A Practical Methodology for Anonymization of Structured Health Data

Amin Aminifar, Yngve Lamo, Ka I Pun and Fazle Rabbi

October 14, 2019

A Practical Methodology for Anonymization of Structured Health Data

Amin Aminifar[1], Yngve Lamo[1], Ka I Pun[1,2], and Fazle Rabbi[3]

[1]Western Norway University of Applied Sciences, Bergen, Norway {amin.aminifar, yngve.lamo, ka.i.pun}@hvl.no

[2]University of Oslo, Oslo, Norway

[3]University of Bergen, Bergen, Norway {fazle.rabbi}@uib.no

Abstract

Hospitals, as data custodians, have the need to share a version of the data in hand with external research institutes for analysis purposes. For preserving the privacy of the patients, anonymization methods are employed to produce a modified version of data for publishing; these methodologies shall not reveal the patient's information while maintaining the utility of data. In this article, we propose a practical methodology for anonymization of structured health data based on cryptographic algorithms, which preserves the privacy by construction. Our initial experimental results indicate that the methodology might outperform the existing solutions by retaining the utility of data.

Keywords

Anonymization, privacy-preserving data sharing, structured health data, data mining, cryptography.

1 INTRODUCTION

Hospitals, nowadays, are increasingly collecting data from patients as it allows to provide better treatment and precise diagnosis. Analyzing such data by sharing it with researchers can be useful for society. However, the shared data should not compromise the privacy of the individuals. Removing the identifier fields like name and address, is not enough for preserving privacy from certain attacks, e.g., linking attack [1]. Such attacks can re-identify the individuals and reveal specific information based on the raw data. One solution to this is that the data custodians, e.g., hospitals, anonymize such data before sharing.

1.1 Anonymization

Having access to high-quality data is a necessity for medical and pharmaceutical experts and researchers for facilitating decision making. Sharing healthcare data can benefit several parties, including hospitals, medical and pharmaceutical researchers outside the hospital, patients, and data mining researchers. Hospitals, more precisely, medical experts and researchers, can make use of the result of data analysis performed by external research centers. Medical practitioners and pharmaceutical researchers outside the hospital need the data for analysis leading to informed decision making. Patients, indirectly through this, will receive better services from hospitals and medical centers outside the hospital. Finally, data mining researchers will have access to real health data and use them as benchmarks for their methods. However, raw health data contains patients' sensitive information and can compromise their privacy. Therefore, health data holders are looking for anonymization techniques that prepare the health data for release, while keeping the quality of data and preserving the privacy of patients.

Patients consider hospitals as trustworthy entities, so they are willing to share their data with hospitals. Nevertheless, this trust is not transitive to other entities such as research centers outside the hospitals. Many believe that removing specific identifying information including name, telephone, and social security number, is sufficient for releasing the data. As several previous studies show [1, 2], merely removing the identifier fields is deficient for preserving the privacy of individuals. Sweeney [1] shows, an adversary by having limited information from an individual, say from another dataset, can match other attributes, called quasi-identifiers (QID), and reidentify the individual. Three prominent examples about this are provided in [1, 3-6, 7].

At some points, hospitals, instead of analyzing the data by themselves and sharing the analysis results, e.g., statistics or classifiers, need to share the data with external research centers, e.g., universities and pharmaceutical companies, in order to make use of other professional resources outside. Therefore, they should share the data with external researchers specialist in data analysis. Moreover, having the data give much freedom to external research centers for data analysis. Frequent requests from hospitals for providing statistical information and fine-tuning the data mining results is not feasible [2].

1.2 Motivational Example

Hospitals are considered to be the trusted party, and thus have access to the raw data. However, they, in general have limited resources for some specific data analyses. Therefore, it is common to delegate the analysis process to external research institutions. To preserve privacy of individuals, data should be anonymized in the hospitals, and only anonymized data can be shared with external institutions or released to the public. Note that any party external to hospitals can be the adversary, as illustrated in Figure 1.

After analyzing the published data, the results will be released to the hospital, which can be, for instance, a discriminator function as the outcome of the learning from anonymized data. With this function, the hospital can

Paper from the Scandinavian Conference on Health Informatics, 12 -13 Nov, Oslo Norway.

Conference Proceedings published by Linköping University Electronic Press.

classify new raw records as follows: firstly, the new record should be anonymized in the same way as the published data anonymized; secondly, the new anonymized record can be passed to discriminator function, shared by the external institutions, for classification. In this way, hospitals can make use of services outside without compromising the privacy of their patients.

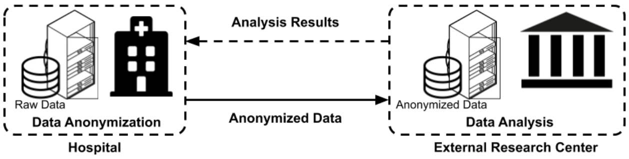

Figure 1 medical data anonymization and analysis.

In this paper, we propose a methodology to anonymize structured health data based on cryptographic algorithms and without assumptions on the characteristics of the encryption method. Adopting cryptographic algorithms guarantees privacy preservation by construction. Moreover, the comparison results of the data utility between raw and anonymized data generated based on our proposed methodology and the existing methods are promising. The proposed methodology can have a complementary role in combination with previous methods as well.

The organization of the rest of this article is as follows. In Section 2, a short review of previous methods for anonymization of the structured data is provided. Section 3 addresses the proposed approach for anonymization, along with providing some preliminary information. Section 4 presents the necessary information and settings concerning the experiments. Section 5 is devoted to the evaluation and experimental results. Finally, in Section 6, conclusions and future research directions are provided.

2 RELATED WORKS

For research purposes, data custodians need to release a version of data in a way that individuals cannot be re-identified. Statistical and multi-level databases are among the other approaches for addressing these kinds of needs. Despite the assumption made in [1], statistical disclosure control [8] is an active research area for addressing today's needs to provide accurate information while protecting the privacy of the various parties involved [9, 10]. On the other hand, anonymization techniques are between other solutions in this regard. For sharing the data records, microdata, in anonymization, we try to irreversibly alter the personal data until the re-identification of data subjects is no longer possible [11].

Anonymization methods provide a new class of acceptable solutions to this problem. Typically, anonymization techniques for structured data make use of generalization method. More specifically, such techniques modify or generalized the data records components in a way that a data record is hardly distinguishable from others. Some important related studies are k-anonymity [1], l-diversity [12], t-closeness [13], and LKC-privacy [2]. To date, k-anonymity remains the most widely known privacy model for anonymization during the past two decades. To thwart privacy threats, k-anonymity privacy model generalizes and suppresses data record components or features into equivalence groups so that any record is indistinguishable from at least k other data records [14, 2]. However, in this method, when the dimensionality of data is high, most of the data must be generalized or suppressed for achieving k-anonymity; this negatively affects the utility of data and degrades it [2]. Other methods try to rectify the issue, for instance, by imposing limitations on the problem, such as the supposition of limited knowledge of the adversary about the patient. For example, in the LKC-privacy model, the adversary is supposed to have only the values for a part of the QID attributes of the victim's record, L attributes [2].

The proposed approach in this study described in Section 3 tries to provide a solution for the above problem, i.e., anonymization of structured data. The problem here is the same as the one described in the above research studies, while we formally define the problem in Section 3. The proposed approach of this study for the solution is completely different from that provided in the previous studies. This study investigates the application of cryptographic algorithms, which is distinguishing from previous works. The majority of previous studies consider performing machine learning over homomorphically-encrypted data [15-18], while in this paper we do not make such assumptions.

3 METHOD

In this section, we first define the anonymization problem and then propose a practical solution to this problem. Two main concerns for data anonymization is privacy preservation and data utility, discussed in the following subsection. There is often an inherent trade-off between these two metrics. At one extreme, all data can be released, for maximizing the utility, and as a result, violate the privacy entirely. On the other extreme, releasing no data can maximize privacy; however, there would be no data utility [14]. The proposed methodology in this section provides an approach for addressing this problem, which is based on cryptography for data anonymization.

3.1 Problem Definition

In the following two subsections we discuss the two criteria for this problem. We define the anonymization problem as guaranteeing the privacy while maximizing the utility of the data for the statistical and machine learning data analysis.

3.1.1 Privacy Preservation

This section explains the privacy threats for sharing the raw information through an example; there exist two types of privacy concerns, namely identity linkage and attribute linkage. Table 1 shows the raw patient data. The raw data does not have the identifier features but is still vulnerable to the violation of privacy. Education, sex, and age are quasi-identifying attributes [1]. Disorder is the sensitive feature that the adversary does not know about the victim patient and tries to infer it. Finally, there exists one class for every record in the dataset.

Based on the following assumptions about the adversary, there are two types of privacy concerns to address. As mentioned in Introduction, the adversary is assumed to have anonymous data for all the patients. Moreover, the adversary has parts of the victim patient's record, in its raw format; this information is part of or all the quasi-identifying attributes and is only for one patient. The extent

ID	Quasi-identifier (QID)			Sensitive	Class
	Education	Sex	Age	Disorder	
1	BSc	F	40	Depression	cat. #1
2	MSc	M	53	ADHD	cat. #1
3	HS-grad	F	40	Depression	cat. #2
4	PhD	F	31	Social Anxiety	cat. #1
5	MSc	M	31	Bipolar	cat. #2

Table 1 An example of raw data table.

of adversary's information about the victim patient is assumed differently in different studies. For instance, in [1] the author for k-anonymity model assumes that the adversary has all the values for quasi-identifying attributes, but in [2] in LKC-privacy model limits the adversary's information to only the values of L number of the quasi-identifying attributes. Finally, the adversary does not know about the sensitive information of the victim and is willing to infer it. Accordingly, hospitals face two common privacy concerns [2] described below:

- **Identity Disclosure:** If the record is highly specific, matching the records with the victim's information is simple, which lead to the inference of the patient's sensitive information. For instance, in Table 1, the raw data table, if the adversary knows that the victim's education and age are 'MSc' and '31', respectively, then s/he confidently identifies that record number 5 is the victim's and infers that the victim's disorder is 'Bipolar'.
- **Attribute Disclosure:** If with some quasi-identifying attributes, the sensitive value happens repeatedly, it makes the inference of the sensitive value easy, although the accurate data record of the victim is not identifiable. For instance, in Table 1, the raw data table, if the adversary knows that the victim's sex and age are 'F' and '40', respectively, then, s/he can match the victim's information to records number 1 and 3. However, since both sensitive values for record number 1 and 3 are the same, 'Depression', then, the adversary can infer with 100% confidence that the victim's disorder is 'Depression'.

3.1.2 Utility of Data

To make sure that the anonymization method is not degrading the utility of the data, a comparison of the utility of raw data with the anonymous data is essential. The classification performance is a valid criterion for making a comparison between the utility of data before and after anonymization. Since the main concern of this study is sharing the data for data mining purposes, the difference between the classification performance for the raw and anonymized data shows the excellence and efficiency of the algorithm.

Information gain [19] is another criterion that indicates how much a method may degrade or improve the data quality for every feature of the data individually. Information gain was first introduced for decision trees and is based on the information entropy [20]. Nevertheless, since it does not consider the correlation and combination of the attributes, it is not as reliable as the classification performance criterion.

3.2 The Anonymization Method

For the preservation of privacy, we seek a function to map each unique record of raw data to another unique record, different from the raw record and in the same feature space. The anonymized data records must be different enough to prevent identity and attribute attacks. The anonymized data must not allow the possibility for the adversary to map back to the raw data. Therefore, the utilized function for mapping the raw data must not be reversible, or in other words must be one-way, for those with whom the anonymized data will be shared.

Cryptography fulfills the privacy objectives by construction. Mapping a number to another unique number through one-way functions is the main purpose of cryptography. Therefore, by such intrinsic features of cryptographic algorithms, we can make sure of the preservation of privacy criterion without taking further actions. Since, after encryption, the values would be meaningless numbers for the adversary, and it is not possible for one without a key to map back to the raw data.

Due to the objective of this study for anonymization of the structured health data containing categorical and numerical features, encryption is entirely feasible. Since in both cases there are numbers, more precisely category numbers and numerical values, which are mapped to other numbers. The sensitive attribute is not an exception and is encrypted as well. Normalization of data is the second phase of anonymization. Normalization, in addition to the positive impact on learning, reinforces preserving the privacy as this is a hashing phase after encryption.

As described earlier the anonymization methods should fulfill two criteria, namely privacy preservation and data utility. Application of cryptographic algorithms guarantees the privacy preservation criterion by construction. However, we also need to make sure about the performance of this methodology in regard to the utility of data. In this study, we experimentally show that our proposed methodology for anonymization of structured data is also efficient regarding the data utility.

The utility of the data needs to be preserved and this is related to the correlation of attributes and labels in data samples and the algebraic distance of samples from each other. To ensure satisfying this criterion after encryption and normalization of the dataset, the utility of the data is compared before and after anonymization based on two measurements described previously in this section. If the results for raw and anonymized data are close, then in addition to the preservation of the privacy, there also would be a confidence about the utility of data. A loss to a limited extent in the utility of data is acceptable as there exists a trade-off between privacy and data utility in data anonymization [14].

4 EVALUATION SETUP

4.1 Dataset for Evaluation of the Methodology

Adult dataset [21] is the de facto benchmark for evaluation of anonymization models [2, 12, 22-27]. In this dataset, the samples belong to two different classes; the rates of the positive and negative classes are 76.07% and 23.93%. The total number of records is 48842 (train=32561, test=16281), and the train and test sets were separated when shared. Each record has 14 attributes, including eight categorical and six numerical ones. Furthermore, the dataset contains missing values. This study considers all the attributes as QID, although it is possible to suppose part of them as QID, like in [2] which considers marital-status as sensitive and others as QID attributes.

4.2 Encryption Algorithms

For the evaluation of the proposed approach, four cryptographic algorithms, including two from symmetric and two from asymmetric encryption systems, are considered. The symmetric algorithms are Advanced Encryption Standard (AES) and Data Encryption Standard (DES); the input and output data and key size for each is 128 and 64 bits, respectively. The Asymmetric algorithms are RivestShamirAdleman (RSA) and ElGamal, which both are also homomorphic over multiplication. The key size for each is 2048 and 1024 bits, respectively. All the keys are generated randomly for every iteration of experiments, based on the toolbox.

4.3 Comparison with K-Anonymity

In order to evaluate the results of our methodology, a comparison between the results of the proposed and former methods of anonymization is necessary. K-anonymity is one of the most popular privacy models. In [28], the authors propose Mondrian for obtaining k-anonymity. This study considers this work for anonymizing the data based on the k-anonymity model for comparison with the proposed methodology. The corresponding parameters for these methods are k, set of QID, and the mode of the algorithm, which can be either relaxed or strict. In the experiments, k is set to 10 and QID are set to all the attributes, and the results for both relaxed and strict modes are provided.

4.4 Utility Measure

Two measures employed here for evaluation of data utility are information gain and classification performance. Information gain is based on information entropy and is being used to evaluate how well an attribute alone predicts the classes for samples in comparison to other attributes. In other words, every attribute is used to categorize samples, then the information entropy of the classes of the categorized samples are calculated. The lower the entropy of the samples' classes in each category of samples categorized based on that specific attribute, the higher the information gain of that attribute. The loss of information gain after anonymization can indicate the extent of deterioration of data. However, since this measure does not consider the combination of attributes, it is not as reliable as classification performance. For calculation of classification performance, we used the geometric mean of the ratios of correctly classified samples to the number of samples in that particular class. Geometric mean is the only correct average for normalized measurement [29].

5 EVALUATION RESULTS

To evaluate the efficiency of our proposed methodology, the Adult dataset [21] is anonymized with the proposed methodology by this paper. Afterward, the information gain and classification performance for raw and anonymized data are calculated and recorded for comparison and evaluation. The closer the results of raw and anonymized data the higher our confidence to the anonymization methodology regarding the preservation of data utility.

As mentioned earlier, after one level of encryption, we need to normalize the data in order to obtain the anonymized data. The normalization method used for our experiments is min-max normalization:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{1}$$

where x_{new} is the normalized value of x, the encrypted number, and x_{min} and x_{max} are respectively minimum and maximum values of the corresponding column in the matrix of encrypted numbers.

Furthermore, for more certainty, the experiments for every method iterates for ten times, and the average results are measured. In every, iteration the key for encryption algorithms are generated separately and randomly, to ensure the classification results are independent of the keys.

5.1 Information Gain

The encryption is particularly useful when the attribute is numerical since, concerning the learning results, encryption of the number of categories is similar to mapping each specific category number to another random number specific for that category; therefore, for such attributes, encryption is not a necessary process. However, in this study's experiments, we encrypted all the attributes and normalized the data afterward. Before and after anonymization by this methodology, the information gain of categorical attributes always remains the same, because of the characteristics of this measure, so there would be no points in reporting them here.

Table 2 presents the information gain of the numerical attributes of raw and anonymized datasets; the results are from the average for ten independent iterations. The results in this table show that our anonymization methodology does not reduce the information gain of the numerical attribute unless in attributes 1 and 13, albeit negligible. Considering the information gain, the proposed methodology preserves the utility of data to a considerable extent.

5.2 Classification Performance

In addition to the anonymization with the proposed methodology of this paper, for comparison, we also anonymized the Adult dataset with Mondrian multidimensional k-anonymity approach [28]. Then, the results of these methods, along with the raw dataset, are used for learning a classification function. The learning algorithm used in this research is the random forest algorithm [30]. The training and testing sets for the raw data and anonymized data based on our proposed methodology are the same as published in [21]. However, for Mondrian multidimensional k-anonymity approach for every iteration, we take 70% of randomly shuffled data as the training set and the remaining 30% as the testing set; splitting the train and test sets for learning and evaluation in this setting is conventional and valid, considering the studies in the field [31].

Table 3 exhibits the classification performance based on the geometric mean measure, i.e., geometric mean of the ratios of correctly classified samples to the number of samples in

DATASET	INFORMATION GAIN					
	Attribute 1	Attribute 3	Attribute 5	Attribute 11	Attribute 12	Attribute 13
Raw Data	0.09754	0	0.09328	0.11452	0.05072	0.05814
Anonymized Data (RSA Alg.)	0.096839	0	0.093379	0.118778	0.051108	0.057001
Anonymized Data (ElGamal Alg.)	0.097563	0	0.093507	0.118503	0.05157	0.056479
Anonymized Data (DES Alg.)	0.096581	0	0.093452	0.118688	0.051163	0.05713
Anonymized Data (AES Alg.)	0.096755	0	0.093434	0.118512	0.051061	0.057325

Table 2 Information Gain for numerical attributes of the Adult dataset [21] before and after anonymization.

that particular class, for raw and anonymized data obtained adopting several methods. All the results in Table 3 are the average of the results of ten independent iterations. The information gain table provided in this article is calculated using WEKA software [32]. The difference between the classification performance of anonymized data based on our methodology and the raw data is less than 3%; our proposed methodology, however, outperforms Mondrian multidimensional k-anonymity regarding classification performance for adult dataset as the results show that the geometric mean measure for our anonymization approach, in the worst case, is higher for at least 5%.

Dataset	Geometric Mean (%)
Raw Data	75.37
Anonymized Data (K-Anonymity Mondrian [21], Relaxed, K=10, QI = Attribute 1-14)	67.87
Anonymized Data (K-Anonymity Mondrian [21], Strict, K=10, QI = Attribute 1-14)	68.08
Anonymized Data (RSA Alg.)	73.30
Anonymized Data (ElGamal Alg.)	73.59
Anonymized Data (DES Alg.)	73.22
Anonymized Data (AES Alg.)	73.57

Table 3 Classification performance based on geometric mean for all methods for Adult dataset [21].

The results in Tables 2 and 3 show that our proposed methodology only deteriorates the data to a negligible extent depending on the application; this is justifiable as there exists a cost for preserving the privacy of individuals. A comparison between the classification results of the anonymized data obtained by our proposed methodology and Mondrian multidimensional k-anonymity approach, in Table 3, indicates that our methodology outperforms theirs as the prediction results, with the same learning algorithm, are more accurate. Moreover, the results suggest that maintaining the utility of data is not dependent on a specific cryptographic algorithm.

Comparisons of two data utility measures for raw and anonymized data show that this methodology preserves the relations of values in the data table to a considerable extent. Therefore, analyses dependent on the relations of the data attributes to each other, and the labels are feasible and supported, e.g., learning tasks through machine learning algorithms. Such analyses are not dependent on the exact values in raw data since the anonymization changes the range of values for each attribute. The anonymized data is a matrix of numbers, likewise to the raw data, and it can be used the same way as the raw data. Moreover, regarding the privacy concerns described in the Problem Definition Section, if one manages to change the values in the raw data until the adversary cannot map it back to the original values, then the desired purpose is achieved. Using cryptographic algorithms for anonymization along with the fundamental property of these algorithms, i.e., mapping numbers by one-way injective functions, dismisses the described privacy concerns, in other words, matching data values from what the adversary has and what is published as anonymized data is not possible.

6 CONCLUSION

In this study, we investigated the approach of anonymizing the structured health data by utilizing cryptographic algorithms, which is, to the best of our knowledge, the first application of these algorithms in anonymization. Anonymization methods must fulfill two criteria, namely privacy preservation and data utility. We evaluated the presented methodology on the de facto benchmark dataset for anonymization. The results are promising and indicate that such an approach may be employed in real-world applications by the healthcare sector. However, similar to the majority of anonymization techniques, our proposed methodology impacts the quality of data mining results, even though we have shown that this degradation is less than the previous works in the data anonymization domain. This methodology is particularly practical for anonymizing the data for data mining applications. For future works, the applicability of this approach may be investigated for unstructured types of health data, e.g., physiological signals. Moreover, automatic de-identification of clinical notes and overcoming the particular challenges is another closely related research area that can be tied up with natural language processing [33, 34]. Further studies on the field mentioned above would be analogous to this study and worthwhile.

7 ACKNOWLEDGMENTS

This research is supported by INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology) Project [35], funded by the Norwegian Research Council (259293/o70). The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

8 REFERENCES

[1] Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), pp.557-570.

[2] Mohammed, N., Fung, B., Hung, P.C. and Lee, C.K., 2009, June. Anonymizing healthcare data: a case study on the blood transfusion service. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1285-1294). ACM.

[3] Health Data in an Open World. (2019, July 17). Retrieved from https://arxiv.org/ftp/arxiv/papers/1712/1712.05627.pdf

[4] Research reveals de-identified patient data can be re-identified. (2019, July 17). Retrieved from https://about.unimelb.edu.au/newsroom/news/2017/december/research-reveals-de-identified-patient-data-can-be-re-identified

[5] The simple process of re-identifying patients in public health records. (2019, July 17). Retrieved from https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records

[6] Understanding the maths is crucial for protecting privacy. (2019, July 17). Retrieved from https://pursuit.unimelb.edu.au/articles/understanding-the-maths-is-crucial-for-protecting-privacy

[7] Narayanan, A. and Shmatikov, V., 2008. Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset). University of Texas at Austin.

[8] Willenborg, L. and De Waal, T., 1996. Statistical disclosure control in practice. Springer Science & Business Media.

[9] Domingo-Ferrer, J. and Montes, F. eds., 2018. Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings. Springer.

[10] Fienberg, S.E. and van der Linden, W.J., Statistics for Social and Behavioral Sciences.

[11] International Organization for Standardization: ISO 25237:2017 Health informatics – Pseudonymization.

[12] Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M., 2006, April. l-diversity: Privacy beyond k-anonymity. In 22nd International Conference on Data Engineering (ICDE'06) (pp. 24-24). IEEE.

[13] Li, N., Li, T. and Venkatasubramanian, S., 2007, April. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd International Conference on Data Engineering (pp. 106-115). IEEE.

[14] Davis, J.S. and Osoba, O., 2019. Improving privacy preservation policy in the modern information age. Health and Technology, 9(1), pp.65-75.

[15] Bost, R., Popa, R.A., Tu, S. and Goldwasser, S., 2015, February. Machine learning classification over encrypted data. In NDSS (Vol. 4324, p. 4325).

[16] Graepel, T., Lauter, K. and Naehrig, M., 2012, November. ML confidential: Machine learning on encrypted data. In International Conference on Information Security and Cryptology (pp. 1-21). Springer, Berlin, Heidelberg.

[17] Bos, J.W., Lauter, K. and Naehrig, M., 2014. Private predictive analysis on encrypted medical data. Journal of biomedical informatics, 50, pp.234-243.

[18] Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M. and Wernsing, J., 2016, June. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International Conference on Machine Learning (pp. 201-210).

[19] Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1(1), pp.81-106.

[20] Shannon, C.E., 1948. A mathematical theory of communication. Bell system technical journal, 27(3), pp.379-423.

[21] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[22] Bayardo, R.J. and Agrawal, R., 2005, April. Data privacy through optimal k-anonymization. In 21st International conference on data engineering (ICDE'05) (pp. 217-228). IEEE.

[23] Fung, B.C., Wang, K. and Philip, S.Y., 2007. Anonymizing classification data for privacy preservation. IEEE transactions on knowledge and data engineering.

[24] Iyengar, V.S., 2002, July. Transforming data to satisfy privacy constraints. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 279-288). ACM.

[25] Mohammed, N., Fung, B., Wang, K. and Hung, P.C., 2009, March. Privacy-preserving data mashup. In Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (pp. 228-239). ACM.

[26] Wang, K. and Fung, B., 2006, August. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 414-423). ACM.

[27] Wang, K., Fung, B.C. and Philip, S.Y., 2007. Handicapping attacker's confidence: an alternative to k-anonymization. Knowledge and Information Systems.

[28] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., 2006, April. Mondrian multidimensional k-anonymity. In ICDE (Vol. 6, p. 25).

[29] Fleming, P.J. and Wallace, J.J., 1986. How not to lie with statistics: the correct way to summarize benchmark results. Communications of the ACM, 29(3), pp.218-221.

[30] Ho, T.K., 1995, August. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[31] MITCHELL, T. M. (2017). Machine learning. New York, McGraw Hill.

[32] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[33] Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L. and Solti, I., 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. Journal of the American Medical Informatics Association, 20(1).

[34] Liu, Z., Tang, B., Wang, X. and Chen, Q., 2017. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics, 75.

[35] INTROMAT (INtroducing personalized TReatment Of Mental health problems using Adaptive Technology). (2019, July 17). Retrieved from https://intromat.no/