Neural Computing and Applications (2022) 34:11491–11505
https://doi.org/10.1007/s00521-020-04873-z

ORIGINAL ARTICLE

MNSSp3: Medical big data privacy protection platform based on Internet of things

Xiang Wu[1,2] · Yongting Zhang[1,2] · Aming Wang[1,2] · Minyu Shi[1,2] · Huanhuan Wang[1,2,3] · Lian Liu[2]

Received: 23 December 2019 / Accepted: 20 March 2020 / Published online: 23 May 2020

Abstract

How to transform the growing medical big data into medical knowledge is a global topic. However, medical data contains a large amount of personal privacy information, especially electronic medical records, gene data and electroencephalography data; the current methods and tools for data sharing are not efficient or cannot be applied in real-life applications. Privacy disclosure has become the bottleneck of medical big data sharing. In this context, we conducted research of medical data from the data collection, data transport and data sharing to solve the key problems of privacy protection and put forward a privacy protection sharing platform called MNSSp3 (medical big data privacy protection platform based on Internet of things), which attempts to provide an effective medical data sharing solution with the privacy protection algorithms for different data types and support for data analytics. The platform focuses on the transmission and sharing security of medical big data to provide users with mining methods and realizes the separation of data and users to ensure the security of medical data. Moreover, the platform also provides users with the capacity to upload privacy algorithms independently. We discussed the requirements and the design components of the platform, then three case studies were presented to verify the functionality of the platform, and the results of the experiments show clearly the benefit and practicality of the proposed platform.

Keywords Privacy protection · Platform · Gene data · EMR · EEG data

1 Introduction

Medical data is characterized by volume, velocity, variety and value (4V). With the rapid development of information technology, data mining makes medical data become an information asset with stronger decision-making power, insight, discovery and process optimization capability. Moreover, it is an important force to promote the development of medical research [1–4]. Medical big data has various forms [5–7], among which EMR and gene data have become the main research hot spots. In addition, with the in-depth study of brain–computer interface and deep learning, EEG signals have become an important data resource for medical data. However, while medical big data is fully utilized, great risks of privacy disclosure are exposed [8–12]. Especially, the mining process of EMR data, gene data and EEG data as the main research objects is very easy to disclose personal privacy information.

✉ Xiang Wu
wuxiang@xzhmu.edu.cn

Yongting Zhang
yazimai1018@163.com

Aming Wang
wamsinx@163.com

Minyu Shi
shiminyu163@163.com

Huanhuan Wang
whhxzhmu@163.com

Lian Liu
15050846931@163.com

1 School of Medical Information and Engineering, Xuzhou Medical University, Xuzhou 221000, Jiangsu, China

2 Institute of Medical Information and Health Big Data, Xuzhou Medical University, Xuzhou 221000, Jiangsu, China

3 School of Information and Control, China University of Mining and Technology, Xuzhou 221000, Jiangsu, China

A large number of health management companies have actively collected users' sensitive information, which greatly increases the risk of consumers' privacy being leaked. In 2016, the personal health data of 918,000 seniors was leaked online for months, after a software developer working for Health Now Networks uploaded a backup database to the Internet.[1] Likely, there have been many security accidents caused by hacking of medical equipment or related mobile devices. For example, the hacker organization has stolen and published 180,000 patient medical records through three illegal incursions, which caused great harm to the patients.[2] Therefore, medical big data has exploded over the past decade; the development of privacy protection methods and tools did not keep pace with its growth. It is difficult to make full use of medical big data in the context of privacy disclosure.

The application of medical big data has its particularity. First of all, medical data is more likely to involve users' privacy than other types of data, so medical data needs to be managed uniformly by specialized agencies, and ordinary visitors do not have access to these data, ordinary users cannot get access to the data. Secondly, the application of medical big data is mainly used in medical institutions and scientific research institutions. Thirdly, the mining process is more likely to expose private information, and simply deleting personal logo information cannot achieve the purpose of privacy protection. Finally, the possibility of medical data privacy disclosure comes out at the moment of data generation. In summary, how to comprehensively protect the medical data privacy information is an urgent problem to be solved [13, 14].

Therefore, a privacy protection sharing platform integrated with processing, management, query and analysis is of great value and significance for the effective use of medical big data, and it can achieve distributed deployment of data collection, processing and sharing while ensuring data privacy. However, there are three major problems with the existing medical data sharing platform. Firstly, there are many data sharing platforms, but most of them focus on updating and sharing data without considering data privacy disclosure. In [15], the author developed a sharing biobank that integrates personal health, genome, and omics data along with biospecimens donated by volunteers of 150,000. And the University of California, Santa Cruz (UCSC) Genome Browser website (see Footnote 2) provided a large database of publicly available sequence. Until now, the site has continued to enrich its database. Secondly, the privacy protection algorithms are separated from the platform. At present, there are many privacy protection algorithms for EMR, genetic and EEG data, but they are not integrated into practical applications. In addition, at the time of data processing, it is very difficult to choose the most appropriate privacy protection algorithms according to the different data types and the desired privacy protection level [16–18], so it is very difficult to build a common intermediate platform between many privacy protection algorithms and various types of medical data. For different types of medical data, the data mining objectives and requirements are different, so the privacy protection methods and strategies are different. Even for the same type of medical data, the privacy protection methods and strategies are also different. At present, there are many research results in this field. Literature [17, 19, 20] do not only present potential genetic privacy risks, but also described three techniques for protecting human genetic privacy: controlled access, differential privacy and cryptographic solutions. In [21], authors reviewed some techniques carried out in the basis of e-Healthcare privacy protection. It also explored whether the existing researches offer any possible solutions for either patient privacy requirements on e-Healthcare or possibilities to address the (technical as well as psychological) privacy concerns of the users. Literature [22] is a review of differential privacy methods. Differential privacy emerged as a new model for privacy preserving with strong privacy guarantees. By resisting adversaries with any background knowledge and preventing attacks from untrustworthy data collector, differential privacy can protect private information thoroughly [23]. Thirdly, the existing platform cannot guarantee multiple privacy protection levels of medical data from perception to application. From the generation of medical data, it is exposed to multiple risks including management negligence, network attack and mining technology attack. At present, most data sharing platforms are only able to withstand unilateral risks [24, 25].

To meet above challenges, this paper proposes a privacy protection scheme for building a platform that can provide strong privacy protection for medical data sharing. Our scheme is implemented by involving the differential privacy technology and encryption techniques. In this framework, the platform authenticates the data submitted from devices and provides different privacy protection services according to different users' query requirement. We exploited the properties of modular arithmetic to design a data sharing platform which is efficient and has the capability of privacy preserving.

The privacy protection data sharing platform proposed in this paper is designed to enhance the privacy protection of the medical data life cycle including perception layer, transport layer and application layer. The first layer is mainly to complete the perception and collection of data.

[1] http://www.healthcareitnews.com.

[2] https://www.sohu.com/a/197611957_104421.

The data collected in this layer has its particularity in its collection methods. The security of the collected data and the original data uploading process of the database is completed by the second layer; encryption technology is used in this layer. And the application layer is the third layer that realized data privacy protection and user defined code standards.

In summary, the contributions of this paper are:

1. We proposed a novel privacy protection scheme for the secure sharing of medical big data. This scheme enables users to mine data according to different needs without touching the original data.
2. We built a security data sharing platform. The platform is mainly used for the research related to personal medical data such as EMR, genes and EEG, and provides privacy protection for the research process. Analyzed the characteristics of different data to meet the users' requirement by selecting the matching algorithms according to different mining tasks, and realized the sharing of medical big data privacy protection algorithms.
3. Our platform allows users to use customized codes for personalized data analysis, established the upload standard of codes and data, including charts, tables, command lines and other standards, designed the online visual analysis methods.
4. The privacy information of medical data is protected by adding noise to the query results, and the requirements of user mining query are met at the same time. The case studies shows that the scheme is safe and reliable.

The remainder of the paper is structured as follows: Sect. 2 summarizes a comprehensive survey of the existing related work. Section 3 explains the proposed medical data privacy protection solution in detail. Section 4 uses three case studies to verify the practicability of the platform. Section 5 includes a discussion on the results of the experiments. Finally, Sect. 6 presents the conclusions of our work.

2 Related work

This paper focuses on the security sharing of data mining results. The focus of sharing mining results is to study different types of medical data, so as to design the sharing scheme to meet the requirements of privacy protection. In this section, we described the classification of medical data types, introduced the privacy protection research methods and tools for different data types, and then presented the work related to the solution proposed in this paper.

Medical big data has various forms, among which EMR data, gene data and EEG data have become the main research hot spot.

2.1 EMR data

EMR represents longitudinal data (electronic format) that are collected during routine medical data. EMR data mining plays an important role in medical diagnosis, medical management and scientific research. However, EMR is rich in personal information, and there is a great risk of privacy disclosure in the process of mining and sharing. The EMR research is mainly about mining and prediction, both of which require data to be obtained and then mined; such privacy leakage risks are very high. For the statistical analysis of EMR [26] introduced the differential privacy method, which improved the accuracy and privacy of electronic medical record. The implementation of the prediction algorithm is based on a large amount of EMR data to build a prediction model. This process needs to obtain EMR data, which will lead to privacy disclosure. Therefore, it is urgent to find a privacy protection method to solve the above problems during the EMR data mining process.

2.2 Gene data

With the development of genome sequencing technology, human genome data has been widely used in biomedical research with great biological value. Gene data mining can effectively promote the development of biomedicine. DNA genes contain a large amount of personal information, and the genetic data cannot be fully utilized without effective privacy protection methods. At present, there are many research methods of genes, just like motif finding, genome-wide association study (GWAS), etc. In the motif finding process, since the DNA sequence used contains a lot of information about personal characteristics, physiological functions, and diseases, it is easy to leak personal privacy. Homer et al. [27] proved that a person's specific identity could be identified from a set of DNA data. After that, Gymrek et al. [28] showed that it is possible to re-identify 50 DNA participants from the 1000 Genomes Project dataset. The privacy disclosure problem like this happened all the time [17, 29]. Homer et al. [30] publicly released GWAS statistics could be used to estimate a GWAS participant's disease status from knowing his/her genotypes at certain risk factors. In order to reduce the accident of these risks of privacy disclosure, there are three main methods from the status of genetic privacy protection technology that purpose to protect genetic privacy from various perspectives: controlled access [31], differential privacy preservation [32, 33] and cryptographic solutions [34, 35].

The current gene data privacy protection study insufficient and many problems have not been solved well. Since ordinary data sharing platforms cannot provide targeted privacy protection for medical data, our platform should meet the following requirements: ensure the privacy of genetic data sharing; integrate different types of data mining algorithms to meet user query needs; provides basic functions such as data upload, data usage, and data sharing.

2.3 EEG data

Brain–computer interface (BCI) is a kind of communication system that does not depend on the normal output pathway which is composed of peripheral nerves and muscles. The purpose of BCI is to enable humans to express ideas or manipulate devices directly through the brain [36]. Through BCI equipment, we can capture neural signals and extract features from them, then use these features to train through various machine learning and artificial intelligence models, and finally achieve the goal of prediction and inference. Researchers focused on how to record neuron information and stimulate neurons through noninvasive and invasive ways. In the non-intrusive field, the Carnegie Mellon University team has developed a noninvasive brain–computer interface that allows people to manipulate a robotic arm with ideas. Recently, Facebook released the research results of brain–computer interface, which detects the oxygen consumption of neurons through pulse oximetry, detects the expected speech in brain activity in real time and realizes "intelligent typing" [37]. In the invasive field, scientists have long carried out tests on brain implants, allowing patients to move cursors or robotic arms. Cochlear implants, artificial retinas and other brain–computer interface products have played an active role in helping people with disabilities to restore their impaired ability. With the rapid development of BCI, people can quickly get a person's idea which is the privacy of this person through the stored EEG data. EEG data is so rich with information that researchers can easily gain knowledge beyond the professed scope from unprotected EEG signals, including passwords [38]. Giving access to a user's brain signals, or features extracted from them, can seriously harm the user's privacy. There is plenty of evidence that anonymizing data does not offer sufficient protection [39]. Attackers can still use the attack (background attack, attribute attack, differential attack, link attack, etc.) to obtain valid information. Therefore, researchers proposed to use encryption to achieve the privacy and security of brain–computer interface applications [40]. Literature [41] uses secure multiparty computation (SMC) to perform linear regression (LR) over EEG signals from many users in a privacy-preserving fashion. In general, there are relatively few studies on EEG signals. In order to protect the privacy of users' EEG signals, we will completely isolate the EEG data from the researchers, so that users can complete the training and research of EEG signals without getting the EEG data in the cloud.

However, the above kinds of data privacy protection methods and tools have different defects and limitations, moreover, most of the existing data sharing platforms do not involve the data perception layer, which lacks the dynamic update link of data. Aiming at the above three kinds of data privacy disclosure problems, this paper proposed a unified solution to achieve the separation of users and data, provided mining solutions for different data types and designed personalized services. Considering the leakage of medical big data related to perception, transmission and application, we put forward a privacy protection scheme based on the Internet of things framework. In the transmission layer, a cryptographic mechanism is used to focus on data processing at the application layer, and the platform standard is formulated.

3 Model and system

This section explains in detail the proposed medical data privacy protection solution. Privacy protection in medical data sharing is a key innovation of our platform, we focus on the data privacy protection algorithms and platform services in the application layer; then, the perception layer and transmission layer of medical data are described closely. The platform architecture is shown in Fig. 1.

3.1 Application layer

The application layer of the platform provides data mining services for users, and its privacy policies are mainly divided into two categories: one kind is that "separating users from data" + "protecting with differential privacy algorithm," which mainly provides gene data and EMR data mining query services. We call this scheme as "*SP*1" method. Another kind is "separating users from data" + "predicting," which mainly provides intelligent diagnostic research services by this privacy strategy; it can be used for EMR data or EEG data. We call this scheme as "*SP*2" method.

SP1 is based on the method of isolating the data owner from the data usage process and provides a mechanism for adding noise (Fig. 2). The mining process usually involves various complicated mining methods, even without direct access to the data, this privacy information can still easily lead to privacy leakage. For instance, the research on motif finding of gene data, almost all network attacks (such as background attack, attribute attack, differential attack and

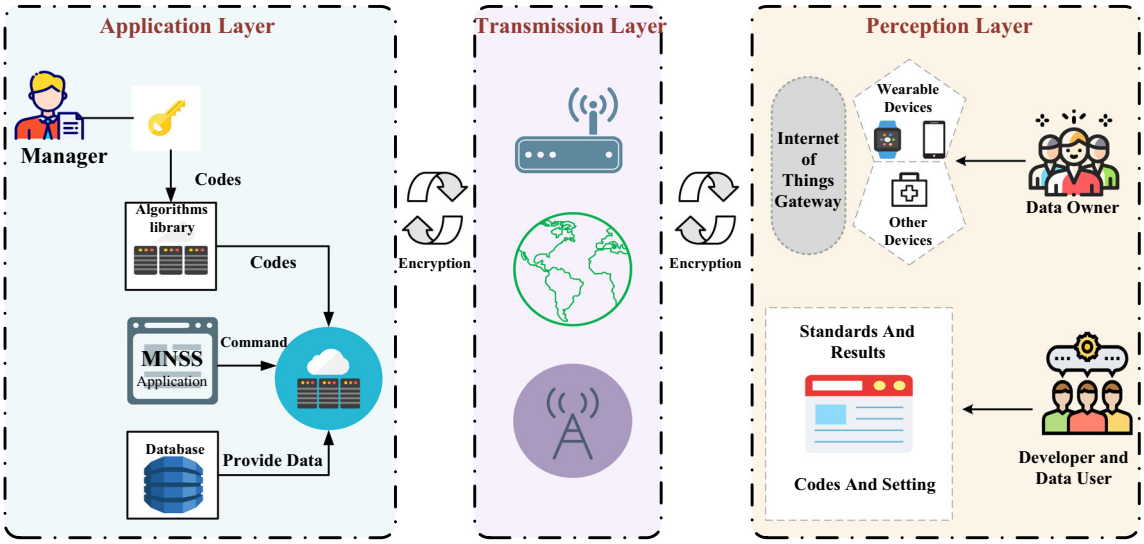

Fig. 1 The medical privacy protection platform architecture

link attack) can produce privacy attacks on the process of DNA motif finding. Especially in dynamic interactive query, by repeatedly adjusting parameters or DNA motif for search query, the personal privacy information contained in DNA data can be easily obtained by using the output information, which leads to even if we do not get DNA data. It will also cause privacy disclosure through the remote query output information [23, 42]. For such queries, even if the data is separated from the users, the query results need to be protected. Differential privacy technology is a relatively strong privacy protection technology. Datasets that satisfy differential privacy can resist any analysis of privacy data. Therefore, we used the differential privacy algorithms in the platform application layer to protect the mining results.

Definition 1 (*Differential privacy* [43]) Given arbitrary two database D_1 and D_2 differing by at most one record, a randomized privacy mechanism G achieves ε-differential privacy if for D_1 and D_2, for any possible output $O \in Range(G)$:

$$P_r[G(D_1) \in O] \leq e^{\varepsilon} \times P_r[G(D_2) \in O] \quad (1)$$

where ε indicates privacy budget and ε-differential privacy guarantees powerful privacy protection against wide background knowledge. To append moderate noise and achieve differential privacy, global sensitivity plays a significant role in determining the count of the noise.

However, different users have different mining methods, and the accuracy requirements of query results are also different. Therefore, we have built a library of popular privacy protection algorithms. The users can select data and algorithms to mine information and set relevant parameters to get the mining results. In addition, users can upload the mining codes according to their own requirements, which must obey the platform standards and meet privacy budget requirements. This method has three advantages. First, it solves the security problem of data sharing. Second, it provides medical database for users. The third is to integrate different privacy protection algorithms, which can be used by users through the platform interface without looking for code debugging, and greatly improving the efficiency.

The *SP*2 privacy protection method is mainly for the research of disease prediction and diagnosis using medical data, which generally requires training a large amount of medical data to obtain a better prediction model. It is obviously unreasonable to provide data directly to users for research, but there is currently no good way to ensure the privacy of the data in the case of sharing data. In order to solve such problems, the platform separates users from medical data and provides different training prediction models for different data types (EMR or EEG data). Users can directly use the disease prediction model provided by the platform or upload users' own codes to train new prediction models by choosing the medical database in the platform. The platform provides abundant computing resources which can dynamically increase CPU and memory with the method we proposed in [55, 56], this ensures that users can get greater efficiency. For the related research of disease diagnosis or prediction, the privacy strategy provided by the platform can greatly improve the security of data usage for the scientific research. Figure 3 shows the security strategy of *SP*2.

To achieve more functions, the platform reserved the interfaces for users to upload the customized code, which

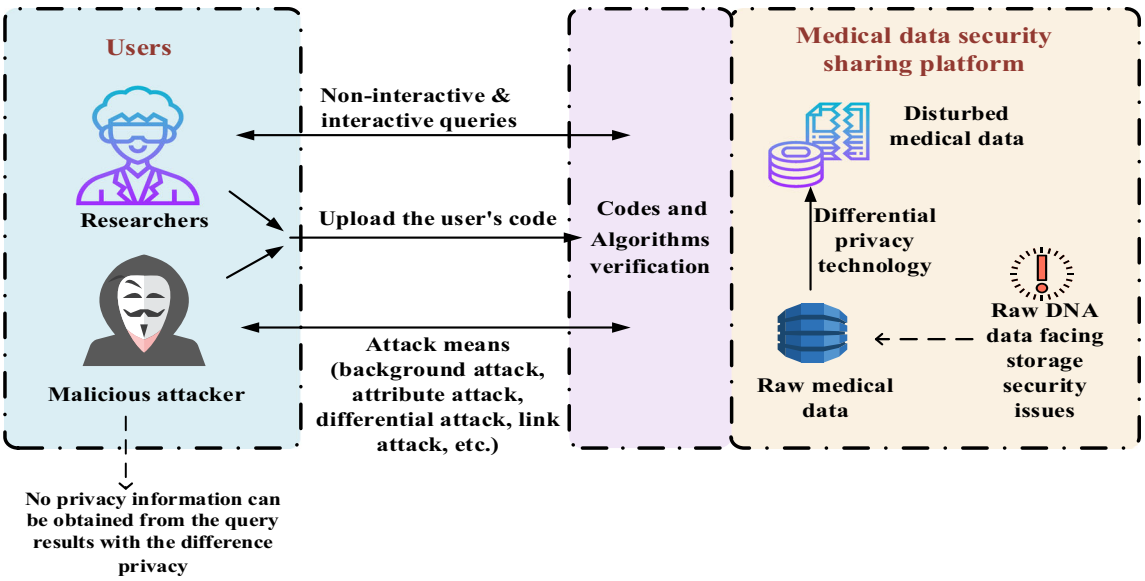

Fig. 2 Service procedure for *SP1* with security strategy

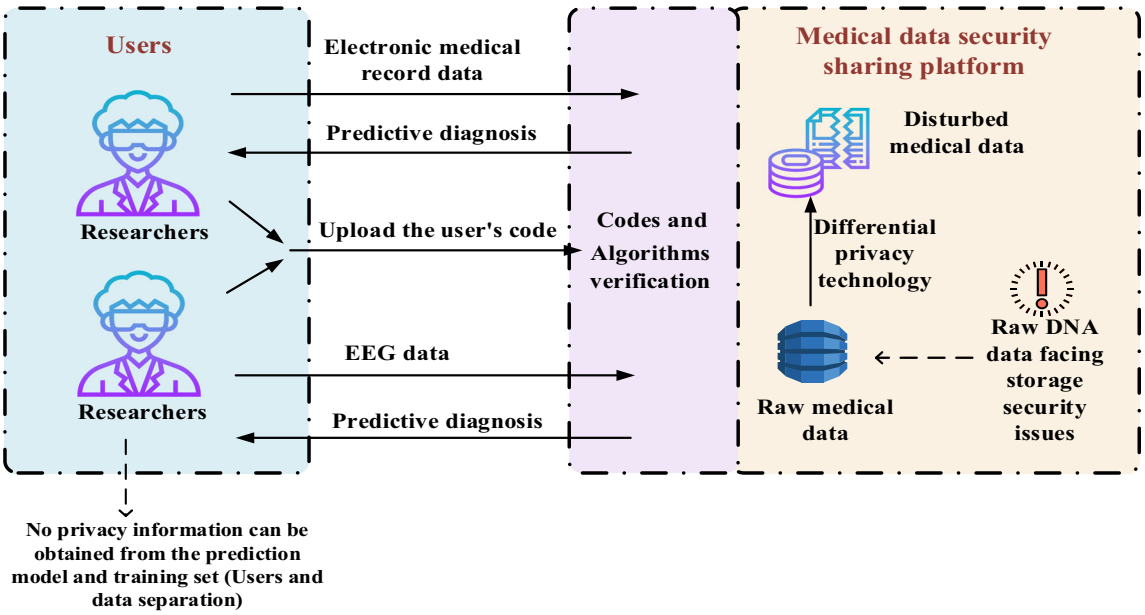

Fig. 3 Service procedures for *SP2* with security strategy

stipulates that the code uploaded by the user must conform to the platform standard. In the standard of uploading code, the setting of privacy budget ε is very strict. As long as the code uploaded by users includes mining function, the privacy budget must be greater than 0.3 [44, 45]. The platform verified that whether the privacy budget of the user uploading code meets the standard value.

3.2 Transmission layer

In this layer, all medical data collected from the previous layer is conveniently stored in the database to determine different solutions, this layer can preprocess real-time data, which ensures the quality of the data collected.

The privacy protection platform in this paper mainly realizes data protection in data transmission and data application. The encrypted process of updating data is

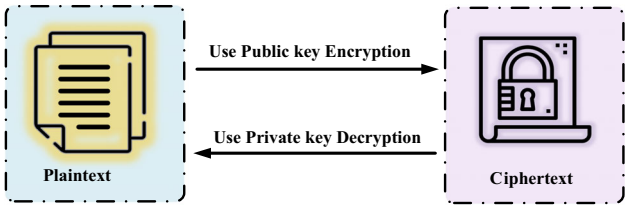

Fig. 4 The encrypted process of updating data

shown in Fig. 4. Privacy protection algorithm used in the application layer mainly to protect data to ensure the separation of users and data. At the transport layer, we used a lightweight privacy approach for uploading data to Internet of things devices [46–48].

The RSA encryption algorithm is a typical public key cryptography algorithm (also known as asymmetric cryptography). The main idea is to use different keys for encryption and decryption. The conventional key is divided into two, that is, the encryption key K_e and the decryption key K_d. K_e can be offered as a public key, and K_d as a secret private key. The algorithm can guarantee that the private key cannot be derived from the public key, nor can the plaintext be derived from the ciphertext. The RSA encryption process has the three main steps.

3.2.1 Key generation

First, let $R = p \times q$. p and q are two prime numbers that are large enough and adjacent. Strictly keep p and q secret, but open R. Next, calculate the value of the Euler function by the formula $\varphi(R) = (p-1) \times (q-1)$, and open $\varphi(R)$. Then, an integer e is selected in the range of $(1,\varphi(R)))$, which also satisfies $\gcd(\varphi(R), e) = 1$ (indicating that $\varphi(R)$ and e are prime to each other). Calculate $d = e^{-1} \bmod (\varphi(R))$. Finally, get the public key $K_e = <e, R>$ and the private key $K_d = <p, q, d, \varphi(R)>$.

3.2.2 Encryption

The message M is encrypted by raising it to the *eth* power modulo R. The result is called the ciphertext of M

$$C = M^e \bmod R. \tag{2}$$

3.2.3 Decryption

A ciphertext C, for a given message M, is decrypted by raising it to the dth power modulo R. From Lagrange's theorem, it follows that:

$$C^d \bmod R = M^{ed} \bmod R = M \bmod R = M \tag{3}$$

where $ed = 1 + k \times \varphi(R)$ and k is a positive integer.

3.3 Perception layer

The rapid development of the smart medicine indicates that connecting personal health data to the Internet through IoT devices will become the norm in the near future [49], which can provide accurate medical services for individuals and provide more medical data resources for scientific research institutions or medical institutions by the efficient and secure algorithms and communication mechanisms. This platform database stores the data mainly derive from the public database, cooperative hospitals and scientific research institutions, the perception layer must consider the issue of data updating. The platform provides a data update interface, which supports to update the medical data and access into the platform in real time. Personal health information can be monitored and transmitted to the platform through the IoT device to support the user's personal health management.

The platform mainly provided three types of medical data, electronic medical records, gene data and EEG data. The original data is stored in the database after being preprocessed, while the later updated data needs to be monitored and collected by sensing equipment, and different data types have different update processes.

3.3.1 EMR data

The original data storage database of EMR has been preprocessed and stored according to the data format standard. However, the information of personal physical diseases changed over time. The updated information can be detected by wearable or embeddable devices regularly, which can facilitate the platform application layer to update data mining results.

3.3.2 Gene data

Gene data mainly comes from sequencing institutions, and a large amount of gene data are generated everyday. Therefore, API interface is reserved in the platform to automatically update the gene data provided by automatic sequencing equipment, which can continuously expand the gene database.

3.3.3 EEG data

The collection of EEG signals is mainly from wearable devices and medical database. The data is encrypted and transmitted to the cloud database, where it is cleaned, processed and annotated.

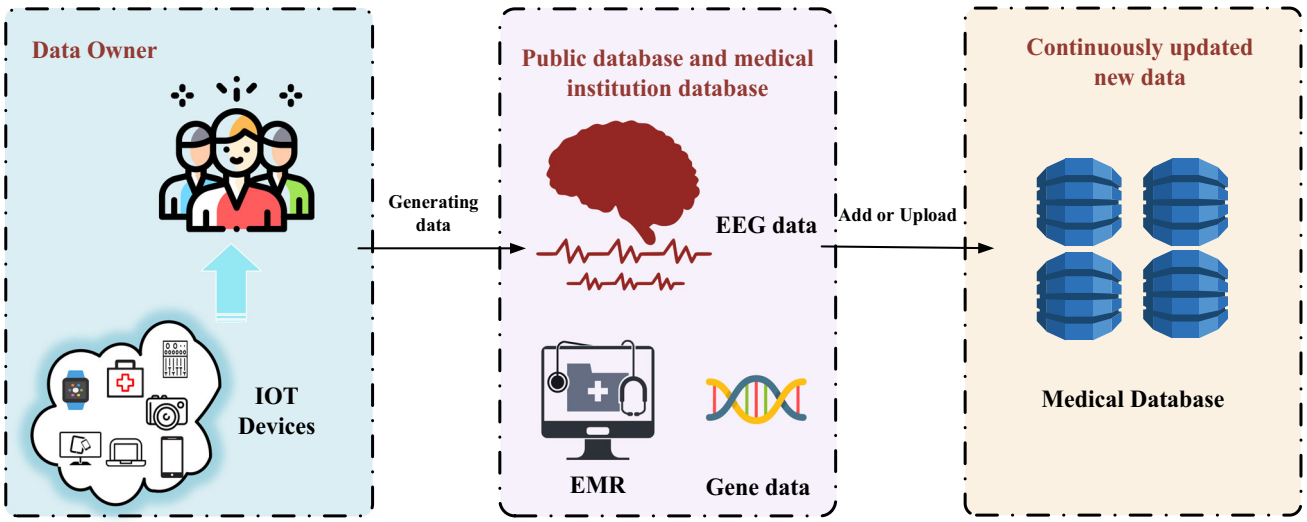

Fig. 5 Data perception layer based on the IoT

Data perception layer based on the IoT is shown in Fig. 5. This layer is in charge of connecting physical devices or actuators that are going to provide the updated or increased medical data to the platform. Once this is done, it will map the collected data to the platform and send the mapped information to the storage database.

3.4 Platform standard

The platform requires that mining process must be performed online through cloud computing without data download service. Therefore, the platform developed the standards for input data, output data and pecifications for different types of data.

3.4.1 Input standards

The platform provided developers with an interface to get training data, which is limited to server local access. The developers can integrate the SDK provided by the platform into the program. When the developer's program is running in the cloud, the data can be used through the interface, and the data structure is json.

3.4.2 Output standards

The platform integrated the data visualization module, which can be used by developers to visually display the results of cloud computing, and the calculation results are allowed to be downloaded and saved. In the SDK provided by the platform, the files in the results folder can be downloaded at the end of the program. Developers can save program logs, running results or other files that need to be downloaded in this path.

3.4.3 Data standards

In order to ensure the data versatility of the platform, we required data owners to share data in accordance with data format standards. Taking the EEG data as an example, the format of the EEG data and the data label must be uploaded in accordance with the platform requirements. The cleaned EEG data and training model need to be stored separately and set the corresponding access permissions.

4 Validation of the platform

Based on the malleable network system simulator (MNSS)[3] platform, we have developed a medical big data privacy protection platform called MNSSp3 (MNSS privacy protection platform),[4] which integrates many data sharing and data mining tools. The MNSS platform is a cloud computing platform, which integrates many excellent open-source software such as Eve-ng, GNS3 GUI, Dynamips, Dynagen, QEMU, GNU Health, OpenLIS and OpenSourcePACS. We added many resource scheduling optimization algorithms at the backend of the platform, which further improves the efficiency of the platform [55, 56].

In order to test the feasibility of the proposed platform, three case studies have been instantiated that allowed us to evaluate functionalities of the platform. Here, we provide some details of the evaluation scenario. The display of the privacy protection platform is shown in Fig. 6.

[3] http://www.mnss.xzhmu.edu.cn.

[4] http://www.mnssp3.xzhmu.edu.cn.

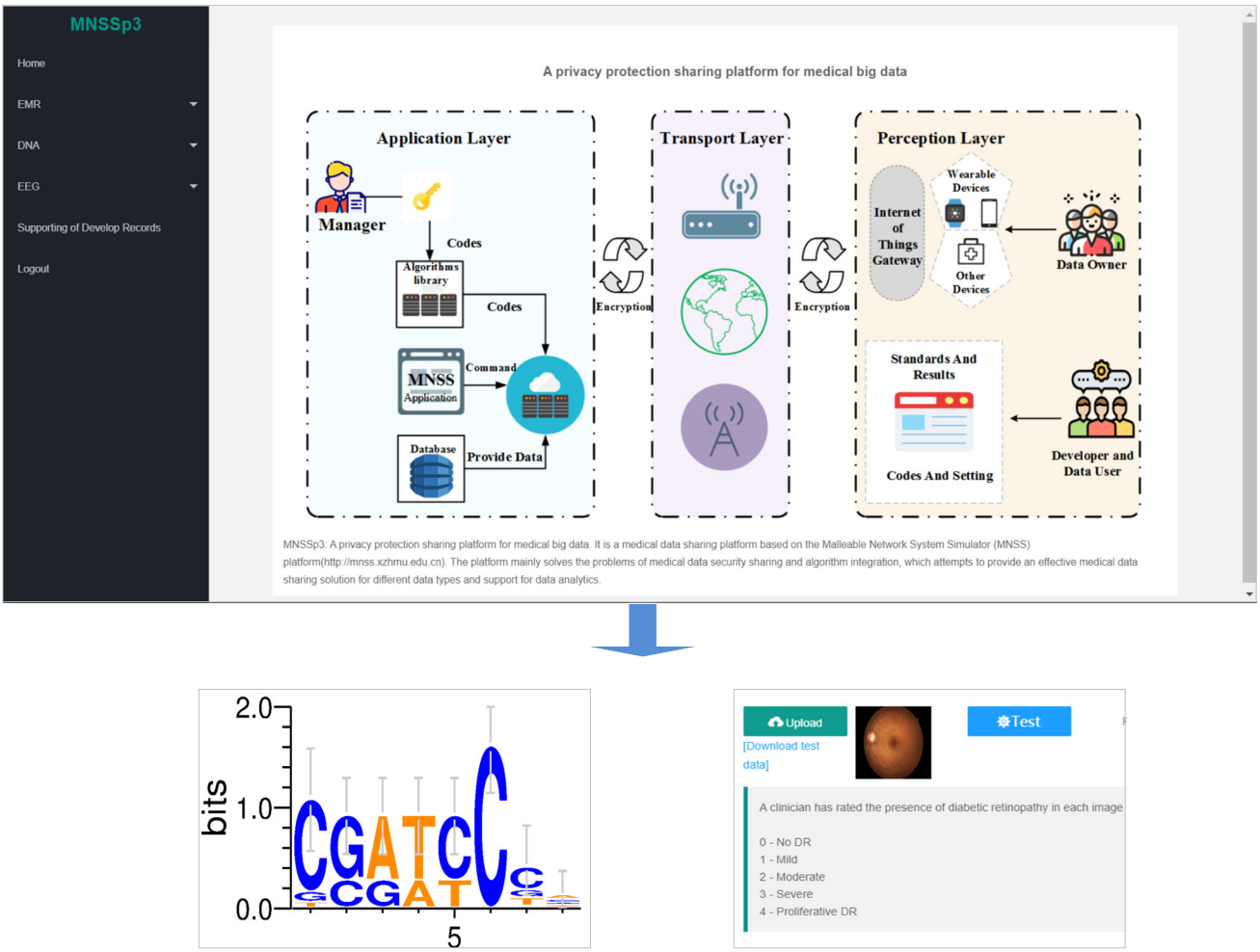

Fig. 6 The display of the privacy protection platform

4.1 Case study 1

Case study 1 belongs to *SP*1 service, which mainly provides data mining service. The query results need to be further processed in privacy to prevent malicious attacks from attackers with specific background knowledge. In gene data research, the privacy disclosure of DNA gene data is becoming more and more serious [27, 50], and an individual's private information is easily leaked in the process of discovering motifs because DNA sequences contain a large amount of private information about personal characteristics, functions, illnesses and personality disorders [28, 51, 52]. In recent years, DNA datasets have caused a serious problem about privacy disclosures. The privacy protection research of motif finding has aroused widespread social concern. Therefore, protecting gene data on the research process does not reveal private information is an important task of this platform. We took the *N*-gram algorithm of motif finding algorithm as an example to illustrate the privacy protection process of DNA motif finding. *N*-gram algorithm belongs to the exact algorithm; its calculation result is the frequency statistics of the motif's conservative sites [53, 54]. *N*-gram algorithm has powerful capabilities for modeling sequence data and has been widely used in computational biology.

All input files and the outputs generated for this case study are contained in supplementary file. The datasets utilized are two real-life DNA datasets, Washington[5] and Upstream.[6] Both datasets were preprocessed by the established measure. In the final test dataset, Washington and Upstream contain 14,126 and 487,760 sequences, respectively.

The platform supported direct input of DNA gene sequences or fasta format files for DNA motif finding. Figure 7 shows the process of motif finding with secure *N*-gram algorithm. After the DNA gene dataset for motif finding is ready, set Hamming distance, motif length and privacy budget ε. Different Hamming distances, motif

[5] http://www.bio.cs.washington.edu/assessment/download.html.

[6] http://www.hgdownload.soe.ucsc.edu/downloads.html.

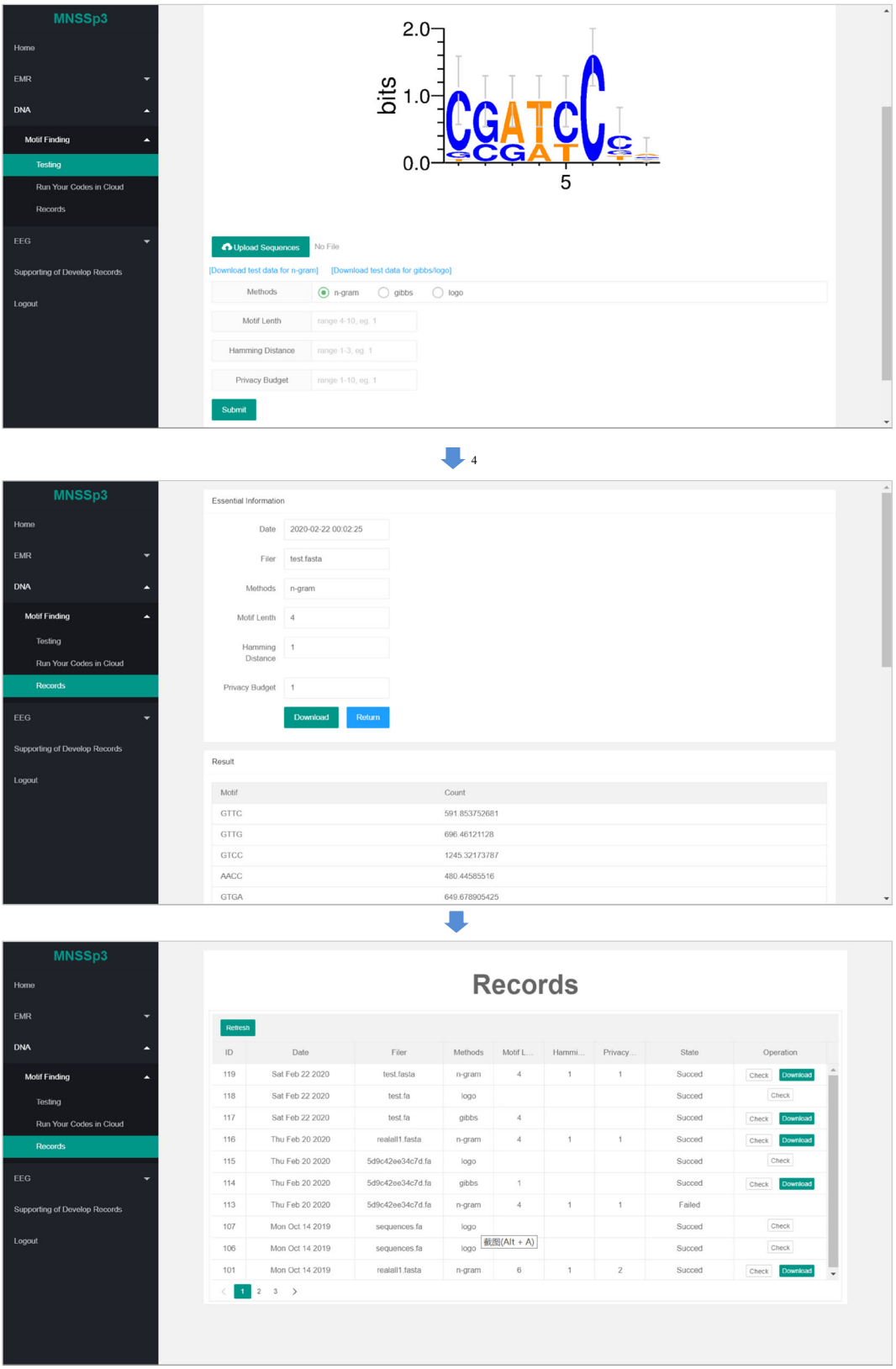

Fig. 7 The process of motif finding with secure *N*-gram algorithm

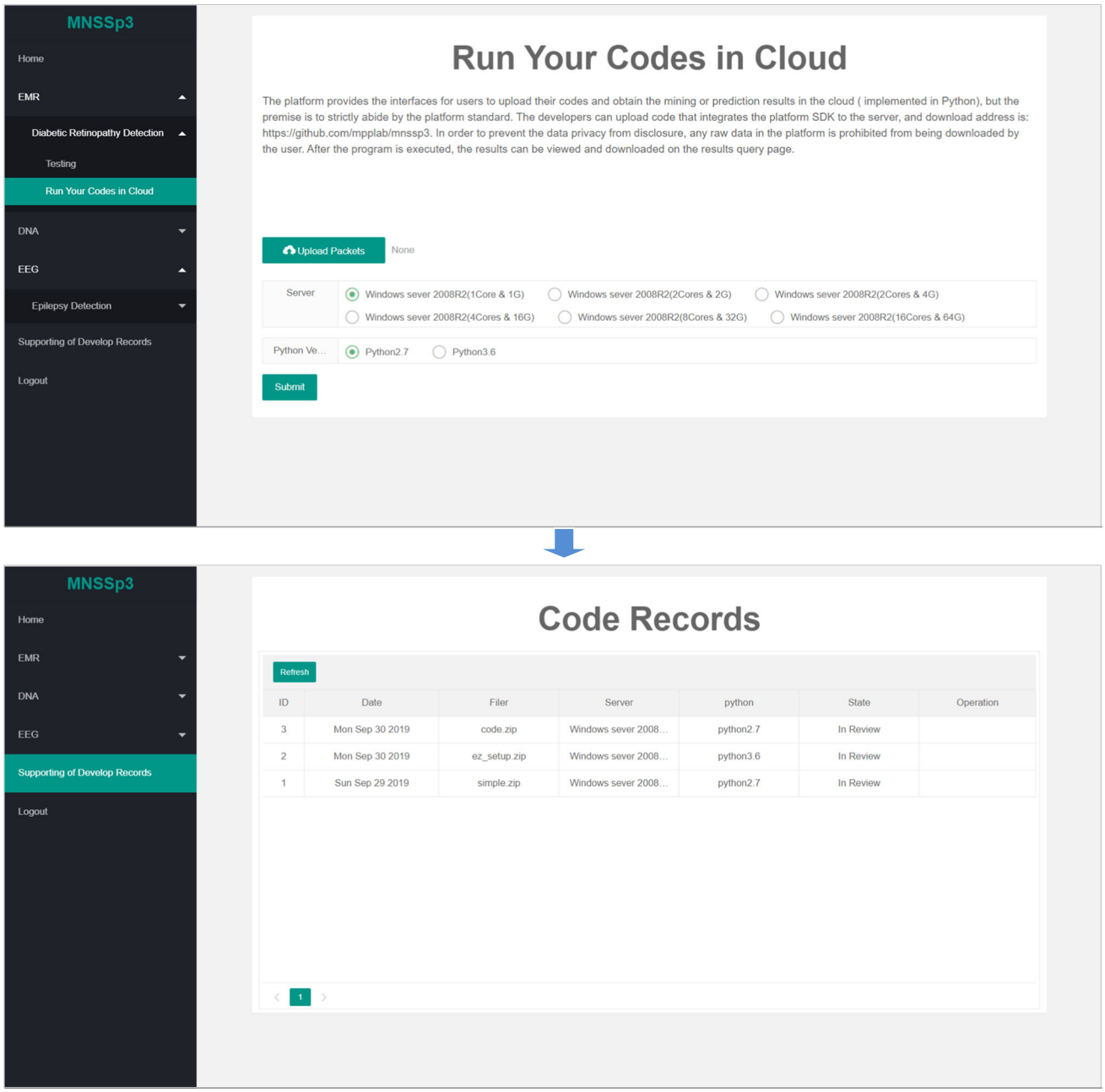

Fig. 8 The process of uploading codes

length and privacy budget ε together determine the calculation time, which can vary from a few minutes to several hours. In addition, the corresponding classic literature is provided at the bottom of the page, which can be downloaded and read. When the motif finding calculation is completed, the platform will send a reminder message to the user's mailbox, users can view at the user center.

4.2 Case study 2

Case study 2 belongs to SP1 and SP2 service. The platform provided a variety of data mining and predicting services (implemented in Python). For each services, we reserve the interface for the users to upload the code and obtain the mining or prediction results in the cloud.

Figure 8 shows the process of uploading codes; users can run their codes in cloud. In developing the support module, developers can upload codes that integrate the platform SDK to the server. The SDK of MNSSp3 can be

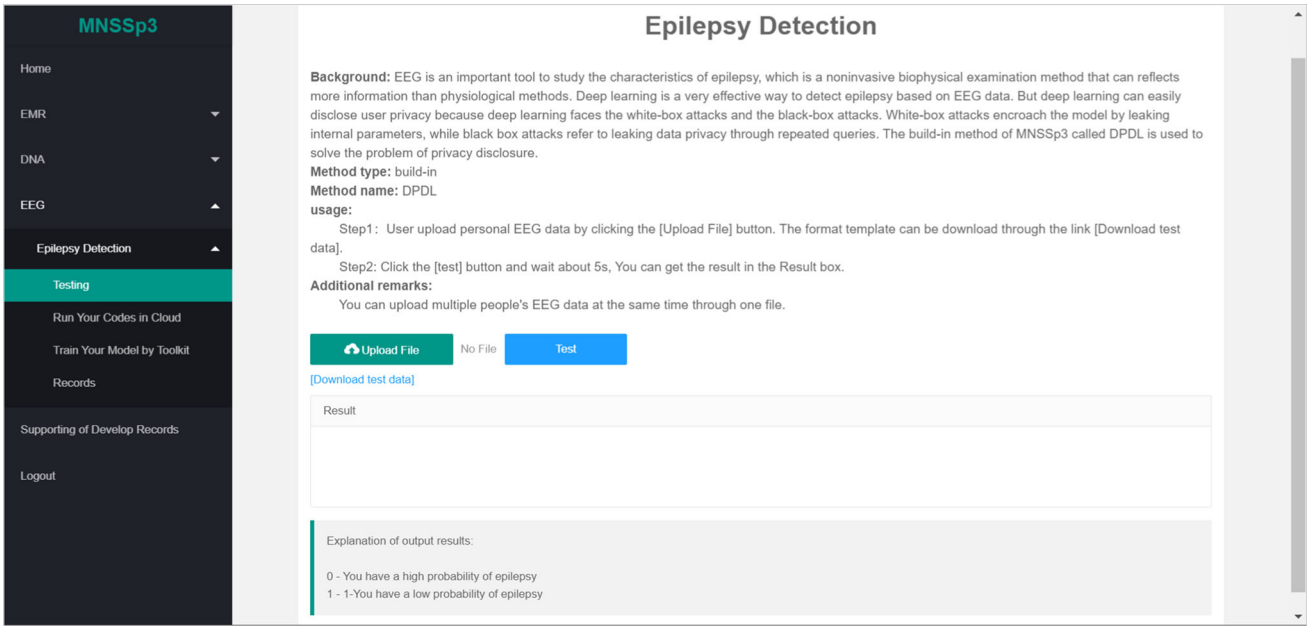

Fig. 9 Prediction process through built-in security model of MNSSp3

download in the URL https://github.com/mpplab/mnssp3. In order to prevent the data privacy from disclosure, any raw data in the platform is prohibited from being downloaded by the user. The program integrated with the SDK obtains the dataset in the database when executed on the server. In addition, the SDK specifies the format standard of data visualization and output. Developers can save the output results in the SDK and check the execution of the uploaded code in the personal center. The results that meet the criteria can be viewed and downloaded on the results query page.

4.3 Case study 3

Case study 3 belongs to *SP2* service. The *SP2* privacy protection method is mainly for the research of disease prediction and diagnosis using medical data, which generally requires training a large amount of medical data to obtain a better prediction model. In this case, we take epilepsy detection as an example; EEG is an important tool to study the characteristics of epilepsy, which is a noninvasive biophysical examination method that can reflects more information than physiological methods. Deep learning is a very effective way to detect epilepsy based on EEG data, but deep learning can easily disclose user privacy because deep learning faces the white-box attacks and the black-box attacks. White-box attacks encroach the model by leaking internal parameters, while black-box attacks refer to leaking data privacy through repeated queries. In order to illustrate the problem, we use a relatively simple but effective privacy protection method: The data used in the training of deep learning is processed by differential privacy method before training, it can prevent the internal parameters or output results of deep learning from disclosing privacy, and this method called DPDL is a build-in method in MNSSp3. Figure 9 shows the prediction process through built-in security model of MNSSp3.

In MNSSp3, there are three ways to use the privacy protection method: (1) Users can use the build-in method of MNSSp3. (2) Users can write their own privacy codes and run the codes in cloud. (3) If users have a large amount of privacy data and they think the training effect of the built-in method in MNSSp3 is not ideal, but users are not willing to write there own code and download the Toolkit of MNSSp3, which contains the built-in privacy protection method, users can train the new model according to the instructions and upload the new trained model to MNSSp3 for use. Figure 10 is the process of train users' model by Toolkit of MNSSp3.

5 Discussion

The personal information usage is a topic of global discussion with regard to the privacy protection while promoting scientific advancement. People increasingly need a highly secure platform to collect, analyze and share personal health data. The scheme proposed in this paper provides a highly secure platform and rich computing medical resources.

The design idea of the platform is to protect medical information privacy according to the needs of query results

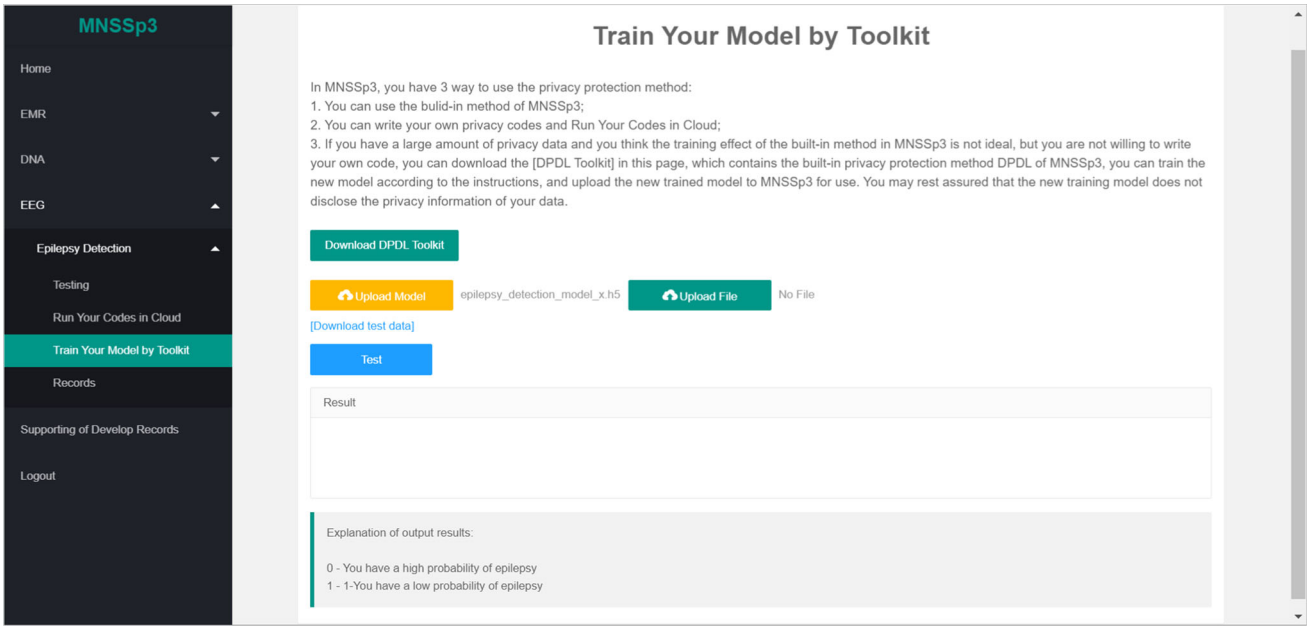

Fig. 10 Train your model by Toolkit of MNSSp3

while ensuring the separation of users and data. Because we find that different data mining methods may reveal privacy, such as motif finding methods, the privacy can still be leaked without touching the gene data. Besides health data, specificity of genome data results from certain essential features: (1) an association with traits and certain diseases, (2) identification capabilities and (3) revelation of family relations. On the research of DNA motif finding, almost all network attacks (such as background attack, attribute attack, differential attack and link attack) can produce privacy attacks on the process of DNA motif finding. Especially in dynamic interactive query, by repeatedly adjusting parameters or the number of DNA motifs for search query, the personal privacy information contained in DNA data can be easily obtained by using the output information, which leads to even if the attackers do not get DNA data. Therefore, for such mining queries, differential privacy technology is generally used to add noise to the query results, which ensures that it is impossible to disclose private information in any query mode. Similarly, data mining methods for EMR data and EEG data also face their own privacy disclosure problems.

Integrating data mining support programs as part of the platform makes it easy to develop the ultimate service for medical data mining. In this sense, developers only need to focus on the actual functionality of the services associated with the prediction algorithm or privacy protection algorithm, as the platform already provides other important tasks for data analysis, such as data preprocessing or algorithm analysis. In addition, the ability to run their own algorithms in the cloud enhances the versatility and scalability of the platform.

6 Conclusion

Although the fact that many researchers have studied the privacy and security of medical data, but there is no comprehensive scheme that fully satisfies the privacy requirements. And there are also various problems in the combination of privacy protection methods and data sharing platforms. Therefore, the security sharing of medical data is still a hot issue that has not been fully solved. In this paper, we built a unified multi-functional security sharing platform to solve the above problems from data security sharing, the actual requirements of researchers and the platform efficiency, which integrates the functions of medical data mining, model training, disease diagnosis, etc. The platform focuses on designing different privacy protection schemes based on the privacy risks that may arise from different medical data and encrypts the transmission process of updated data at the transport layer. The innovation of the proposed scheme in this paper is that the platform can provide users with different mining methods, models and computing resources while ensuring the security of medical data. At the same time, users can upload codes according to their own mining needs, complete data mining in the cloud and download the query results. The entire platform architecture uses a range of flexible APIs to enable users to use platform services and share data. In next step, the main direction of platform optimization is

still to improve the performance of the platform and the security of data sharing, more importantly, to expand the build-in algorithms and shared medical resources, and provide data support for more users and scientific research institutions.

Acknowledgements This work was supported by the Natural science fund for colleges and universities of Jiangsu Province under Grant No. 18KJB520049 and the industry University-Research-Cooperation Project in Jiangsu Province under Grant No. BY2018124. In addition, the project received the funding support from National Scientific Data Sharing Platform for Population and Health.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Perez JA, Poon CCY, Merrifield RD et al (2015) Big data for health. IEEE J Biomed Health Inform 19(4):1
2. O'Driscoll A, Daugelaite J, Sleator RD (2013) 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 46(5):774–781
3. Zhang Y, Guo SL, Han LN et al (2016) Application and exploration of big data mining in clinical medicine. Chin Med J 129(6):731–738
4. Pashazadeh A, Navimipour NJ (2018) Big data handling mechanisms in the healthcare applications: a comprehensive and systematic literature review. J Biomed Inform 82:47–62
5. Chen Y, Ding S, Xu Z et al (2018) Blockchain-based medical records secure storage and medical service framework. J Med Syst 43(1):5
6. Vayena E, Blasimme A (2017) Biomedical big data: new models of control over access, use and governance. J Bioeth Inq 14(4):501–513
7. Bibault JE, Giraud P, Burgun A (2016) Big data and machine learning in radiation oncology: state of the art and future prospects. Cancer Lett 382:S0304383516303469
8. Jagadeesh KA, Wu DJ, Birgmeier JA et al (2017) Deriving genomic diagnoses without revealing patient genomes. Science 357(6352):692–695
9. Wang S, Jiang X, Singh S et al (2016) Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 1387:73
10. Glenn T, Monteith S (2014) Privacy in the digital world: medical and health data outside of HIPAA protections. Curr Psychiatry Rep 16(11):494
11. Nia A, Sur-Kolay S, Raghunathan A et al (2015) Physiological information leakage: a new frontier in health information security. IEEE Trans Emerg Top Comput 4:1
12. Ibrahim MHA, Zhou K, Ren J (2018) Privacy characterization and quantification in data publishing. IEEE Trans Knowl Data Eng PP(99):1
13. Adane K, Gizachew M, Kendie S (2019) The role of medical data in efficient patient care delivery: a review. J Risk Manag Healthc Policy 12:67–73
14. Peat G, Riley RD, Croft P et al (2014) Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med 11(7):e1001671
15. Rhead B, Karolchik D, Kuhn RM et al (2010) The UCSC genome browser database: update 2010. Nucleic Acids Res 38(Database Issue):D613
16. Hamid HAA, Rahman SMM, Hossain MS et al (2017) A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography. IEEE Access 5:1
17. Shi X, Wu X (2016) An overview of human genetic privacy. Ann N Y Acad Sci 1387(1):61
18. Dwork C, Roth A (2013) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407
19. Wu X, Wang H, Wei Y, Mao Y, Jiang S (2018) An anonymous data publishing framework for streaming genomic data. J Med Imaging Health Inform 8(3):546–554
20. Wu X, Wei Y, Jiang T, Wang Y, Jiang S (2019) A micro-aggregation algorithm based on density partition method for anonymizing biomedical data. Curr Bioinform 14(7):667–675
21. Gkoulalas-Divanis A, Loukides G, Sun J (2014) Publishing data from electronic health records while preserving privacy: a survey of algorithms. J Biomed Inform 50(Sp. Iss. SI):4–19
22. Sarwate AD, Plis SM, Turner JA et al (2014) Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Front Neuroinform 8:35
23. Wu X, Wei Y, Mao Y, Wang L (2018) A differential privacy DNA motif finding method based on closed frequent patterns. Clust Comput 21:1–13
24. Woodward B (1997) Medical record confidentiality and data collection: current dilemmas. J Law Med Ethics 25(2–3):10
25. Claerhout B, Demoor GJE (2005) Privacy protection for clinical and genomic data: the use of privacy-enhancing techniques in medicine. Int J Med Inform 74(2–4):257–265
26. Li Z, Roberts K, Jiang X, Long Q (2019) Distributed learning from multiple EHR databases: contextual embedding models for medical events. J Biomed Inform 92:103138
27. Malin B (2004) Protecting dna sequence anonymity with generalization lattices. Carnegie Mellon University, School of Computer Science (Institute for Software Research International)
28. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y (2013) Identifying personal genomes by surname inference. Science 339(6117):321–324
29. Angrist M (2013) Genetic privacy needs a more nuanced approach. Nature 494(7435):7
30. Homer N, Szelinger S, Redman M et al (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4:e1000167
31. Erlich Y, Williams JB, Glazer D et al (2014) Redefining genomic privacy: trust and empowerment. PLoS Biol 12:e1001983
32. Dwork C, McSherry F, Nissim K et al (2006) Theory of cryptography. In; Lecture notes computer science, vol 3876. Calibrating noise to sensitivity in private data analysis. Springer, Berlin, pp 265–284
33. Dwork C (2011) A firm foundation for private data analysis. Commun ACM 54:86–95
34. Djatmiko M, Friedman A, Boreli R et al (2014) Proceedings of the 13th workshop on privacy in the electronic society. In: Secure evaluation protocol for personalized medicine. ACM, New York, pp 159–162
35. He D, Furlotte NA, Hormozdiari F et al (2014) Identifying genetic relatives without compromising privacy. Genome Res 24:664–672
36. Boto E et al (2018) Moving magnetoencephalography towards real-world applications with a wearable system. Nature 555:657–661

37. Moses DA, Leonard MK, Makin JG et al (2019) Real-time decoding of question-and-answer speech dialogue using human cortical activity. Nat Commun 10:3096
38. Martinovic I, Davies D, Frank M, Perito D, Ros T, Song D (2012) On the feasibility of side-channel attacks with brain–computer interfaces. In: Kohno T (ed) USENIX security symposium. Proceedings. USENIX Association, pp 143–158
39. Vu K, Zheng R, Gao J (2012) Efficient algorithms for *K*-anonymous location privacy in participatory sensing. In; Proceedings—IEEE INFOCOM, pp 2399–2407
40. Cramer R, Damgård I, Nielsen JB (2015) Secure multiparty computation and secret sharing. University Press, Cambridge
41. Agarwal A, Dowsley R, Nicholas D et al (2019) Protecting privacy of users in brain–computer interface applications. IEEE Trans Neural Syst Rehabil Eng 27(8):1534–4320
42. Wenyong G, Yingjie WU, Lan S et al (2015) Frequent pattern mining with differential privacy based on transaction truncation. J Chin Comput Syst. https://doi.org/10.1007/978-3-319-89500-0_38
43. Dwork C, Roth A et al (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407
44. Huang D, Han S, Li X (2015) Achieving accuracy guarantee for answering batch queries with differential privacy. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, Cham
45. Shen H, Lu Z (2017) A new lower bound of privacy budget for distributed differential privacy. In: 2017 18th international conference on parallel and distributed computing, applications and technologies (PDCAT). IEEE
46. Sun HM, Wu ME, Ting WC et al (2007) Dual RSA and its security analysis. IEEE Trans Inf Theory 53(8):2922–2933
47. Zhang C, Zhu L, Xu C, Sharif K, Du X, Guizani M (2019) LPTD: achieving lightweight and privacy-preserving truth discovery in CIoT. Fut Gener Comp Syst 90:175–184
48. Xu B, Xu LD, Cai H et al (2014) Ubiquitous data accessing method in IoT-based information system for emergency medical services. IEEE Trans Ind Inf 10(2):1578–1586
49. Zhang H, Li J, Wen B et al (2018) Connecting intelligent things in smart hospitals using NB-IoT. IEEE Internet Things J 5:1550–1560
50. Lumley Thomas (2010) Potential for revealing individual-level information in genome-wide association studies. JAMA 303(7):659
51. Homer S, Szelinger M, Redman D, Duggan W, Tembe J, Muehling JV, Pearson DA, Stephan SF, Nelson DW (2008) Craig, resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4(8):e1000167
52. Chen R, Acs G, Castelluccia C (2012) Differentially private sequential data publication via variable-length *N*-grams. In: ACM conference on computer and communications security (CCS). ACM
53. Chen R, Peng Y, Choi B et al (2014) A private DNA motif finding algorithm. J Biomed Inform 50(Sp. Iss. SI):122–132
54. Kevin O, Seidman R (2016) Personal information security and exchange tool. Interaction processor and exchange tool
55. Wu X, Wang H, Wei D et al (2020) ANFIS with natural language processing and gray relational analysis based cloud computing framework for real time energy efficient resource allocation. Comput Commun 150:122–130
56. Wu X, Wang H, Tan W et al (2020) Dynamic allocation strategy of VM resources with fuzzy transfer learning method. Peer Netw Appl. https://doi.org/10.1007/s12083-020-00885-7