Hindawi
International Journal of Antennas and Propagation
Volume 2021, Article ID 9317018, 7 pages
https://doi.org/10.1155/2021/9317018

Research Article

Construction and Application of Media Corpus Based on Big Data

Bochun Yin[1] **and Lei Fu**[2]

[1]*School of Foreign Languages, Changsha University, Changsha 410205, China*
[2]*The First Clinical College, Changsha Medical University, Changsha 410205, China*

Correspondence should be addressed to Bochun Yin; z20200811@ccsu.edu.cn

Received 17 August 2021; Revised 17 September 2021; Accepted 28 September 2021; Published 21 October 2021

Academic Editor: Fangqing Wen

Copyright © 2021 Bochun Yin and Lei Fu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the problems of poor data quality and low application rate in the construction of existing media corpus, this paper proposes the construction and application research of media corpus based on big data. Media corpus data are collected, the data are divided into four categories, the heuristic data item column sorting algorithm is introduced to sort all collection processes, the minimum value of data item collection rate is determined, on this basis, the maximum value of quantity in media corpus is determined, and data collection is realized in media corpus data through sliding window. Then, the state characteristics and probability distribution of feature data are determined by dynamic Bayesian network, the relationship between the state variables and dimensions of media corpus data is determined, and the media corpus data state is processed by component to complete the preprocessing of media corpus data; finally, through the application research of storage and encryption of the designed database through big data technology, the storage structure data and encryption secret key are designed to realize the construction and application of media corpus. The experimental results show that the data quality of the media corpus constructed by the proposed method is high, and its application rate has been improved to a certain extent.

1. Introduction

With the rapid development of Internet technology, database is becoming more and more important in people's mind. As the core of future technology development, database has attracted much attention. Different software companies have developed different database software, and the data models of the same software are also different. Because the data models are not unified, it is difficult to transmit data between database software. Data conversion between heterogeneous databases is an important means needed at present. The conversion between heterogeneous data can effectively improve work efficiency and reduce cost [1]. The timeliness of language database is an important attribute of data. In data mining, data analysis, and data value-added applications, accurate data timeliness determines the effect of a series of algorithms such as time series analysis, association, and recommendation. Relevant scholars have studied the problem of data quality by means of direct observation, social investigation, and theoretical derivation. The properties that have a great impact on data availability are accuracy, integrity, consistency, timeliness, and entity identity. In the field of business data, due to the change of customer information, 2% of business data are outdated every month. A large number of imprecise data are filled in the data set. If you cannot identify which is the latest, data query may return wrong results, and data analysis may produce contrary conclusions, followed by the decline of data quality and data value. In the era of big data, people's various data are distributed in various platforms and systems, forming data islands. The problems caused by inaccurate data timeliness and outdated data are becoming more and more serious. In the era of big data and artificial intelligence, personal big data contain immeasurable social and economic value. The personal data banking model is a new model that can effectively sort and integrate personal data, improve the quality and value of personal data, enhance the controllability and availability of personal data, and effectively protect personal data privacy [2]. In the process of data aggregation, due to the high dispersion of personal data at the same time, personal data are typical dynamic data, and various data reflecting

personal attributes and status are constantly changing. This feature is also the biggest challenge in the process of personal data cleaning. In the data mode, in order to ensure data quality and improve data value, it is necessary to gather data from many sources, and its time attribute is often inaccurate [3]. For some attributes of data, different times correspond to different values or states, such as a person's educational background change and marital status change. If the timestamp is incomplete or inaccurate, it is impossible to determine the sequence of records, which will bring great difficulties to data value-added applications. Therefore, how to improve the construction effect of media corpus and widely apply it has become the focus of current research [4]. Therefore, relevant researchers have conducted research on the construction of media corpus and obtained some results.

Literature [5] designed a database selection based on the hesitant language information aggregation algorithm and studied the construction and application of media corpus. In order to improve the efficiency of database selection, a database selection method based on the hesitation language multiattribute group decision-making algorithm is proposed for multiattribute group decision-making. Firstly, a database selection model based on generalized hesitation language Heronian average (GHLHM) operator is constructed; secondly, Archimedean norm is introduced into hesitation language environment, and a new hesitation language algorithm is defined; thirdly, based on the new algorithm and Heronian average, GHLHM operator is proposed, some basic properties of GHLHM operator are discussed, several common operator forms of GHLHM operator are studied, and generalized hesitant language weighted Heronian average (GHLWHM) operator is proposed; and finally, a new hesitant language multiattribute group decision-making method based on GHLWHM operator is constructed and applied to database selection. Experiments show that this method can realize the comprehensive optimization and ranking of database performance and has a wide application prospect in other fields. However, this method still needs some improvement in order to consider too much data security and storage occupancy. Literature [6] designed a method for automatic conversion of heterogeneous data in language database. This method provides an important method and means to operate the database for the database management system. In view of the problems of long data conversion time, low information utilization, and conversion accuracy after data conversion in the traditional data conversion method, a new heterogeneous data conversion method is proposed, which is based on the automatic heterogeneous data conversion method in the language database. The language data with the same collection cycle are placed in the corresponding data column, the heuristic algorithm is introduced to sort the data column, and each data item is adjusted according to the arrangement order to complete the automatic heterogeneous data collection in the language database. The experimental results show that the proposed method can effectively reduce the time-consuming of data conversion and improve the utilization and accuracy of data conversion. However, the amount of data and safety performance in the construction of this method has not been effectively controlled, and there are some deficiencies. Through the investigation of 33 Arabic language corpora, reference [7] found that although Arabic language corpus has made great progress, the Saudi dialect corpus still needs to be further expanded. This paper makes a contribution to the literature of SD corpus by creating the Saudi Corpus (KSUSC) of King Saud University, the largest corpus in Saudi Arabia. The total number of words in this corpus is +1 B, in which SD words are +119 M. KSUSC is not only the latest and largest SD corpus in China but also a corpus with rich and diverse contents, covering 26 fields from 5 different sources. Reference [7] proposed a method to semiautomatically construct a corpus that includes Japanese youth slang called Wakamono Kotoba. The process of semiautomatic corpus construction is composed of the first step which is automatic collection of example sentence, the second step is tag annotation to collected sentences, and the final step is manually modifying tag and noise reduction.

Therefore, this paper proposes a media corpus construction and application research method based on big data. Firstly, the media corpus data are collected, the data are divided into four categories, the heuristic data item column sorting algorithm is introduced to sort the whole collection process, and the minimum value of the collection rate of the data item is determined. On this basis, the maximum value of the number in the media corpus is determined, and the data collection in the media corpus data is realized through the interaction window; then, the state characteristics and probability distribution of feature data are determined by dynamic Bayesian network, the relationship between the state variables and dimensions of media corpus data is determined, and the media corpus data state is processed by component to complete the preprocessing of media corpus data; finally, the designed database is effectively applied through big data technology. The technical route of this paper is as follows:

(i) Step 1: collect media corpus data, divide the data into four categories, introduce the heuristic data item column sorting algorithm to sort all collection processes, determine the minimum value of data item collection rate, on this basis, determine the maximum value of quantity in media corpus, and realize data collection in media corpus data through interactive window.

(ii) Step 2: determine the state characteristics and probability distribution of feature data through dynamic Bayesian network, determine the relationship between media corpus data state variables and dimensions, process the media corpus data state by components, and complete the preprocessing of media corpus data.

(iii) Step 3: effectively apply the designed database through big data technology.

(iv) Step 4: analysis of the experimental process.

(v) Step 5: experimental conclusions and future prospects.

Our contribution includes the following three points:

(1) Aiming at the problems of poor data quality and low application rate in the construction of existing media corpus, this paper proposes the construction and application research of media corpus based on big data.

(2) Media corpus data are collected, the data are divided into four categories, the heuristic data item column sorting algorithm is introduced to sort all collection processes, the minimum value of data item collection rate is determined, and on this basis, the maximum value of quantity in media corpus is determined.

(3) The state characteristics and probability distribution of feature data are determined by dynamic Bayesian network, the relationship between the state variables and dimensions of media corpus data is determined, and the media corpus data state is processed by component to complete the preprocessing of media corpus data.

2. Research on the Media Corpus Construction

2.1. Media Corpus Data Collection. Data collection of media corpus is the most critical mitigation in the construction process. Only complete and massive data support can make the construction of media language database more complete. Only by fully ensuring the real-time and effectiveness of the collected data can we fully understand the operation status of the whole media corpus and better apply it [8].

Data items in the media corpus are composed of different separate cells, each with their own attribute characteristics. Therefore, in the process of building the media corpus, setting up the collected data consists of four categories. Express its attributes based on different language data as $S/T/T_d/D$. Among them, S represents the data in the media corpus, T represents the sampling cycle, T_d represents the heterogeneous data acquisition time delay, and D represents the collection of data items in the media corpus. Heuristic data items are introduced to the column sorting algorithm to sort all the collection processes, and then the common relationship of different data items is sorted. In the process of sorting, the collection rate of "reused" data items is the lowest, that is:

$$A = \left(a_{ij}\right)_{n\times m}. \tag{1}$$

After obtaining the most effective frequency of data acquisition in the media corpus data, it is necessary to determine the number of data items required by the media corpus [9], and the maximum number of media corpus obtained is as follows:

$$R_{\max} = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} a_{ij} - 1\right). \tag{2}$$

After the quantity of data in the media corpus data is determined, reflect the data relationship existing in the database and reflect the data relationship through the relationship matrix to obtain the following:

$$B = \left(x_{ij}\right)_{n\times m}. \tag{3}$$

According to the determined data relationship matrix in the media corpus data, the data collection in the media corpus data is completed through the sliding window, that is:

$$R = \sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij}\delta_{ij} - 1. \tag{4}$$

In the process of media corpus construction, first collect the data of media corpus, divide the data into four categories, introduce the heuristic data item column sorting algorithm to sort all the collection processes, determine the minimum value of data item collection rate, on this basis, determine the maximum value of quantity in media corpus, and realize data collection in media corpus data through interactive window.

2.2. Media Corpus Data Preprocessing. According to the media corpus data obtained above, there are many data in the data collected in the media corpus, and there are great differences between the data, which affects the construction of the media corpus. Therefore, it is necessary to preprocess the media corpus data. In the obtained media corpus data, the characteristics of media corpus data cannot become the key data due to the rapid change of media corpus data [10]. Therefore, this paper preprocesses the differences in media corpus data. In this paper, dynamic Bayesian network is used to preprocess feature data.

Assume that the p-dimensional implied state variable in the media corpus data can be expressed as follows:

$$P = \{p_1, p_2, \ldots, p_{n-1}\}. \tag{5}$$

Here, the probability distribution is expressed as follows:

$$Q = \{q_1, q_2, \ldots, q_n\}. \tag{6}$$

The relationship between the media corpus data state variables and the dimension can be expressed as follows:

$$G(P,Q) = \varepsilon\prod_{i=1}^{P-1} P\left(p_n \mid q_n\right)Q\left(p_n \mid q_n\right). \tag{7}$$

In formula, $P(p_n \mid q_n)$ represents the state transition probability distribution, $Q(p_n \mid q_n)$ represents the actual information volume of the media corpus data, and ε represents the initial state of the media corpus data.

Based on this basis, the media corpus data state is processed, and the preprocessing of the data characteristics of the media corpus is completed.

$$D_i(u) = \frac{1}{N}\sum_{i=1}^{N} Q(t). \tag{8}$$

In formula, $D_i(u)$ represents the acquired media corpus data.

The media corpus data preprocessing process is shown in Figure 1.

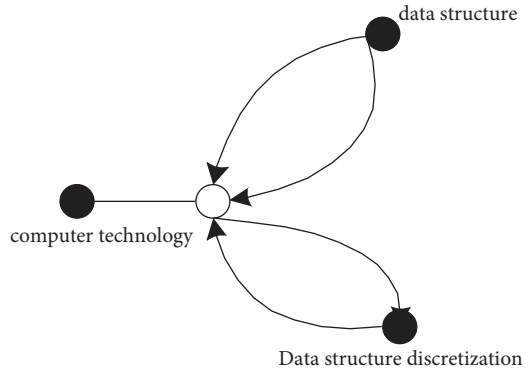

Figure 1: Data preprocessing process of the media corpus.

In the media corpus data preprocessing, the characteristic data state characteristics and probability distribution are determined through the dynamic Bayesian network, the relationship between the media corpus data state variables and the dimensions is determined, the media corpus data state is carried out for component processing, and the preprocessing of the media corpus data is completed [11].

3. Application Study of the Media Corpus Based on Big Data

3.1. Media Corpus Storage Applications. Based on the construction of the above media corpus data, the storage and application of the database are studied. By constructing the data information flow model of media corpus, the data storage of media corpus is studied by using nonlinear time series analysis [12]. In the media corpus data storage, the time model of media corpus data storage and scheduling information flow is as follows:

$$s(t) = \sum_{i=1}^{n} a_{\min} g_{\min}. \tag{9}$$

In formula, $s(t)$ represents the media corpus data storage space envelope amplitude.

In order to improve the efficiency of media corpus data storage application and meet the needs of media corpus data structure, the fitness function of constructing media corpus data storage point with the multiple regressive regression model is as follows:

$$f_{ij} = w_i \delta_t + w_c + \delta_t + w_q \delta_q. \tag{10}$$

In formula, f_{ij} represents the scheduling time for storage and access, δ_t represents the time cost of the storage, w_i represents the quality of storage of media corpus data, and w_q represents the database security performance.

Based on this basis, using the phase space reconstruction model, reorganize the storage subsets of the media corpus data for spatial characteristics and determine the probability distribution [13] of the fitness function satisfied by the media corpus data storage node, that is:

$$x_p(u) = \sqrt{\frac{1-\mu}{2\pi}} dt\ . \tag{11}$$

In formula, $x_p(u)$ represents the number of stages of the media corpus data storage structure. Data information flow characteristics in the media corpus data are classified as noninteger times of a. According to this calculation, the information flow time series reconstruction [11] in the database storage system is reconstructed by the nonlinear time series system analysis method, and the reconstruction expression is as follows:

$$F^p = F^4[f(t)]v. \tag{12}$$

According to the above reconstruction data flow time series, complete the media corpus data storage application as follows

$$F^p = x_p(u)(u - v \sin a). \tag{13}$$

In formula, F^p represents the spectral analysis features of the data store, v represents the distributed structure scaling properties, and equation (13) represents the orthogonal odd function of fractional transformation.

3.2. Media Corpus Data Encryption Application. In the application of media corpus data, its encryption is also a key link. Therefore, in the application of this paper, the encryption of media corpus data is studied. The secure encryption design of media corpus information is carried out under the advanced encryption standard protocol. The random linear coding scheme is adopted to construct the encryption key model of media corpus data storage information. The ciphertext construction and key design of database storage information are carried out under the logistics chaotic mapping, and the optimal encryption of database storage information is realized under the piecewise linear combination model [15].

The encryption and decryption key of the media corpus data storage information is set to H, the random phase reorganization method is used to encode and design the key of media corpus data information, the ciphertext protocol is constructed in logistics chaotic mapping, the block encryption method is used to encode the symbols of media corpus data, and the entanglement state model of media corpus data is established.

$$\beta \in H_1 \mid \beta \in H_1, \ldots, \mid \in H_n. \tag{14}$$

The cumulative probability distribution interval of the media corpus data encryption within the statistical interval I_i meets the following conditions:

$$\begin{aligned} \sum_{i=1}^{n} Q_i &= 1, \\ I_i\left[\sum_{i=1}^{n} Q_i = 1\right] &= \sum_{i=1}^{n} Q_j. \end{aligned} \tag{15}$$

The security key for media corpus data encryption is built, adaptive feature classification and vector quantification coding design with arithmetic coding design scheme are conducted, combined with the segmented Logistics encryption public key [16] of media corpus data, and the media corpus data encryption is completed. According to the Hash ciphertext distribution, the cryptographic coding protocol is used to design the media corpus data in the finite domain, improve the stability of the encryption process, and realize the research of encryption application.

4. Experimental Analysis

4.1. Experimental Design. In order to verify the effectiveness of the proposed method in the construction and application of media corpus data, simulation experiments are carried out. Assuming that the block length of media corpus data is 100, the sampling length of time series samples of media corpus data is 1200, the bandwidth of statistical feature sequence distribution set of media corpus is 14 dB, the number of layers of media corpus data encryption is set, and 30, 50, 60, 70, and 75 are used as the segmented sample set size of information coding stored in media corpus. The intensity of media corpus data attack is 20 dB. The waveform shape of sample media corpus data is shown in Figure 2.

4.2. Experimental Index Design. Based on the experimental scheme designed above, the indexes of this experiment are set as the space occupation of media corpus data storage and the security of media corpus data encryption. In order to promote the effectiveness of the experiment, the experiment is carried out in the form of comparison. The methods in this paper, literature [5], and literature [6] are compared, respectively. Many iterations are carried out in the comparison process to improve the accuracy of the experiment.

4.3. Analysis of Experimental Results. In order to verify the effectiveness of the design method in this paper, the experiment compares the method in this paper, the method in literature [5], and the method in literature [6] and compares the space occupation of sample media corpus data storage. The results are shown in Figure 3.

By analyzing the data in Figure 3, it can be seen that there are some differences in the space occupation of sample media corpus data stored by this method, literature [5] method, and literature [6] method under the same data

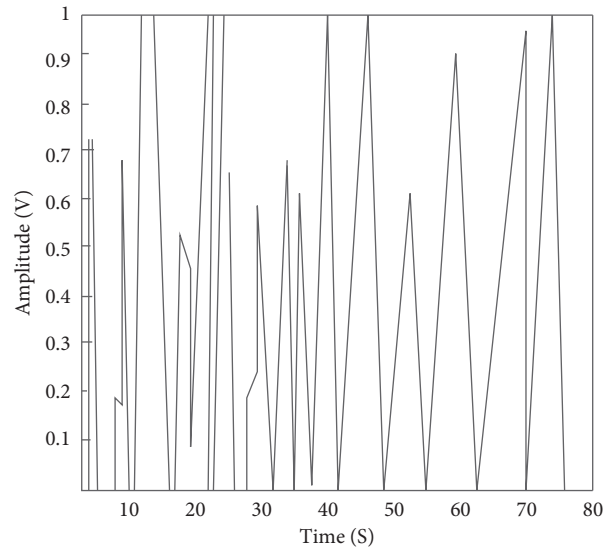

Figure 2: Waveform form of the sample media corpus data.

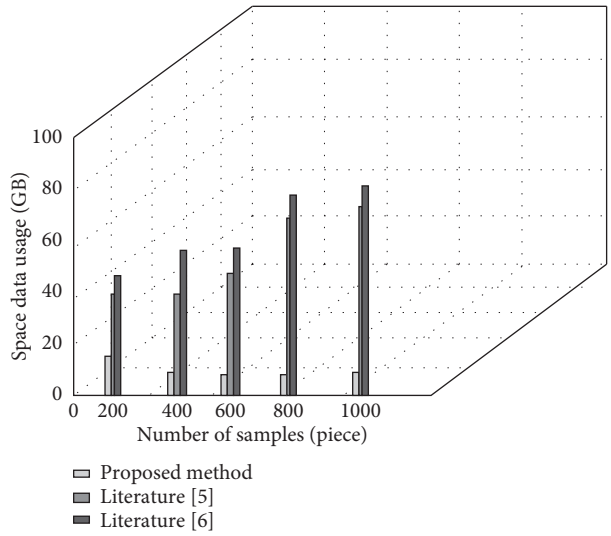

Figure 3: Comparison of space occupation of media corpus data storage by different methods.

volume. Among them, the space occupation of sample media corpus data storage in this method is always less than 20 GB, while the space occupation of sample media corpus data storage in the other two methods is always higher than that in this method. This is because this method has processed the data in the construction of data to verify the effectiveness of this method.

In order to further verify the effectiveness of the proposed method, the experiment compares the method in this paper, the method in literature [5], and the method in literature [6] to analyze the security of data encryption of sample media corpus. Taking the encrypted security coefficient as the measurement standard, the value range of the security coefficient is [0, 1]. The closer the security coefficient is to 1, the better the security is. The security results after encryption by the three methods are shown in Figure 4.

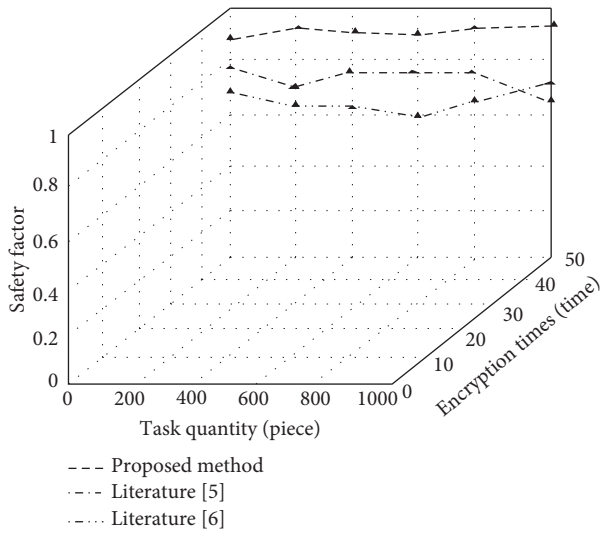

Figure 4: Comparison of data security coefficients of media corpora with different methods.

By analyzing the curve trend in Figure 4, it can be seen that under the same experimental conditions, the security of data encryption of sample media corpus is analyzed by using the methods of this paper, literature [5], and literature [6], and it can be seen that there are some differences in the security coefficients of the three methods. Among them, the security factor of sample media corpus data encryption by this method is higher than 0.9, while the security factor of sample media corpus data encryption by the other two methods is lower than that by this method. This is because this method uses big data technology to design the encryption key on the basis of the media corpus, which improves the security of the media corpus data and has a certain reliability. As can be seen from Figure 4, our method is more stable and fulfils our expected assumptions. In addition, the results in the figure also reflect the effective improvement of big data storage and encryption technology to our model.

5. Conclusion

This paper proposes the construction and application of media corpus based on big data. Media corpus data are collected, the data are divided into four categories, the heuristic data item column sorting algorithm is introduced to sort all collection processes, the minimum value of data item collection rate is determined, on this basis, the maximum value of quantity in media corpus is determined, and data collection in media corpus data through sliding window is determined; then, the state characteristics and probability distribution of feature data are determined by dynamic Bayesian network, the relationship between the state variables and dimensions of media corpus data is determined, and the media corpus data state is processed by component to complete the preprocessing of media corpus data; finally, through the application research of storage and encryption of the designed database through big data technology, the storage structure data and encryption secret key are designed to realize the construction and application of media corpus. The experimental results show that the data quality of the media corpus constructed by the proposed method is high, and its application rate has been improved to a certain extent.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Aprent, "Eine digitale analyse sterreichischer printmedien auf basis des Austrian media corpus," *Zeitgeschichte*, vol. 46, no. 4, pp. 501–534, 2018.

[2] S. Coats, "Language choice and gender in a nordic social media corpus," *Nordic Journal of Linguistics*, vol. 42, no. 1, pp. 1–25, 2019.

[3] S. Goodman and S. Kirkwood, "Political and media discourses about integrating refugees in the UK," *European Journal of Social Psychology*, vol. 49, no. 7, pp. 1145–1152, 2019.

[4] M. T. Odenkirk, D. M. Reif, and E. S. Baker, "Multiomic big data analysis challenges: increasing confidence in the interpretation of artificial intelligence assessments," *Analytical Chemistry*, vol. 93, no. 11, pp. 45–53, 2021.

[5] T. Gao and X. Wang, "The database selection based on hesitant linguistic information aggregation algorithm," *Control Engineering of China*, vol. 26, no. 8, pp. 1444–1449, 2019.

[6] Y. Wu, "Research on automatic conversion method of heterogeneous data in language database," *Computer Simulation*, vol. 36, no. 7, pp. 380–384, 2019.

[7] H. Elgibreen, M. Faisal, M. A. Sulaiman et al., "An incremental approach to corpus design and construction: application to a large contemporary saudi corpus," *IEEE Access*, vol. 9, pp. 88405–88428, 2021.

[8] L. Zhang, Y. Guan, and S. C. Jiang, "Investigations of soil autotrophic ammonia oxidizers in farmlands through genetics and big data analysis," *Science of the Total Environment*, vol. 777, no. 3, Article ID 146091, 2021.

[9] M. J. Y. Suh, H. J. Yi, H. J. Kim, and S. H. Kim, "Is asymmetric hearing loss a risk factor for vestibular dysfunction? lesson from big data analysis based on the Korean national health and nutrition survey," *Otology & Neurotology*, vol. 40, no. 10, pp. 1339–1345, 2019.

[10] P. Kumar and A. Kumar Bhatt, "Enhancing multi-tenancy security in the cloud computing using hybrid ECC-based data encryption approach," *IET Communications*, vol. 14, no. 18, pp. 3212–3222, 2020.

[11] R. Geetha, T. Padmavathy, T. Thilagam, and A. Lallithasree, "Tamilian cryptography: an efficient hybrid symmetric key encryption algorithm," *Wireless Personal Communications*, vol. 112, no. 1, pp. 21–36, 2020.

[12] H. Zhang, Z. Xu, F. Tao, Y. Li, Y. Cui, and X. Li, "New barbituric acid derivatives for data encryption and decryption based on the mechanochromic fluorescence effect," *Analyst*, vol. 145, no. 12, pp. 1109–1121, 2020.

[13] W. Sun, L. Wang, J. Wang, H. Li, and Q. Wu, "Optical hyperspectral data encryption by using gamma distributed phase masks in gyrator domain," *Optical Review*, vol. 26, 2019.
[14] Y. Chang, S. B. Zhang, L. L. Yan, and G. Wan, "A quantum secure sharing protocol for cloud data based on proxy re-encryption," *Scientific Reports*, vol. 10, no. 1, pp. 78–85, 2020.
[15] K. Wang, W. Li, Y. Liu, and T. Chen, "A high-efficient and low-cost secure AMBA framework utilizing configurable data encryption modeling against probe attacks," *IEICE Electronics Express*, vol. 18, no. 7, pp. 123–128, 2021.
[16] M. H. Saracevic, S. Z. Adamovic, V. A. Miskovic et al., "Data encryption for internet of things applications based on catalan objects and two combinatorial structures," *IEEE Transactions on Reliability*, vol. 16, no. 99, pp. 1–12, 2020.