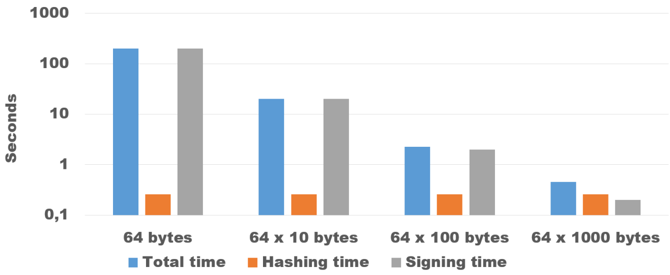

Fig. 3. Time necessary to sign chunks of data of different sizes for an overall of 64000 bytes. Please note the logarithmic scale.

B. The recruiting test

The recruiting test (R2) is used by the *Clinical Research Institute* to discriminate whether a candidate is suitable to enroll for a given clinical trial. At a preliminary investigation with physicians, in many practical cases, the recruiting test is relatively simple and mostly based on threshold functions. However, in our experiments, we performed some tests to evaluate the ability of our PoC to allow more complex and general recruiting tests based on machine learning algorithms. We consider two main scenarios:

S1 The data in the private space, possibly pre-processed to extract relevant features, are used to learn some model. The parameters of such model are given as inputs to the recruiting test.

S2 A given pre-computed model is embedded by the *Clinical Research Institute* into the recruiting test, and some features of the data in the private space are given in input to such model in order to classify the candidate.

In S1, both features extraction and machine learning are performed by the *gateway*, while in S2, the model for the machine learning phase is computed by the *Clinical Research Institute* and provided to the *gateway* to execute the classification based on the features extracted.

We used four real-world datasets provided by the UCI Machine Learning Repository [18]. They vary both in the number of samples and in the number of features, accordingly to table I. Two of them target health issues as cancer and heart diseases (*Arcene* and *Heart Diseases - Cleveland*). *EEG Eye State* correlates the EEG with open or closed eyes while *Gisette* has been selected to stress the performance analysis.

TABLE I
DATASETS USED TO EVALUTE PERFORMANCES

#	Dataset Name	Number of Samples	Number of Features
1	Arcene	100	10000
2	EEG Eye State	13444	14
3	Heart Disease	270	13
4	Gisette	6000	5000

Machine Learning. The main goal of this experiment, is to evaluate the ability of our reference architecture, to support the computation of complex recruiting functions on realistic data-set. We use 6 well-known machine-learning algorithms available on the *Python* package *scikit-learn* [19] and their execution times are shown in table II.

The fine tuning of such algorithms, in order to improve the effectiveness of the results will be the objective of a future work. However, our preliminary investigation shows that the accuracy can change significantly for different algorithms; as an example the accuracy of KNN on [20] is already up to 95%, while k-means on the same dataset is less than 55%. Most remarkably for our purpose, the accuracy is the same both in the PC and in the *gateway*.

TABLE II
EXECUTION TIMES OF COMMON MACHINE LEARNING ALGORITHMS. EACH ROW CORRESPONDS TO THE EQUIVALENT DATASET IN TABLE I

	#	PC	Gateway
Support Vector Machines	1	0.147322s	0.695988s
	2	45.561s	565.288s
	3	0.005111s	0.061623s
	4	250.205s	1180.824s
Logistic Regression	1	0.133114s	1.639809s
	2	0.177369s	2.124157s
	3	0.001498s	0.020984s
	4	1.707945s	22.940s
k Nearest Neighbors	1	0.010120s	0.086751s
	2	0.009555s	0.125092s
	3	0.000355s	0.002438s
	4	1.315992s	18.851s
Gaussian Mixture Models	1	0.245765s	2.103226s
	2	0.667314s	7.934804s
	3	0.011209s	0.087351s
	4	30.658s	Memory Error
k-Means	1	0.116964s	0.859489s
	2	0.029887s	0.293893s
	3	0.003146s	0.035679s
	4	8.616173s	Memory Error
PCA	1	0.463008s	3.250940s
	2	0.055881s	0.662027s
	3	0.000539s	0.003660s
	4	40.488s	Memory Error

[1] Notice that this curve is also used in the blockchain

In all cases the *gateway* is an order of magnitude slower than the PC but it can execute the algorithms in reasonable time. The algorithms that are more demanding in terms of memory requirements can easily exhaust the memory generating significant delays (e.g., due to swaps) and in some cases create *Memory Errors* on the resource constrained *gateway*.

Features Extraction. Raw data generated by the IoT devices in the private space are elaborated in order to extract relevant features provided as input to the machine learning algorithms. In this experiment we use ECG data analyzed following the methodology of [21]. We first apply a Butterworth Band Pass Filter of the 5th order. As a second step we calculate the Fast Fourier Transform (FFT) and the associated Power Spectral Density (PSD). Finally, we evaluate the mean, the standard deviation, the variance and the maximum peak of the signal as statistics.

TABLE III
EXECUTION TIMES TO ELABORATE ECG

	PC	**Gateway**
Filtering	0.114648s	1.483026s
FFT and PSD	0.621372s	4.410873s
Statistics	0.016048s	0.115040s
Total	0.752068s	6.008939s

In Table III we reported the execution times required to analyze one hour of real ECG data sampled at 360Hz [22]. Again, our findings indicate that the *gateway* is roughly an order of magnitude slower than the *PC*, but it can extract important and complex features in absolutely reasonable times.

C. Interacting with the blockchain

Data acquired in the private space are periodically hashed and stored in the blockchain (R7). This allows the *Clinical Research Institute* to validate the authenticity, integrity and conformance of collected data (R3). The blockchain is also used to store the public key generated by the devices and published by the *gateway* on the blockchain to control which devices are acurate enough for generating health data (see Section V-A). Providing guarantees on the performance of validating transactions in the blockchain is not possible by definition, but usually is in the order of minutes. On the contrary, the time necessary to read the blockchain is negligible.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a digital health application that enables recruitment of clinical trials to use IoT data. Blockchain technologies are used to guarantee that personal data are not shared publicly until an individual is enrolled. We demonstrate that our solution can be implemented in real-world IoT devices and achieves adequate performance over real-world datasets. We believe that by suitably combining blockchain and IoT technologies we can provide significant results over the digital health domain. As future work we wish to extend our work by using IoT over the other steps of clinical trials.

ACKNOWLEDGMENTS

This work has been partially supported by the research project Designing Human-Agent Collectives for Sustainable Future Societies (C26A15TXCF) of Sapienza University of Rome.

REFERENCES

[1] M. Hay, D. W. Thomas, J. L. Craighead, E. Celia, and J. Rosenthal, "Clinical development success rates for investigational drugs," *Nature Biotechnology*, vol. 32, no. 1, pp. 40–51, 2014.

[2] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of r&d costs," *Journal of Health Economics*, vol. 47, pp. 20–33, 2016.

[3] "Trends, Charts, and Maps at ClinicalTrials.gov," https://clinicaltrials.gov/ct2/resources/trends.

[4] "The Center for Information and Study on Clinical Research Participation (CISCRP) ," https://www.ciscrp.org.

[5] K. T. Pickard and M. Swan, "Big desire to share big health data: A shift in consumer attitudes toward personal health information," in *2014 AAAI Spring Symposium Series*. AAAI, 2014.

[6] A. Smith, "U.S. smartphone use in 2015," http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/, pewResearchCenter, 1 April 2015.

[7] A. Edney and C. Chen, "Big pharma hands out fitbits to collect better personal data," http://www.bloomberg.com/news/articles/2015-09-14/big-pharma-hands-out-fitbits-to-collect-better-personal-data, bloomberg, 14 September 2015.

[8] S. F. Terry and P. F. Terry, "Power to the people: Participant ownership of clinical trial data," *Science Translational Medicine*, vol. 3, no. 69, pp. 69cm3–69cm3, 2011. [Online]. Available: http://stm.sciencemag.org/content/3/69/69cm3

[9] C. on Strategies for Responsible Sharing of Clinical Trial Data; Board on Health Sciences Policy; Institute of Medicine, *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, National Academies Press (US), Washington (DC), 4 2015, guiding Principles for Sharing Clinical Trial Data. "https://www.ncbi.nlm.nih.gov/books/NBK285999/".

[10] J. C. Kvedar and A. L. Fogel, "mhealth advances clinical research, bit by bit," *Nat Biotech*, vol. 35, no. 4, pp. 337–334, 4 2017.

[11] O. Sangpetch and A. Sangpetch, *Security Context Framework for Distributed Healthcare IoT Platform*. Cham: Springer International Publishing, 2016, pp. 71–76.

[12] N. Vithanwattana, G. Mapp, and C. George, "Developing a comprehensive information security framework for mhealth: a detailed analysis," *Journal of Reliable Intelligent Environments*, pp. 1–19, 2017.

[13] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, 2016. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2016.2566339

[14] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control," *Journal of Medical Systems*, vol. 40, no. 10, p. 218, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10916-016-0574-6

[15] "RIOT-OS: The friendly Operating System for the Internet of Things," https://www.ethereum.org/.

[16] "Ethereum: Blockchain App Platform," https://www.ethereum.org/.

[17] "IoT-LAB: a very large scale open testbed," https://www.iot-lab.info/.

[18] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] "EEG Eye State Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

[21] I. Saini, D. Singh, and A. Khosla, "QRS detection using k-nearest neighbor algorithm (KNN) and evaluation on standard ECG databases," *Journal of Advanced Research*, vol. 4, no. 4, pp. 331 – 344, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S209012321200046X

[22] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.