Privacy-Preserving Data Aggregation Scheme for E-Health

Matthew Watkins, Colby Dorsey, Daniel Rennier, Timothy Polley, Ahmed Sherif and Mohamed Elsersy

September 6, 2022

Privacy-Preserving Data Aggregation Scheme for E-Health

Matthew Watkins
University of Southern Mississippi
Hattiesburg, Mississippi
Matthew.Watkins@usm.edu

Colby Dorsey
University of Southern Mississippi
Hattiesburg, Mississippi
Colby.Dorsey@usm.edu

Daniel Rennier
University of Southern Mississippi
Hattiesburg, Mississippi
Daniel.Rennier@usm.edu

Timothy Polley
University of Southern Mississippi
Hattiesburg, Mississippi
Timothy.Polley@usm.edu

Ahmed Sherif
University of Southern Mississippi
Hattiesburg, Mississippi
Ahmed.sherif@usm.edu

Mohamed Elsersy
Computer Information Systems Department
Higher Colleges of Technology
Al Ain City, UAE
Melsersy@hct.ac.ae

***Abstract*—E-Health is the use of digital services and communication technology in support of healthcare. E-Health services are becoming increasingly popular. With E-Health, large amounts of data need to be collected, stored, and sent to other places all while remaining private. This arises the need for privacy-preserving data aggregation schemes to be implemented. Many other privacy-preserving data aggregation schemes already exist for E-Health services utilizing tools such as homomorphic encryption which can be slow with large amounts of data. This paper proposes a privacy-preserving scheme to aggregate data in an E-Health setting. Our scheme allows for all of the patients' individual data to remain private. Doctors can utilize partial decryption in our scheme to collect specific information about patients such as how many patients have high blood pressure without seeing all of the patients' data.**

***Index Terms*—aggregation, encryption, k-Nearest Neighbor**

I. Introduction

E-Health is a rapidly expanding healthcare service. E-Health involves the use of electronic systems and communication in healthcare that carries many benefits, one of them being easy to access for doctors to view and modify patient information. Within E-Health, there is a need to collect lots of information which calls for the use of data aggregation. Data aggregation is a technology that needs to be used in E-Health architectures to collect mass amounts of medical data for storage and processing. By aggregating data, mass amounts of data are collected and summarized to be easier to read and understand.

There are numerous different types of proposed data aggregation schemes for use in the burgeoning E-Health industry today. One major category of these schemes uses Homomorphic encryption, which is a form of encryption that permits users to perform computations on its encrypted data without first decrypting it. These resulting computations are left in an encrypted form which, when decrypted, results in an identical output to that produced had the operations been performed on the unencrypted data. Other schemes use Asymmetric encryption, also known as Public-key encryption, which uses pairs of keys. Each pair consists of a public key (which may be known to others) and a private key (which may not be known by anyone except the owner). In such a system, any person can encrypt a message using the intended receiver's public key, but that encrypted message can only be decrypted with the receiver's private key. Already proposed examples of such data aggregation schemes include HMAC, which uses a keyed hash for message integrity and authentication, AVISPA, which provides security and authentication with the local server and establishes session keys, and PCDA, which provides integrity of data through a cryptographic hash algorithm, DDPA, which uses hashing, cryptographic mechanisms, and both types of encryption keys, as well as the older RSA, which uses relies on the practical difficulty of factoring the product of two large prime numbers. K Nearest Neighbor is a non-parametric method used for classification. The principle is that known data are arranged in a space defined by the selected features. When new data is supplied to the algorithm, the algorithm will compare the classes of the k closest data to determine the class of the new data. For data aggregation, the aggregated distance to neighbouring observation would be by aggregating the results from the closest users to the cloud.

One privacy problem that arises with data aggregation is collecting information without knowing all of the details of all of the individual data. In E-Health architectures, privacy preservation and security are of extreme importance. This means that it is necessary to be able to complete data aggregation over encrypted data. This allows for more privacy preservation and security. By the use of a K-Nearest Neighbors (kNN) algorithm, aggregation can be done over encrypted data.

The k-Nearest Neighbors (kNN) algorithm is capable of performing aggregation on individual bits. This allows for more scalability within the scheme which means that the aggregation can be performed on data of different types and sizes. The benefit of using a K-Nearest Neighbors algorithm is that the aggregation can be done over the encrypted data and the contents of the data are not revealed. The users of the system can request the aggregated data and not all of the individual data is revealed allowing for more privacy and security in the forties.

A strong and secure privacy-preserving data aggregation scheme in E-Health needs to be secure to maintain the confidentiality of the user's medical data to preserve privacy. The scheme should also maintain the integrity of the user's medical data to ensure that the data has not been tampered with or altered in any way. This means that a strong encryption technique needs to be used to encrypt the data and protect it. The scheme proposed in this paper accomplishes these requirements.

The remainder of this paper is organized as follows: The related work is discussed in section II. The network model, attack model, and design goals are discussed in section III. The proposed scheme is discussed in section IV. The security and privacy analysis is discussed in section V. The conclusion are made in section VI.

II. Related Work

Many researchers have proposed data aggregation schemes to try and solve the problem of privacy-preserving data aggregation. With privacy-preserving data aggregation schemes, it is important to be able to aggregate the data without the aggregator knowing either the individual or whole aggregated data. Many schemes that already exist utilize cryptographic techniques such as homomorphic encryption.

In [1] and [11], privacy-preserving data aggregation schemes are proposed utilizing homomorphic encryption. Utilizing techniques such as homomorphic encryption can become slow when dealing with large amounts of data however it does allow for strong privacy preservation. The authors in [2] and [3] present schemes that utilize an encryption scheme based on the k-Nearest Neighbor (kNN) algorithm. The authors in [2] propose a scheme for aggregation of data sets and the authors in [3] proposes a scheme for autonomous cab management. Utilizing the kNN algorithm allows for individual bits to be aggregated instead of using multi-bit addition such as other schemes. CNN also allows for privacy preservation since the aggregator can aggregate the data without learning the data. The author's schemes in [2] and [3] also allow for partial decryption which in our scheme would allow doctors to only decrypt certain information about the patients. A modified version of these kNN schemes is utilized in our scheme dealing with E-Health.

[4] Due to the direct participation of patients' health data, Zhang makes it clear that privacy issues and security are very sensitive in HWSNs. Authenticity and confidentiality are the basic requirements of secure communication. On the one hand, only medical professionals and patients must be able to see the raw medical data, on the other hand, there must be an efficient data authentication mechanism to deal with the increasing number of patients. A public-key cryptosystem is designed to solve the above problems of confidentiality and authentication. Zhang touched bases on identity-based cryptosystem to simplify the certificate management problems existed in the certificate-based systems. [5] The wireless body sensor network (WBSN) technology is an application of IoT in healthcare, whereas data security and privacy impediments have raised some concerns. The main contribution of this research paper is a cryptographic accumulator based on the authenticated additive homomorphic encryption which can collect and accumulate data from IoT wireless wearable devices. These encrypted data can be used for analysis in an encrypted form so that the information is not revealed. To validate security and efficiency, Rezaeibagha et. al present security analysis and performance evaluations of our proposed scheme for IoT wireless body sensors

[6] This is a survey on frameworks for secure data aggregation in smart cities; all using Fog architecture. Multiple schemes are documented for many applications for not only e-health but for many schemes and smart grids with low computational overhead. Smart devices communicate autonomously to transmit sensitive data securely across network nodes. [7] Secure remote data management is becoming standard modern day. Cloud-let technology aims to ease the computational overhead for networks offering this service. The framework utilizes SQL-type data models.

[8] Proposes a privacy-preserving health data aggregation scheme that securely collects health data from multiple sources and guarantees fair incentives for contributing patients. Combines a Boneh-Goh-Nissim cryptosystem and Shamir's secret sharing to keep data obliviousness security and fault tolerance. [9] Proposes a medical diagnosis system using e-health cloud servers in a privacy-preserving manner when medical datasets are owned by multiple data owners. The proposed system is the first one that achieves the privacy of the medical dataset, symptoms, and diagnosis results and hides the data access pattern even from e-health cloud servers performing computations using the data while it is still robust against collusion of the entities. [10] Proposes an efficient and privacy-preserving medical primary diagnosis scheme based on k-nearest-neighbours classification (kNN), called EPDK. With EPDK, medical users can ensure that their sensitive health information is not compromised during the online medical diagnosis process, and service providers can provide high-accuracy service without revealing their diagnosis model.

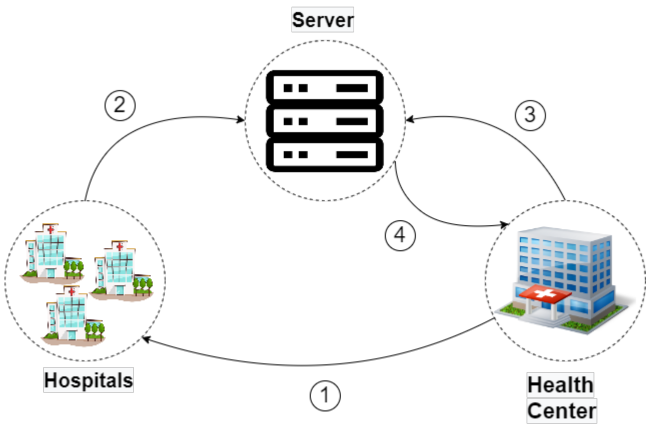
III. System Models and Design Goals

This section will discuss the proposed scheme's network model, attack model, and design goals.

A. Network Model

The network model is shown in figure 1 consisting of three entities the patients, the server, and the health center.

- **Hospitals:** The hospitals collect the data from the patients, encrypt it, and then send it to the aggregation server.
- **Server:** The aggregation server stores the patients' data and is able to perform an aggregation over encrypted data. The aggregation server sends the aggregation results to the health center.
- **Health Center:** The health center is a federal research center. It is responsible for distributing keys. The health



1 The health center distributes encryption keys
2 The hospital uploads patients' encrypted data
3 The health center sends aggregation query
4 The server performs aggregation and sends result back

Fig. 1. Network Model

center sends queries to the aggregation server to receive aggregated data.

B. Attack Model

There could be many different possible attackers involved in a data aggregation scheme including the users involved in the scheme, known as internal hackers, as well as external attackers. Attacks on the scheme could either be passive or active. In a passive attack, the attacker only wishes to collect the data that is being aggregated. In an active attack, the attacker could collect the data and modifies it to produce false information or inserts new data. In our scheme, the aggregation server is a third party that is not related to the health center but is known to all users. With this, a possible attack method would be honest but curious. Honest but curious is a legitimate participant, such as the aggregation server, in a communication protocol who will not deviate from the defined protocol but will attempt to learn all possible information from the network. For example, in our scheme, user A will follow the scheme's design honestly but user A will also try to read user B's information.

C. Design Goals

The following five design goals are to be achieved by this privacy-preserving data aggregation scheme.

- **Efficiency:** The first design goal is efficiency, with the idea of reducing the energy and computational costs of accessing and manipulating data while at the same time providing quick and timely access to the authorized personnel. The encryption, aggregation, and searching should all be performed efficiently and promptly.
- **Security:** The second design goal is security, being able to provide a system that is safe from both active and passive attacks from internal or external hackers and able to filter false data locally at the aggregator. Keeping data secure and providing a high degree of confidentiality is essential to the sensitive nature of the information being stored and accessed.
- **Reliability:** The third goal is reliability, providing a security mechanism that guarantees the availability of data when needed and ensures a sense of confidence that patients' data is both safe and available even when certain nodes may be defective.
- **Scalability:** The fourth design goal is to provide scalability to allow for encryption and aggregation of data being different types and sizes.
- **Privacy Preservation:** The fifth design goal is privacy preservation. The patients' individual information should remain private to all users with only the specific aggregated information being accessible to the health center upon its request.

When dealing with the sensitive nature of E-health data, any system of data aggregation used must provide all three of these design goals with the absolute minimum chance of failure. Keeping computational costs low, while providing both patients and doctors with ease of mind that their health data will not be compromised or lost, is of the utmost importance to our proposed data aggregation scheme.

IV. Proposed Scheme

The proposed scheme consists of four phases. In this section, each phase is discussed in section B through section E with details.

A. Overview

A privacy-preserving data aggregation scheme is important in E-Health for many reasons. A large amount of encrypted data for each patient is collected but certain doctors may not want or need to receive all of the data. This is where aggregation becomes helpful. The doctors can receive only the desired information about patients by receiving the aggregated data from the server that it is stored on.

In this scheme, the key distribution is handled by the health center. The health center distributes keys that will be used to encrypt the data. The encrypted data about each patient is sent to the server, which can perform aggregation. The aggregation is performed at the request of the health center, and the aggregated data is sent to the health center. The health center can then decrypt the information.

In the example given below, there are two binary vectors containing some data about each patient P_1 and P_2. Each cell of the vector contains specific information about the patient such as their blood pressure (BP) or blood sugar (BS). A value of 1 would indicate that the patient has high blood pressure or blood sugar and a value of 0 would indicate that the patient has low/normal blood pressure or blood sugar. The binary vector could also contain information about health problems such as if the patient has diabetes (D), cancer (C), lung disease (LD), etc.

	BP	BS	D	C	LD	
P_1	1	0	0	1	0	...
P_2	0	1	0	1	0	...

B. Key Distribution

The first phase of the scheme consists of the key distribution by the health center to the hospitals. The primary user, which is the health center, will have a key as follows: [SI, X_1Y_1, X_1Y_2, X_2Y_3, X_2Y_4] with SI being a binary vector of size n. X and Y are random invertible matrices with a size of n x n. Each secondary user, being the hospitals, will have keys as follows: [SI, $X_i^{'}Y_1^{-1}$, $X_i^{''}Y_2^{-1}$, $X_i^{'''}Y_3^{-1}$, $X_i^{''''}Y_4^{-1}$]. For the secondary users' keys, $(X_i^{'} + X_i^{''})$ is equal to X_1^{-1} and $(X_i^{'''} + X_i^{''''})$ is equal to X_2^{-1}. The binary vector S is shared between the primary and secondary users and each secondary user has a different $(X_i^{'}Y_1^{-1}$,$X_i^{''}Y_2^{-1})$ and $(X_i^{'''}Y_3^{-1}$,$X_i^{''''}Y_4^{-1})$. This is due to the matrices X and Y being random.

C. Data Encryption and Submission

The second phase of the scheme is the encryption and submission of the patients' data. The hospital is responsible for submitting the encrypted patients' data to the aggregation server.

For the encryption in this scheme, a binary vector SI will be used as a splitting indicator. The splitting indicator SI is used to split the data vectors v_i into two random vectors $v_i^{'}$ and $v_i^{''}$. If the j^{th} bit of SI is one, $v_i^{'}(j)$ and $v_i^{''}(j)$ are set similar to $v_i(j)$. If the splitting indicator SI is zero, $v_i^{'}(j)$ and $v_i^{''}(j)$ are set to two random numbers that add up to equal $v_i(j)$. The data vector pair of the secondary user, $(v_i^{'}, v_i^{''})$ is encrypted to

$$C_i = [{X_i}^{'}{Y_1}^{-1}{v_i}^{'}, X_i^{''}{Y_2}^{-1}{v_i}^{'}, {X_i}^{'''}{Y_3}^{-1}{v_i}^{''}, {X_i}^{''''}{Y_4}^{-1}{v_i}^{''}].$$

The ciphertext C_i is the index and $v_i^{'}$ and $v_i^{''}$ are both column vectors containing data.

D. Aggregation on Patient Data by the Server

The third phase of the scheme consists of the aggregation of the patients' data. The data aggregation is performed by the aggregation server.

For the aggregation, C_i is a column vector containing data about the patient. The number of elements inside C_i is $4 * n$. There are m amount of users U_i who each create an index C_i as shown below. $k_{i,j}$ is the j^{th} element in the i^{th} index.

$$\begin{aligned} C_1 &= [k_{1,1}, k_{1,2}, \ldots, k_{1,4n}]^T \\ C_2 &= [k_{2,1}, k_{2,2}, \ldots, k_{2,4n}]^T \\ &\vdots \\ C_m &= [k_{m,1}, k_{m,2}, \ldots, k_{m,4n}]^T \end{aligned}$$

C_{agg} is the aggregated index that can be computed by the summation of all m users' indices as shown below. This could give the health center information such as the number of patients with high blood pressure.

$$\begin{aligned} C_{agg} &= \sum_{i=1}^{m} C_i \\ &= \left(\sum_{i=1}^{m} k_{i,1}, \sum_{i=1}^{m} k_{i,2}, \sum_{i=1}^{m} k_{i,3}, \ldots, \sum_{i=1}^{m} k_{i,4n} \right) \\ &= [k_1, k_2, \ldots, k_{4n}] \end{aligned}$$

E. Sending Aggregation Request and Data Decryption

	BP	BS	D	C	LD	
P_1	1	0	0	0	0	...

The fourth phase of the scheme consists of sending the requested data and decryption the data. The decryption will be performed by the doctors or nurses who wish the receive the aggregated patient information data. The decryptor uses its key to decrypt and retrieve back the desired aggregated data. In this situation, only partial decryption will be used. To enable partial decryption, an aggregator and the user needs to decrypt the k^{th} element in the aggregated index. Then, the decryptor should use its keys to compute the secret key and send it to the aggregator. When computing the secret key, the decryptor should first create a binary vector, d, where the k^{th} bit in d is set to 1 while all other bits are set to 0. C_{agg} is the aggregated index that can be computed by the summation of all m users as previously stated. "E" will serve as the encryption for data and "d" will serve as a binary vector created by the decryptor. To decrypt the data and view the patients' blood pressure, the dot product between C_{agg} and $E(d)$ would need to be conducted. Doing so would allow for partial decryption to take place. In this situation, one would be placed in the blood pressure category and a zero in all other categories, so only the blood pressure information would be decrypted and accessible.

V. Privacy and Security Analysis

Efficiency. The use of a scheme based on the kNN algorithm ensures efficiency. In our scheme, the time required to encrypt and aggregate data remains low, even with an increased number of data, which allows for these operations to be performed promptly.

Scalability. This scheme is highly scalable and allows for data of different types and sizes to be efficiently encrypted and aggregated.

Patients' data privacy. One of the goals of our scheme is to maintain privacy preservation. Our scheme preserves the privacy of patients' data that is being transmitted. The aggregation server is not able to find out any information about the patients' data because of the encryption used. The patients' data is encrypted using their keys, and it is infeasible to decrypt the data without knowing the keys. This ensures that the patients' data remains private to those involved in the scheme.

The patients' data indices can not be decrypted by the aggregation server. By utilizing a scheme based on kNN encryption, the patients' data indices cannot be decrypted by the aggregation server as long as the keys are unknown to it. Whenever the health center requests the aggregated data, the aggregation server cannot know what is contained within the patients' indices.

The keys remain private between the users. The secret keys used for encryption in this scheme involve random matrices that are unknown between the different users. Since these matrices are unknown, the necessary computations to compute the keys cannot be performed by any attackers. Due to this, the patients' data cannot be decrypted and remains secure and private.

VI. Performance Evaluation

Our scheme was implemented using MATLAB. For the patients' data in the code, random binary vectors of 50 elements were used in different amounts. The execution time of the encryption, aggregation, and searching in the code was tested five times with the four different amounts of binary vectors and then the results were averaged and plotted. The code was tested using 1 vector, 50 vectors, 100 vectors, 150, and 200 vectors simulating different patients' data. The performance evaluation of the MATLAB code was conducted on a PC with an Intel Core i7-10700F processor @ 2.90 GHz and 16.00 GB of RAM.

Size of Data Before Encryption	400 bytes
Size of Data After Encryption	1600 bytes

Fig. 2. Patient Data Vector Sizes

Figure 2 shows the size of the patients' data vectors before and after encryption. Before encrypting the patients' data vectors, the size is 400 bytes. This is due to the patients' data vectors being 50 elements and of the type double. In MATLAB, a double is stored using 8 bytes. So, $50 * 8 = 400$ bytes. After performing encryption, the size is 200 elements which take up 1600 bytes of memory. Performing encryption makes the data size larger in this scheme.

Figure 3 shows the encryption time with different amounts of patient data vectors: 1, 50, 100, 150, and 200. With an increased amount of data, the number of vectors containing patient information, and the encryption time also increase proportionally. Although the time increases, it remains low to perform encryption with it being less than 35 milliseconds.

Figure 4 shows the aggregation time with different amounts of patient data vectors: 1, 50, 100, 150, and 200. The aggregation time is the time it takes for the aggregator to sum the encrypted vectors coming from different hospitals. The time that it takes for aggregation increases proportionally to the number of data vectors that are being aggregated.

Figure 5 shows the time required for the partial decryption operation to be performed with different amounts of patient data vectors: 1, 50, 100, 150, and 200. This figure is related to the dot product operation that is performed for partial decryption where a value of 1 is placed in a certain location with 0's placed in all other locations to get the number of patients with a specific disease.

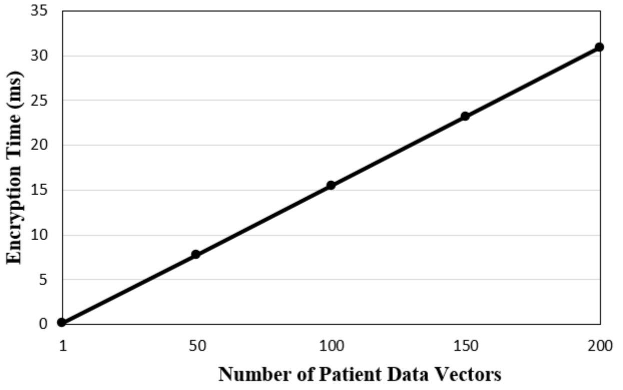


Fig. 3. Encryption Time

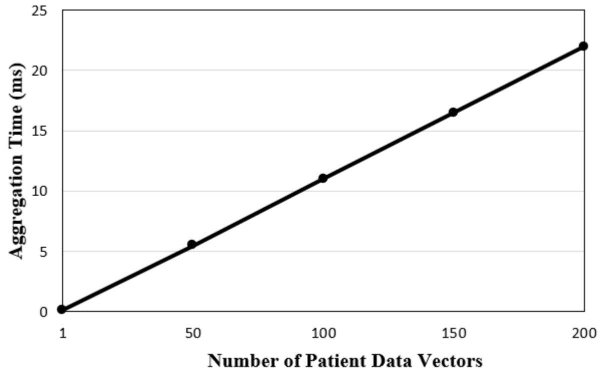


Fig. 4. Aggregation Time

VII. Conclusion

In this paper, we proposed a privacy-preserving data aggregation scheme for E-Health. Our scheme utilizes techniques based on the kNN encryption and aggregation schemes. By utilizing kNN techniques, the patients' data is ensured to remain private and can be aggregated efficiently. The techniques utilized in our scheme allow for data of different sizes to be aggregated, while also maintaining our original design goals of efficiency, security, and reliability. With the E-Health industry expected to grow exponentially in the coming years, schemes such as the one we are proposing are of the utmost importance to the continued growth of the field. Patients and doctors expect nothing less than the complete safety and reliability of their sensitive health information, and our kNN-based scheme is the one we feel will be the best fit for them and the entire industry going forward.

References

[1] F. A. Almalki and B. O. Soufiene, "EPPDA: An Efficient and Privacy-Preserving Data Aggregation Scheme with Authentication and Authorization for IoT-Based Healthcare Applications". *Wireless Communica-*

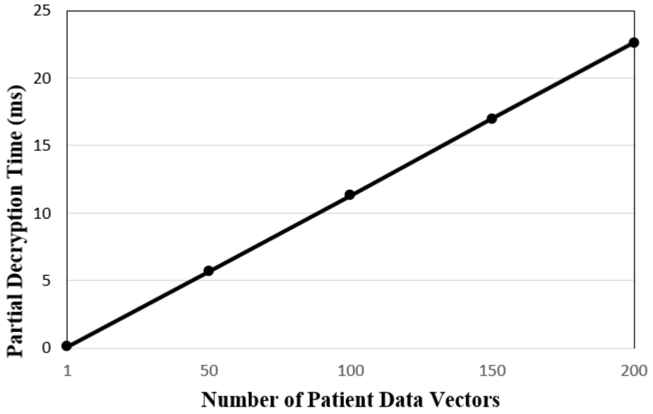


Fig. 5. Partial Decryption Time

tions and Mobile Computing, vol. 2021, Article ID 5594159, 18 pages, March 2021.

[2] A. Sherif, A. Alsharif, M. Mahmoud, M. Abdallah, and M. Song, ”Efficient Privacy-Preserving Aggregation Scheme for Data Sets”. *25th International Conference on Telecommunications (ICT)*, June 2018.

[3] A. Sherif, A. Alsharif, M. Mahmoud, and J. Moran, ”Privacy-Preserving Autonomous Cab Service Management Scheme”. *AMECSE ’17: Proceedings of the 3rd Africa and Middle East Conference on Software Engineering*, Decemeber 2017.

[4] B. Zhang, ”A Lightweight Data Aggregation Protocol With Privacy-Preserving for Healthcare Wireless Sensor Networks”. *IEEE Systems Journal* vol. 15(2), pp. 1705-1716, June 2021.

[5] F. Rezaeibagha, Y. Mu, K. Huang, L. Chen, ”Secure and Efficient Data Aggregation for IoT Monitoring Systems”. *IEEE Internet of Things Journal*, May, 2021, vol. 8(10), pp. 8056-8063.

[6] A. Ullah, M. Azeem, H. Ashraf, A. A. Alaboudi, M. Humayun, and N. Z. Jhanjhi, “Secure Healthcare data aggregation and transmission in IOT—A survey,” *IEEE Access*, vol. 9, pp. 16849–16865, 2021.

[7] S. Sengupta and S. S. Bhunia, “Secure data management in Cloudlet assisted IOT enabled e-health framework in Smart City,” *IEEE Sensors Journal*, vol. 20, no. 16, pp. 9581–9588, 2020.

[8] W. Tang, J. Ren, K. Deng and Y. Zhang, ”Secure Data Aggregation of Lightweight E-Healthcare IoT Devices With Fair Incentives,” in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 8714-8726, Oct. 2019, doi: 10.1109/JIOT.2019.2923261.

[9] Jeongsu Park, Dong Hoon Lee, ”Privacy Preserving k-Nearest Neighbor for Medical Diagnosis in e-Health Cloud”, Journal of Healthcare Engineering, vol. 2018, Article ID 4073103, 11 pages, 2018. https://doi.org/10.1155/2018/4073103

[10] D. Zhu, H. Zhu, X. Liu, H. Li, F. Wang and H. Li, ”Achieve Efficient and Privacy-Preserving Medical Primary Diagnosis Based on kNN,” 2018 27th International Conference on Computer Communication and Networks (ICCCN), 2018, pp. 1-9, doi: 10.1109/ICCCN.2018.8487422.

[11] A. Ara, M. Al-Rodhaan, Y. Tian, and A. Al-Dhelaan, ”A Secure Privacy-Preserving Data Aggregation Scheme Based on Bilinear ElGamal Cryptosystem for Remote Health Montitoring Systems”. *IEEE Access*, vol. 5, pp. 12601-12617, June 2017,