IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

A Novel Ide Based Privacy Preserving Method For Big Data Using Paritial Least Square Regression and ε-Differential Privacy Algorithms

Johny Antony P[1*], Antony Selvadoss Thanamani[2]

[1]Research Scholar, NGM College, Pollachi, Tamilnadu, India
[2]Dept. of Computer Science, NGM College, Pollachi, Tamilnadu, India

Available online at: www.ijcseonline.org

Accepted: 19/Nov/2018, Published: 30/Nov/2018

***Abstract*:** Privacy preservation in big data is a need of the time because of the specialties of the data. Many researches have been made to tackle the issues of privacy in big data still some conflicts arises. Hence, an efficient method for the privacy preservation of data should be introduced. In this proposed work, a novel framework is designed for conserving the data in a secure manner. The personal and medical datasets are taken and are being merged which is under the control of hospital admin. This dataset is preprocessed to remove the noise following the normalization technique in order to convert the string into integers. Then, the efficient partial least square regression model is applied for the extraction of features such as sensitive and non-sensitive attributes. After the identification of this sensitive and non-sensitive attributes, ε-differential privacy preservation algorithm, the sensitive data are encrypted with the use of novel identity based encryption technique by generating the key. By the use of this code the user can decrypt the data which is anonymous format. The performance analysis is made on comparing the existing techniques which shows that the proposed methodology provides a better efficiency in terms of encryption cost, key generation cost, overall execution cost, security scheme, and computation complexity.

***Keywords*-** Privacy preservation, Sensitive attributes, Non-sensitive attributes, ε-differential privacy preservation, encryption, Identity based encryption.

I. INTRODUCTION

In recent years, big data storage has been emerged as the technology buzz. The major focus of this big data storage is the representation of size of the data. The data is capable of undergoing the analytical algorithms which is regarded as the main logic of the data analytics for it's interfere in the useful information. The security [1]and storage of big data plays a significant part because of its enormous amount of data stored and due to the occurrence of many security issues. As there are several types of big data framework such as Hadoop for providing excess space for the processing, there also arises a huge issues regarding security. Also, there are some big data storages like NOSQL database, this also contains some security concerns. For example, the data that are having[2] high-priority will be stored in the flash media. Hence, the locking of data storage represents the strategy of tier-conscious that were to be solved. The security solutions which are drawing logs from endpoints should validate the end points authenticity. The traditional hospital centric health care is not only an inefficient one but also suffers from the excessive [3]time of waiting at the hospital. Therefore, it is necessary to enhance the solution for health care monitoring system.[4] Additionally, the medical record needs novelty. There is a possibility of dispensing of patient's data at different locations because of life proceeding makes them to change from one storage tower of data provider into another provider. Because of this there is a possibility to lose the access of their health history. The ability to exchange and use information among the various hospital and the provider encounters the further added barricades to share the data with effective manner. This may cause the deficiency in the data management coordination and also in the conversion. This leads the disjointed in the health records instead of consistent. While initiate the recovery of data and the distribution, the patients and the data providers are having the possibility to face the substantial issues. This is because of the financial reasons which reassure the "health information blocking". By accusing the excessive costs for the exchange of data, the IT developers in the health sectors are intervene the data transmission flow. This is mentioned in the report of ONC recently. While execute the designing process of innovative system to overcome these kind of issues, there is need to consider the patient agency. [5]The patient can get assistance from the complete and their medical histories crystal clear picture. This shows the critical in executing the hope as well as the participation of in the health care system.

Objectives:

- To generate PLSR algorithm to select features or to identify attributes like sensitive and non-sensitive

attributes using an efficient partial least square regression model
- To generate ε-differential privacy preservation algorithm to improvise the Anonymization.
- To enhance the security of data to be transmitted on employing identity based encryption technique like AES, and cryptographic approach.
- The performance analysis is made in terms of encryption cost, key generation cost, overall execution cost, security scheme, and computation complexity

The remaining portion of the paper is schematized as follows: Section II provides the literature review of the big data security for the maintenance of medical data. Section III describes the proposed Novel IDE based Privacy Preserving method Using Learning Algorithms. Section IV offers the performance analysis and comparative analysis of the proposed work with existing methods. At last, the paper is concluded in Section V.

II. RELATED WORKS

This section describes the literature review of the big data security for the maintenance of medical data.
[6] Cloud computing was regarded as the powerful technology for performing several huge-scale and a difficult computation. The expense for the maintenance of computing hardware, software, and space needs were eliminated. There has been an increasing demand in the huge scale of data and the big data in the cloud computing. As the big data addressing was the time demanding and a challenging factor which usually requires the huge computational infrastructure for the successful ensuring of data analysis and processing. Mainly the analysis was focused on the availability, data integrity, data quality, scalability, privacy, transformation of data, and the regularity issues. Furthermore, the significant challenges and issues were addressed for ensuring the data management in a long term success thereby exploring different territories which has been a major limitation of this work.

[7]In several areas of engineering, medical and scientific studies that were a huge number of advancement in the information technology resulting to data or the information explosion. From such growing data the decision making and discovery of knowledge was the challenging factor in the computation of big data. The computation of big data requires a large storage and the data processing. Thus the evolution of cloud computing and the big data was analyzed. This work needs some improvement in the field of decision making, data organization, and the specific tools that were related to domain.

[8]Big data was concerning the large volume of growing data, complex, and the data collection capability. The model of data-driven involves the aggregation of data sources, analysis, and mining by the consideration of security and privacy. The challenging issues in the revolution of big data is analyzed. The participation of public in the circle of economic events were not done which was the major limitation of this work.

[9]discussed the security, trust and privacy issues in the internet of things. The internet of things provides the security and privacy satisfaction necessities including the confidentiality of data, access control and the authentication. The scalability issues may occur due to the huge number of interconnected devices. Hence, there was a need for static infrastructure to tackle the security threats in this dynamic surrounding. Though many efforts have been taken to analyze the issues regarding this topic, there were still a number of issues that were to be faced further which remains as the limitation of this effort.

[10] The promising possibility of the advent of Internet of Things (IoT) technologies for medical devices and sensors that were interconnected play a key role in health care industry that was yet to emerge in order to assure quality care for patients. Due to the increasing number old and physically challenged people, there was a need of real-time health monitoring infrastructure in order to analyze the patients' healthcare data to decrease mortality rate. A Health Industrial IoT (HealthIoT) enabled monitoring framework was proposed in this method, where ECG and other healthcare data was collected by mobile devices and sensors were secretly sent to cloud for seamless access by healthcare professionals. However, it was suggested that the interconnected wearable patient devices and healthcare data obtained was susceptible to security breaches. Its inability to address the concerns of data security and notification functions was considered as the limitation of this method.
[11] discussed the protection of big data. The distribution system were used as the big data needs high power for computation and the additional storage. The risk of privacy violation was increased there were several number of parties involved in this system. There were huge number of mechanisms for the privacy preserving for its protection at various stages. This work needs some improvement by executing the access control and secure End to End communication.

[12]discussed the big privacy challenges and the opportunities of the privacy analysis in the big data. The major achievements of the privacy field of theoretical angles in order to address the big data challenges thereby providing the operations and roles of the privacy system. However, these efforts remain insufficient for the most incoming application of the big data. Hence, it was necessary to make some efforts for solving the problem which remains as the major limitation of this work. [13]discussed the privacy preservation links of the genomic and clinical sets of data.

The capacity of the link records that were associated with the link record was the major challenge for the data driven research. Privacy-preserving. The preparation of technology standards and the policy for enabling reliability in a high manner and to justify the findings and the recommendations so far. Anyhow, there were some limitations which needs further improvement.

[14] the medical devices that were wearable with the sensor would generate huge number of data that was usually referred as the big data that were incorporated with the structured and unstructured data. The meta-Fog-redirection and the grouping and choosing architecture. The map reduce framework was implemented for the prediction of heart disease. Finally, the performance measure was calculated to enhance the efficiency of the patient health record monitoring system. This work needs some improvement which remains as the major drawback of this effort.

[15] Proposed a mechanism to provide the assurance for the integrity, privacy and also to attain the fine grained access control for the medical information in a flexible and effective way of security. This system mainly based on the concept of Cipher text Policy Attribute-based Encryption (CP-ABE) to attain the superior performance and flexibility. The simulation have been carried out in an extensive way which was shown the effective fine grained and measurable access control in a usual and unusual conditions.

[16] Proposed an innovative system to regulate the admittance for the EHRs which have been stored in a semi-trusted cloud servers. The influence of cipher text-policy attribute-based encryption (CP-ABE) method have been used to attain the fine-grained access control in EHRs. This technique have been used to encrypt the tables which are published by the health care institute like hospitals including the patients' EHRs. This information have been stored with the patient's unique identity by the primary key. This scheme have been evaluated by using the University of California, Irvine datasets.

[17] suggested a need for proactive health care and wellness by the tremendous increasing cost for health care and the insurance of health care premiums. At the health care industry a novel wave of digitizing the records of medical details. The industry of health care is witnessed by the rise of diversity, timeliness, and complexity. For the transformation of health care industry, big data has been emerged as a viable solution. The shifting process was made from reactive to proactive paradigm which results in the decrease of health care cost. As there was an increase in emerging threats and some vulnerabilities the issue of privacy and security begins to grow. However, there were some shortcomings like communication capabilities, storage and computation.

[18]designed and implement a method for creating the privacy and secure electronic health record linkage over multiple sites. The software application was developed and distributed the application of software for performing the process of preprocessing, data cleaning and the identifier of patients for the removal of secured health care data. Furthermore, the performance and security should be enhanced by the use of some efficient filtering approaches.
[19]suggested the record linkage of privacy-preserving for solving the problem of exchanging several data over different organizations. For the detection of fault detection, national security, and this system has been used. In big data collection the confidentiality, and preservation of privacy have been represented. Moreover, for allowing the multiple large database the development of some efficient techniques should be made by facilitating the innovative ways which remains as the major drawback of this method.

[20]presented the privacy-preserving cipher text multi-sharing control for the storage of big data. The guarantee of confidential data was the major necessity of service. The benefits of proxy re-encryption was combined with the anonymous that securely share several times in spite of leaking knowledge of plaintexts. However, there were some limitations which reduces the performance of the presented system.

III. PROPOSED WORK

This section describes the proposed Novel Big data privacy preservation methodology.

The figure 1 represents the overall flow of the proposed system. Initially, the data of owners or the patients are categorized into personal data and the medical data. These data are stored in the hospital admin. By using efficient partial least square regression model, the sensitive and non-sensitive attributes are identified. For, the privacy preservation of the sensitive data, the enhanced ε-differential privacy preservation algorithm is applied from which the encryption process was carried with the help of novel identity based encryption algorithm. Finally, the performance was evaluated to yield the efficient privacy protection regarding the patient health record system.

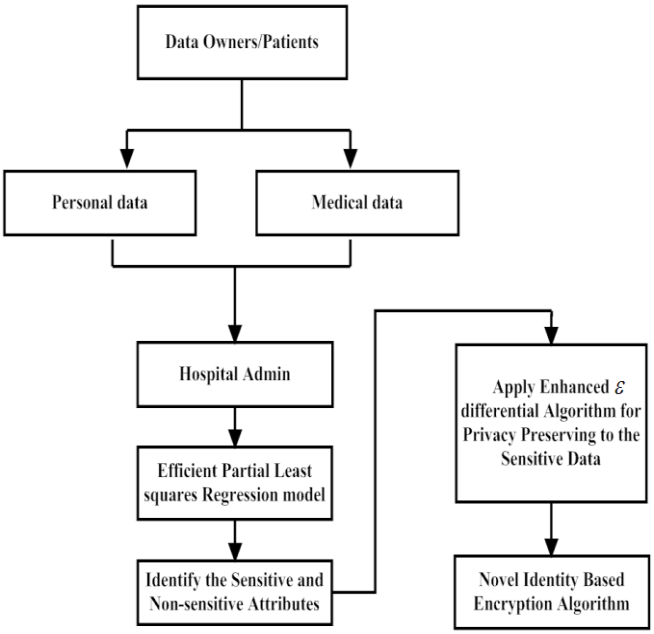

Figure 1 Overall flow of the proposed system

Initially, the electronic record management framework is considered. The data owners are the patients or the authenticated owners who are sharing their clinical or medical data to the cloud with the personal details that has to be given a high grade of security. The hospital admin is responsible for holding the huge size of data records of the whole patients that are related with the particular hospital. Moreover, the big data can be outsourced to the big data researcher's and to the data miner as per their request. The effective sharing of data or publishing thereby preserving the data are the major issues that are to be focused.

A. Pre-processing:

The personal and medical data are merged and are preprocessed. The preprocessing is done to remove the excess noise present in the data. After, the removal of noise in data the normalization techniques are performed. In this, the strings are converted to the integer values. The conversion process is carried till two digit integers are attained.

Algorithm:1 Pre-processing

Input: Input Data (In_{dt})
Output: Pre-processed Data (Pr_{dt})

Procedure:
#Pre-Processing the data.
Noise removal, normalize and rows & column count of the In_{dt}
Line = Content of the file
While line! =null
Alphabets = line split by ","
for alphabet : alphabets
fori=0 to length (alphabet)
fori=0 to length (String)
if 48<String<58
sum=sum+str.character-48
End if
End for
while (sum>0)
temp = sum%10
sum1=sum1+temp
sum = sum/10
End while
End if
End while.
Hence the data was normalized as Numeric Values.
Count no of line and rows.
Reading File
row++;
row = no of rows.
End While
Reading File
C= splitting by ","
Reading File
Wr = {r,c "/n" line}
End While
Wr = Pre-processed Data (Pr_{dt})
Pre-processed Data (Pr_{dt}) as {r,c,line}

After completing the pre-processing step we achieve the file that contain numeric values of the original file data by converting ASCII value. Contain number of rows and columns of the file.

B. Feature Selection:

The sensitive and non-sensitive attributes are selected by the use of an efficient partial least square regression model. On applying this technique, the attributes of sensitive and non-sensitive cases are identified. Initially, a covariance matrix is

formed from which the matrix are generated followed by the efficient partial least square regression model. By this, the features are selected like sensitive and non-sensitive attributes. The sensitive attributes are most essential attribute that are to be classified.

Algorithm 2: Constructing covariance Matrix

Input: Pre-processed Data (Pr_{dt})
Output: Covariance Matrix (Cvm_{dt})

Procedure:

Processing the data.
forito n ,where n is the number of the columns.
$CVM_{X,Y} = {X_i Y_i}/{N}$
Where N = Number of scores in each set of data,
Xi = i[th] raw score in the first set of scores,
Yi = i[th] raw score in the second set of scores,
$CVM_{X,Y}$ = Covariance of corresponding scores in the two sets of data
Calculate Eigen Values and Eigen Vectors
$[Cv_{mat}].[E_{Vec}] = [E_{Val}].[E_{Vec}]$
Where Cv_{mat} = Covariance Matrix
E_{Vec} = Eigen Vector
E_{Val} = Eigen Value
Attribute Scoring$SC = [ori_{dt}].[E_{Vec}]$
Where ori_{dt} = Original Data
E_{Vec} = Eigen Vector
End for

Hence Covariance Matrix was created. It is n*n Matrix depending upon the number of column present in the input data. So, first and last rows of the data will be considered as the future input to the selection step. Let consider Input Covariance Matrix (In_{Cvm}).

Algorithm 3: Feature Selection Using Least Square Matrix

Input: Input Covariance Matrix (In_{Cvm})
Output: Feature Selected Data (Fs_{dt})

Procedure:
Processing the data.
Initialize 1[st] rows as X, 2[nd] row as Y, i, Xb=0, Yb=0, Sx=0, Sy=0
For i to n
$Xb += x[i]$
$Xb /= n$
End For
For i to n
$Yb += Y[i]$
$Yb /= n$
End For
For i to n
$Sx += (X[i] - Xb) * (X[i] - Xb)$
For i to n
$Sy += (X[i] - Xb) * (Y[i] - Yb)$
End For,for
Initiallize A as Sy/Sx and B as $Yb - A * Xb$
Sensitive Value for the Data is B →Sv_{dt} = B
For i to n
Initialize Val as $A * X[i] + Sv_{dt}$
Data = $\begin{cases} Sen & If\ (Val > Sv_{dt}) \\ N.Sen & Else \end{cases}$
Feature Selected Data (Fs_{dt}) = $Data$

After completing the Least Square Matrix Algorithm, we achieved Sensitive value of the whole data and relevant sensitive values for the each attribute. Next we want to select the sensitive attribute for the anonymity task. For that we use following pseudo code.

C. Anonymization Technique:

The ε-differential privacy preservation algorithm for the preservation of privacy to the sensitive data are enhanced and are applied for the sensitive case. Hence, the sensitive attributes are taken from the selected features for conserving the data that are most sensitive.

Algorithm 4: Anonymity using differential algorithm.

Input: Feature Selected Data (Fs_{dt})
Output: Anonymity Data ($Anyn_{dt}$)

Procedure:
processing the data.
Fixing Anonymity type (A_{Ty})
(A_{Ty}) →$Fs_{dt} \epsilon S_{dt} = Cc_{txt}$
$Fs_{dt} \epsilon Nm_{dt} = Hc_{txt}$
Where S_{dt} = String Data, Nm_{dt} = Number Data, Cc_{txt} = Caesar Cipher text,
Hc_{txt} = Hash Code text.
Hence $Anyn_{dt} = \sum A_{Ty}$.

After implementing differential algorithm in data we achieved the anonymity for the sensitive data by using the Caesar cipher for string and hash code for the numbered data. Those Data was re-merged with the original data with relevant column in data set. The New data was Achieved called Anonymized Data set ($Anyn_{ds}$)

D. Encryption Technique:

A novel identity based encryption algorithm is utilized for the process of encrypting the sensitive data. The key is generated with user id and admin id. With the generated key the user can use this id for decrypting the data. The user can attain the decrypted data in an anonymized format.

Algorithm :5IDE based Enc/Dec using AES Algorithm

Input: Anonymized Data set ($Anyn_{ds}$)

Output: Encrypted Data (Enc_{dt}) and Decrypted Data (Dec_{dt})

Procedure:
Processing the data.
Initialization:
$Alm_{aes}$and Tns_{aes}.
Pu_{key} And Pr_{key}= R
Key_{ky} = Pu_{key} + Pr_{key}
Secret Key Generation:
Sc_{key} = Key_{bytes} * Alm_{aes}
Cipher creation:
Cp_c = Cp_c * Tns_{aes}
Crypto Text Creation;
Cy_{txt} = Cp_c * Sc_{key}

Encryption:
Enc_{dt} = Fun (Cy_{txt} * Cp_c ∈ $Key_{ky}$and $Anyn_{ds}$,Enc_{dt})
Function Ends.
Decryption:
Dec_{dt} = Fun (Cy_{txt} * Cp_c ∈ $Key_{ky}$and $Anyn_{ds}$,Dec_{dt})
Function Ends

Abbreviations:
Alm_{aes} as Algorithm of AES, Tns_{aes} as Transportation of AES, Pu_{key}as Public Key, Pr_{key}as Private Key, Key_{ky} as combination of both Private and Public key, R as Random Numbers, Key_{bytes}as bytes of the key, Sc_{key}as Secret Key, Cp_c as Cipher Text, Cy_{txt} as Crypto text, Enc_{dt}as Encryption of the text, Dec_{dt}as decryption of the Text.

After implementing this algorithm, we achieved the Encryption and Decryption of the Anonymized data with the help of AES Algorithm based on IDEwith the use of AES, and cryptographic approach.

IV. PERFORMANCE ANALYSIS

A. *Dataset Description:*

In this, the dataset like wiscoin breast cancer dataset is taken which is the medical dataset. First 5 Columns from Mockaroo Open Dataset. From 6 to 15 was taken from Wiscoin breast Cancer Dataset. Similarly, mockaroo open data is taken which is the personal dataset of the patient.

B. *File Upload Case:*

Key Generation Time of Proposed Algorithm:
The Figure 2 shows the proposed algorithm's Key generation Time. For different types of file size the value of key generation time is obtained and plotted the graph. From this graph, it is observed that, if the file size increases, the value of key generation time is decreases. For the value of file size of 10 to 1000, the value of key generation time is uniform.

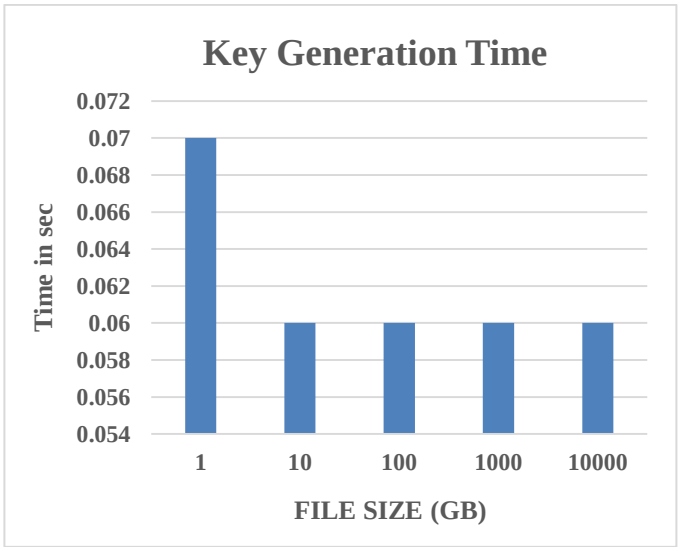

Figure 2. Performance analysis of the key generation time

Encryption Time:
The Figure 3 illustrates the proposed algorithm's Encryption Time. For different types of file size the value of key generation time is obtained and plotted the graph. From this graph, it is observed that, if the file size increases, the value of encryption time is also increases.

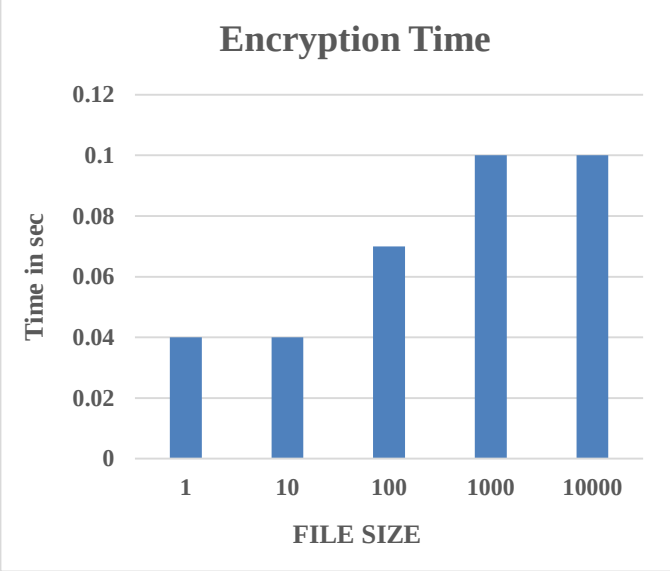

Figure 3 Performance analysis of the Encryption Time

File transmission:

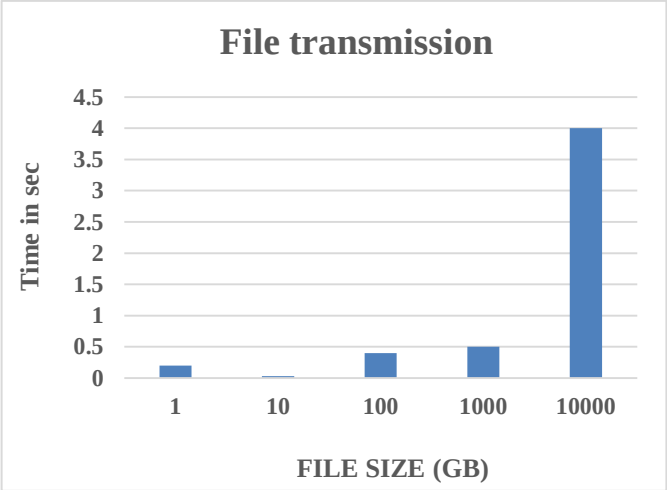

Figure 4 Performance analysis of the file transmission

The plot shows the performance of the proposed algorithm.

V. COMPARATIVE ANALYSIS:

The comparative analysis of the existing and proposed method is performed as follows:

A. File upload:

The comparative analysis of the proposed and existing methods are shown in terms of file upload which proves that the proposed method performs well than the existing methods.

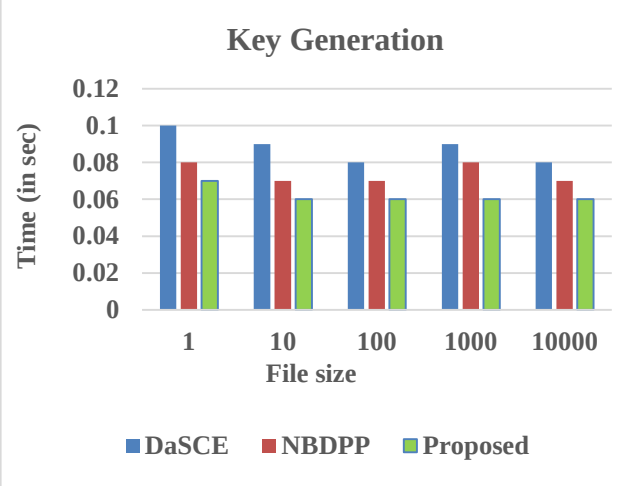

Figure 5 Comparative analysis of key generation

The plot shows the key generation time comparison between the existing algorithms such as DaSCE, NBDPP with the proposed while uploading the file. The x axis exhibits the size of the file. The Y axis denotes the Key Generation Time. It is clearly known that the proposed algorithm shows the better key generation time comparing existing algorithms along the variation of the file size.

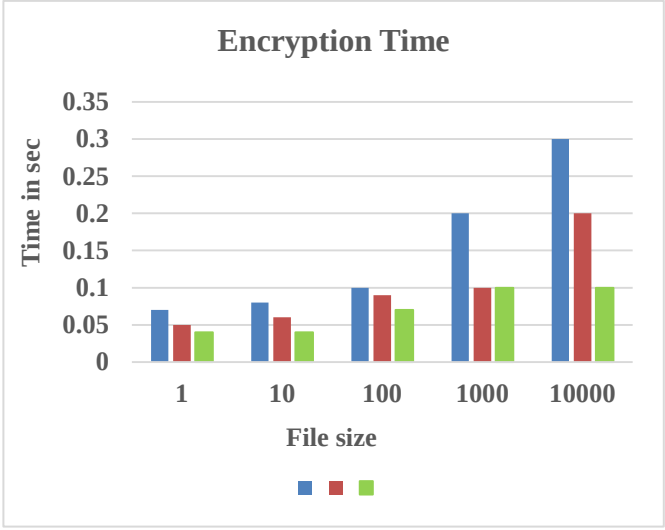

Figure 6 Comparative analysis of Encryption time

The above graph shows the Encryption time while uploading the file between the existing algorithms such as DaSCE, NBDPP with the proposed. The x axis exhibits the size of the file. The Y axis denotes the Encryption Time. It is clearly known that the proposed algorithm shows the better Encryption time while comparing with the existing algorithms along the variation of the file size.

Table 1 Comparative analysis of file transmission time

File Size	DaSCE	NBDPP	Proposed
1	0.4	0.3	0.2
10	0.5	0.04	0.03
100	0.7	0.5	0.4
1000	0.8	0.6	0.5
10000	1	0.5	0.4

The above table shows the values of file transmission time for the existing as well as the proposed. By varying the file size the above values are acquired while uploading the file. This comparison table depicts the proposed algorithm having the optimized time of file transmission.

Table 2 Comparative analysis of file transmission timeof uploading

File size	DaSCE	NBDPP	Proposed
1	0.09	0.07	0.6
10	0.09	0.07	0.6
100	0.09	0.07	0.6
1000	0.09	0.07	0.6
10000	0.09	0.07	0.6

The above table demonstrates the values of the transmission time for key. In this terms also the proposed algorithm shows the better time for the file transmission while uploading the file.

For File downloading cases, the key generation time, Encryption time, key transmission time, file transmission time are compared with the existing algorithms that shown below:

B. File upload updated:

The comparative analysis of the proposed and existing methods are shown in terms of file upload updated which proves that the proposed method performs well than the existing methods.

Table 3 Comparative analysis of key generation time of file download updated

	Key Generation		
File size	DaSCE	NBDPP	Proposed
1	0.1	0.08	0.07
10	0.09	0.07	0.06
100	0.08	0.07	0.06
1000	0.09	0.08	0.06
10000	0.08	0.07	0.06

Table 4Comparative analysis of encryption time of file download updated

	Encryption time		
File size	DaSCE	NBDPP	Proposed
1	0.07	0.05	0.04
10	0.08	0.06	0.04
100	0.1	0.09	0.07
1000	0.2	0.1	0.1
10000	0.3	0.2	0.1

Table 5 Comparative analysis of file transmission time of file download updated

	File Transmission		
	DaSCE	NBDPP	Proposed
1	0.4	0.3	0.2
10	0.5	0.04	0.03
100	0.7	0.5	0.4
1000	0.8	0.6	0.5
10000	1	0.5	0.4

Table 6 Comparative analysis of key transmission time of file download updated

	Key Transmission		
	DaSCE	NBDPP	Proposed
1	0.09	0.07	0.6
10	0.09	0.07	0.6
100	0.09	0.07	0.6
1000	0.09	0.07	0.6
10000	0.09	0.07	0.6

File download:

The comparative analysis of the proposed and existing methods are shown in terms of file download which proves that the proposed method performs well than the existing methods.

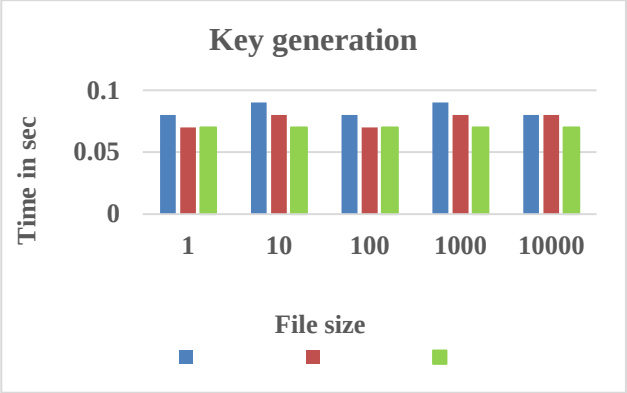

Figure 7 comparative analysis of key generation time

The above figure shows the comparison of key generation time between the existing algorithms such as DaSCE, NBDPP with the proposed. The x axis exhibits the size of the file. The Y axis denotes the Key Generation Time. From this graph, it is clearly known that the proposed algorithm shows the better key generation time comparing existing algorithms along the variation of the file size.

Table 7 Comparative analysis of encryption time

	Encryption Time		
File Size	DaSCE	NBDPP	Proposed
1	0.05	0.03	0.02
10	0.06	0.04	0.02
100	0.07	0.05	0.04
1000	0.08	0.06	0.05
10000	0.2	1	0.9

The above table express the comparison of Encryption time between the existing algorithms such as DaSCE, NBDPP with the proposed. It is clearly known that the proposed algorithm shows the better Encryption time while comparing with the existing algorithms while varying the file size.

Table 8 Comparative analysis of File Transmission

	File Transmission		
File Size	DaSCE	NBDPP	Proposed
1	0.08	0.07	0.06
10	0.09	0.07	0.06
100	0.08	0.07	0.06
1000	0.9	0.07	0.06
10000	0.08	0.07	0.06

By varying the file size the above values are acquired. This evaluation table describes the proposed algorithm having the optimized time of file transmission.

Table 9 Comparative analysis of Key Transmission

	Key Transmission		
	DaSCE	NBDPP	Proposed
1	0.05	0.05	0.04
10	0.1	0.05	0.04
100	0.8	1	0.9
1000	1	5	4
10000	25	8	7

The comparative analysis of key transmission is shown in table which shows that the key transmission time is less for various size of file than the other existing techniques.

C. File download updated:

The comparative analysis of the proposed and existing methods are shown in terms of file download updated which proves that the proposed method performs well than the existing methods.

Table 10 Comparative analysis of Key generation in file download updated

	Key Generation				
File size	Elgamal	DH	DaSCE	NBDPP	Proposed
1	1	1	0.08	0.07	0.06
10	1.2	1.5	0.09	0.08	0.06
100	1.9	1.6	0.08	0.07	0.06
1000	2.54	1.72	0.9	0.08	0.06
10000	3.5	4.5	0.08	0.08	0.06

Table 11 Comparative analysis of encryption time in file download updated

	Encryption Time				
	Elgamal	DH	DaSCE	NBDPP	Proposed
1	0.9	0.7	0.05	0.03	0.02
10	1.4	1.5	0.06	0.04	0.03
100	1.5	1.9	0.07	0.05	0.04
1000	3	4	0.08	0.06	0.05
10000	4.6	5.3	0.2	1	0.9

Table 12 Comparative analysis of file Transmission in file download updated

	File Transmission				
	Elgamal	DH	DaSCE	NBDPP	Proposed
1	1	1	0.08	0.07	0.06
10	1.3	1.6	0.09	0.07	0.06
100	1.5	1.9	0.08	0.07	0.06
1000	2.5	2.8	0.9	0.07	0.06
10000	3.5	3.9	0.08	0.07	0.06

Table 13 Comparative analysis of Key Transmission in file download updated

	Key Transmission				
	Elgamal	DH	DaSCE	NBDPP	Proposed
1	1.2	1.5	0.05	0.05	0.04
10	1.6	1.8	0.1	0.05	0.04
100	1.9	2	0.8	1	0.8
1000	2.5	2.6	1	5	4.9
10000	3.5	3.6	25	8	1.7

D. Comparative analysis of Security:

The comparative analysis of the proposed and existing methods are shown in terms of security which proves that the proposed method performs well than the existing methods.

Table 14 comparative analysis of security

	CO-Res	Revocation		Confidentiality		PrSec	Interg rity
		B	F	Ag cloud	Ag User		
DACC	Yes	Yes	No	Yes	Yes	Yes	No
DAC-MACS	No	No	Yes	Yes	No	Yes	No
NEDAC-MACS	yes	yes	Yes	Yes	Yes	Yes	No
NBDPP	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Proposed	YES	YES	YES	YES	YES	YES	YES

VI. CONCLUSION

Nowadays, big data security is an important aspect in several applications including medical data preservation which was more confidential. To improvise the security of data to be stored, an efficient technique should be introduced in our work. In this, both personal and medical data of the patient was merged and preprocessed by removing noise and

applying some normalization techniques. The feature selection was carried by extracting the sensitive and non-sensitive attributes on employing the efficient least partial square regression model. By the use of ε-differential privacy preservation algorithm, the sensitive attributes are anonymized. Finally, the encryption technique was performed with the utilization of novel identity based encryption algorithm. The performance analysis was made and the comparison of the existing and proposed method shows that the proposed methodology was more efficient than the existing techniques in terms of encryption cost, key generation cost, overall execution cost, security scheme, and computation complexity.

REFERENCES

[1] S. Ananthi and A. Periwal, "Data Security Based On Big Data Storage," *Global Journal of Pure and Applied Mathematics*, vol. 12, pp. 1491-1500, 2016.

[2] K. He, *et al.*, "On the security of two identity-based conditional proxy re-encryption schemes," *Theoretical Computer Science*, vol. 652, pp. 18-27, 2016.

[3] S. Wang, *et al.*, "Attribute-based data sharing scheme revisited in cloud computing," *IEEE transactions on information forensics and security*, vol. 11, pp. 1661-1673, 2016.

[4] A. Azaria, *et al.*, "Medrec: Using blockchain for medical data access and permission management," in *Open and Big Data (OBD), International Conference on*, 2016, pp. 25-30.

[5] J. Luo, *et al.*, "Big data application in biomedical research and health care: a literature review," *Biomedical informatics insights*, vol. 8, p. BII. S31559, 2016.

[6] I. A. T. Hashem, *et al.*, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98-115, 2015.

[7] R. Kune, *et al.*, "The anatomy of big data computing," *Software: Practice and Experience*, vol. 46, pp. 79-105, 2016.

[8] X. Wu, *et al.*, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, pp. 97-107, 2014.

[9] S. Sicari, *et al.*, "Security, privacy and trust in Internet of Things: The road ahead," *Computer networks*, vol. 76, pp. 146-164, 2015.

[10] M. S. Hossain and G. Muhammad, "Cloud-assisted industrial internet of things (iiot)–enabled framework for health monitoring," *Computer Networks*, vol. 101, pp. 192-202, 2016.

[11] A. Mehmood, *et al.*, "Protection of big data privacy," *IEEE access*, vol. 4, pp. 1821-1834, 2016.

[12] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE access*, vol. 4, pp. 2751-2763, 2016.

[13] D. Baker, *et al.*, "Privacy-Preserving Linkage of Genomic and Clinical Data Sets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[14] G. Manogaran, *et al.*, "A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system," *Future Generation Computer Systems*, vol. 82, pp. 375-387, 2018.

[15] A. Lounis, *et al.*, "Healing on the cloud: Secure cloud architecture for medical wireless sensor networks," *Future Generation Computer Systems*, vol. 55, pp. 266-277, 2016.

[16] C. Guo, *et al.*, "Fine-grained database field search using attribute-based encryption for e-healthcare clouds," *Journal of medical systems*, vol. 40, p. 235, 2016.

[17] H. K. Patil and R. Seshadri, "Big data security and privacy issues in healthcare," in *Big Data (BigData Congress), 2014 IEEE International Congress on*, 2014, pp. 762-765.

[18] A. N. Kho, *et al.*, "Design and implementation of a privacy preserving electronic health record linkage tool in Chicago," *Journal of the American Medical Informatics Association*, vol. 22, pp. 1072-1080, 2015.

[19] D. Vatsalan, *et al.*, "Privacy-preserving record linkage for big data: Current approaches and research challenges," in *Handbook of Big Data Technologies*, ed: Springer, 2017, pp. 851-895.

[20] K. Liang, *et al.*, "Privacy-preserving ciphertext multi-sharing control for big data storage," *IEEE transactions on information forensics and security*, vol. 10, pp. 1578-1589, 2015.

AUTHORS BIOGRAPHIES

Mr. Johny Antony P is a research Scholar of the Department of Computer Science, NGM College, Pollachi affiliated to Bharathiyar University, Coimbatore. He has ten years of teaching experience and seven years of administrative experience. His areas of interest are Data Mining, Big Data and Software Engineering. He has published and presented many papers in national and international journals.

Dr. Antony Selvadoss Thanamani is currently working as Associate Professor and Head of the Department of Computer Science, NGM College, Pollachi affiliated to Bharathiyar University, Coimbatore. He has been the Principal Investigator of UGC – Major research project in Computer Science. He has published many papers in national and international journals and written many books. His areas of interest are E-learning, Software Engineering, Data Mining, Net Working etc. He has to his credit 26 years of teaching and research experience. He is a life member of computer society of India, Life Member of Indian Society for Technical Education, Life Member of Indian Science Congress, Life Member of Computer Science Teachers Association, New York and Member of Computer Science , Teachers' Association, India

