RESEARCH ARTICLE OPEN ACCESS

Sensitivity Based Data Anonymization Information Gain

Esther Gachanga

School of Computing Sciences
Riara University Nairobi Kenya

ABSTRACT

Advances in hardware technology have led to an increase in the capability to store, record and share personal data about individuals. In data driven environments it is important to protect the privacy of individuals in published data. Data anonymization approaches have been applied to obfuscate the personally identifiable information in a dataset. However it's not clear what kind of anonymization should be used. This study proposes sensitivity based data anonymization with information gain. Information gain was used to establish redundant features. Experiments were conducted on real-life datasets. The results show that anonymizing redundant features more can reduce the amount of data distortion while enhancing privacy on published datasets.

Keywords: Privacy, Data anonymization, Sensitive Information, Data Publishing.

I. INTRODUCTION

Advances in hardware technology have led to an increase in the capability to store and record personal data about individuals. With increasing volumes of published data privacy preservation issues have emerged [1]. This has raised privacy concerns as personal data may be used for a variety of purposes. Some of the major concerns are pro filing, consumer mistrust and privacy breaches. In order to alleviate these concerns, a number of privacy preservation techniques including the *k*-anonymity, *t*-closeness, differential privacy and the *l*-diversity model have been applied. The k-anonymity model is the basis of more advanced models while still being useful as a stand-alone solution [2]. Data anonymization approaches have been applied to obfuscate the personally identifiable information in a dataset. These approaches include generalization, suppression, anatomization, permutation, or perturbation and are used together with Privacy preserving data publishing (PPDP) techniques to achieve a given privacy condition [3]. Data anonymization preserves privacy by eliminating identifiability from the dataset, i.e., the link between sensitive information and people [4]. However anonymization methods do not specify the approach to be used to generate an anonymized data set in a privacy model. The data publisher must select an approach that maximizes data utility, because satisfying the model already ensures privacy [5]. The rest of this paper is organized as follows; in section 2 related work was reviewed, section 3 presents the proposed approach. In section 4 the experiments are presented while section 5 presents the results and discussions and section 6 concludes the paper.

II. RELATED WORK

PPDP mainly studies how to anonymize data in such a way that after the data is published, individual's identity and sensitive information cannot be re-identified Fung*et.al*, (2010). PPDP focuses on publishing data collected from the record owners rather than data mining results. PPDP anonymizes the data by hiding the identities of record owners. The goal of PPDP is for a data custodian to release some views or statistical computations of the original private data, so that the released data remains practically useful while individual privacy for the data subjects is preserved. Privacy preservation is strongly connected with the idea of preventing information disclosure/leakage/ privacy breach [6].

In [7], several anonymity models have been proposed to protect individual's privacy for micro-data publishing, among them, the k-anonymity model [8], differential privacy [9], the *l*-diversity model [10], and t-closeness models [11].

The k-anonymity model originally proposed by [8] was the first model suggested for anonymizing data while maintaining privacy and has received much attention and has been widely used in practice [1]. The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. k-anonymity ensures that there are at least k people with the same quasi-identifier in an equivalence class such that the risk of identity disclosure is reduced to 1/k [12]. The primary goal of k-anonymity is to protect the privacy of the individuals to whom the data pertains. However subject to this constraint it's important that the data remains as useful as possible [13].

A study by [9] introduced Differential Privacy (DP). DP is a mathematical framework that is widely accepted for protecting data privacy by adding noise to data [14]. DP guarantees that the distribution of query results changes only slightly due to the modification of any one tuple in the database. This allows protection, even against powerful adversaries, who may know the entire database except one tuple. To provide this guarantee, differential privacy mechanisms assume independence of tuples in the database [15].

In [16], a formal definition of differential privacy is given as: a randomized algorithm A, satisfies ε- differential privacy if (eq1);

$$P(A(D1) \in S) \leq e\varepsilon\, A(D_2)) \in S \qquad [1]$$

for any set S and any pairs of databases $D_{1,}$ D_2 where D_1 can be obtained from D_2 by either adding or removing one tuple or by changing the value of exactly one tuple. The degree of privacy protection does not depend on the size of the quasi identifying (QI) group, but instead is determined by the number of distinct sensitive values in each QI group This observation leads to ℓ -diversity which guarantees stronger privacy than k-anonymity [17].

The k–anonymity model suffers from homogeneity and background knowledge attacks. To address the shortcomings of the k–anonymity model, the *l*-diversity model [10] and t-closeness model [11] were introduced.

The *l*-diversity model requires each equivalence group of released table to contain at least *l*-well represented records. It means that every sensitive attribute in each equivalence class should have at least *l* different values. By increasing the diversity of sensitive attributes in every equivalence class, it enhances the difficulty to link a sensitive value to an individual. [10] gives three interpretations of the term "well represented"; first distinct ℓ-Diversity ensures that there exist at least ℓ-distinct sensitive values in each equivalence class, second Entropy ℓ-diversity. The entropy of an equivalence class E is defined as (eq2);

$$Entropy(E) = -\sum_{s \in S} P(logP(E,s) \qquad [2]$$

Where s, is the sensitive attribute domain and P (E,s) is the fraction of records in E with sensitive value S, and lastly, recursive (*c*, *l*)- diversity, which ensures that the most frequent values does not appear too frequently.

The *t*-closeness principle requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table as much as is possible (i.e., the distance between the two distributions should be no more than a threshold ***t***) [11]. Ninghui et al (2007), performed an experiment on *t*-closeness. In the experiment the distance in the distributions of the attributes in a class is measured using the Earth Movers Distance (EMD). Given two distributions P and Q; P= ($p_{1,}$ $p_{2,\ldots,}P_m$) and Q=(q_1, $q_{1,\ldots,}$ q_m) the variational distance in the distribution can be defined as (eq3);

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2} |pi - qi| \qquad [3]$$

Where D is the distance and P,Q are the distributions.

2.1 DATA ANONYMIZATION

Anonymization techniques have been widely used to provide a balance between the beneficial uses of data and privacy [6]. In anonymization methods, it is assumed that data publisher has a table T that includes four subsets of attributes: 1) explicit identifiers (I) containing information that explicitly identifies a record owner and are removed from the released data such as name, social security number and cell-phone number, 2) quasi-identifiers (QID) containing information that could potentially identify a record owner and typically transformed in the released data such as date-of-birth, gender and ZIP code. 3) Sensitive attributes (S) containing sensitive information about data owner such as salary or disease which should be protected, and 4) non-sensitive attributes, which does not fall into the previous three categories and can be published as it is when needed [18].

Anonymity is a safe and effective method for data privacy protection, which can effectively balance the relationship between the efficiency and the security of the data. The basic idea of anonymization is that from a transformed table, the attacker cannot easily analyze the sensitive attribute of a tuple, and therefore cannot identify a specific individual's sensitive information Wang et.al (2016).

To preserve sensitive information, anonymization techniques need to be applied to a published micro-data table. The anonymization approaches tries to protect the identity and the sensitive information of a user. The sensitive data must be preserved for data analysis. In the privacy-preserving data publishing setting, a data publisher releases the data to the public, and it is open to everyone. An attacker also receives the published data, and could use some background knowledge to identify a person by linking with some publicly available data sources. Hence, the demand for anonymity is necessarily present in the privacy-preserving data publishing context [19]. An anonymized view of data protects an individual's data/record from unwanted disclosure. Data anonymization preserves privacy by eliminating identifiability from the dataset, i.e., the link between sensitive information and people [4] .

2.2 SENSITIVITY OF DATA

In the work of [20], it is well explained that the sensitivity of data tends to be connected with the potential harm of any confidentiality breach and that for a disclosure to be meaningful something has to be learned. These works further explain that personal data becomes sensitive according to its context and that the sensitivity of data can be reduced by removing the sensitive attributes. While there are several data release options available, the one you choose depends on the data you plan to release, the sensitivity of the data, and the proposed usage of the data. [21], studied the value of individual using the terminology the insensitive value model, where users do not care about leaking their privacy valuations and a sensitive model where users may care about leaking their privacy.

2.3 INFORMATION LOSS

The process of privacy preservation and anonymization causes information loss, which can be considered as a loss of utility. To produce useful output the data publisher has to balance the competing requirements of sufficient privacy protection and maximum possible utility [22]. In [23] anonymity is an optimization problem accompanied by information loss based on the actual distortion of data based on a taxonomy tree. Information loss is defined as; given a micro-data table $T = \{ t_1, \ldots \ldots \ldots . t_n \}$ where $Q_1 = \{ Q_1, \ldots \ldots \ldots \ldots Q_m$ is the amount of data distortion that would occur by generalizing T, and is denoted by the equation 4;

$Distortion\ (T) = \sum_{i=1}^{n} \sum_{j=1}^{m} |h_{ij}|$ ------- [4]

If the value in Q_j of t_i has been generalized $|h_{ij}|$ levels up in the taxonomy, the height of the value generalized is equal to $|h_{ij}|$.

[5] explains that the utility of an anonymized dataset is evaluated in terms of information loss, that is, the discrepancies between the original and the anonymized data set.

2.4 INFORMATION GAIN

Feature selection involves identifying and selecting a subset of the most useful features that produces compatible results as the original entire set of features [24]. Feature selection with information gain can be used to identify and remove the redundant features [25]. Information gain has been used as a measure of feature relevancy for filter based feature selection and evaluates the worth of an attribute by measuring the information gain with respect to a class [26]. Given two attributes X and Y that belong to dataset D, the Information Gain or mutual information between attributes can be calculated using conditional entropy as eq5;

$IG[X;Y] = H(X) - H\left(\frac{X}{Y}\right) + H(Y) - H(\frac{Y}{X})$ -- [5]

The above equation can be re-written as eq6;

$H(X) + H(Y) - H(X, Y)$ ------------- [6]

Where $IG\ [X;Y]$ measures the degree of uncertainty about x due to the knowledge of y. $IG\ [X;Y]$ also measures the two way association between the attributes x and y. Further [27] define IG as a measure reflecting additional information about a variable C provided by A that represents the amount by which the entropy of C decreases. This measure, is an indicator of the dependency between A and C, known as IG eq 7;

IG (A) = E(C) − E (C|A) = E (A) − E (A|C) ----- [7]

2.5 RISK MODEL

According to [28], risk models are used to estimate re-identification risks, which are an inherent aspect of many privacy models. [28] further explains that, the attacker models in data anonymization; the prosecutor, the journalist and the marketer model.

In the *prosecutor* model the attacker targets a specific individual and it is assumed that the attacker already knows that data about the individual is contained in the dataset. In the *journalist* model the attacker targets a specific individual but it is not expected that the attacker possesses background knowledge about membership. In the *marketer* model the attacker does not target a specific individual but aims at re-identifying a high number of individuals. An attack can therefore only be considered successful if a larger fraction of the records could be re-identified. In [29] marketer risk refers to the proportion of records that are correctly re-identified. A measure of this risk is computed only from the disclosed database. Under this type of attacks, the risk metric can be quantified as the largest probability that an individual in a given dataset is correctly re-identified. The risk is measured by the records at risk, the highest risk and the success rate for a given dataset [30].

III. PROPOSED APPROACH.

In this section, the study presents sensitivity based data anonymization with information gain. The approach first categorizes the different attributes in a dataset into quasi identifiers and sensitive attributes. The distinct attributes within the sensitive attribute are identified and grouped into two i.e. highly sensitive and less sensitive. Third the information gain for the quasi identifier attributes is established. Only Quasi identifiers with very low information gain for tuples with less sensitive attributes are generalized, while all the quasi identifiers for tuples with highly sensitive attributes are all generalized. Finally the anonymized dataset is evaluated.

IV. EXPERIMENT

This research adopted the adult dataset from UCI (Dheeru et al., 2017). It is a real-world dataset and has already been utilized for benchmarking previous work on k-anonymity [1]. The Dataset is conceptually organized as a table of rows (or records) and columns (or fields). Each row is termed a tuple. Tuples within a table are not necessarily unique. Each column is called an attribute and denotes a set of possible values within its domain. Data cleaning was done by removing tuples/records with missing values similar to [31], [32] [33] and [3]. After removing tuples with missing values, a total of 30162 tuples /records were left and these were utilized for the experiments. Eight attributes among them; Age, Work-class, Education, Marital-status, Occupation, Relationship, Race and Sex were used. Occupation was used as the sensitive attribute while the other seven attributes were the quasi identifier attributes.

Next an identification of distinct values of the sensitive attribute occupation was done. 14 attributes were identified. These values were, Craft-repair, Prof-specialty, Machine-op-inspct, Farming-fishing, sales, Other-service, Exec-managerial, Adm-clerical, Transport-moving, Protective-service, Tech-support, Handlers-cleaners, Priv-house-service and Armed-Forces.

The research considered the values Protective-service, Farming-fishing, Priv-house-service and Armed-Forces to be sensitive attributes. Tuples containing this sensitive attributes were grouped together into a table T1. The rest of the tuples were grouped together into a table T2. All the QIDs for the table T1 were anonymized so as to give table T1*. Information gain was used to determine the QIDs that should be anonymized in Table T2. Information gain for the eight attributes using feature selection and ranking with occupation as the target attribute was established (Table 1).

Table 1: Information Score for the different attributes	
Attribute	Score with Ranking
Education	0.3364
Work Class	0.1682
Sex	0.1496
Relationship	0.1217
Marital Status	0.077
Age	0.066
Race	0.0193

The attributes with an information gain of less than 0.14 in table T2, i.e race, age and marital status were anonymized so as to give T2*. Anonymization for both table T1 and T2 was done on the k-anonymity model with the value of k being 5 and the value of l as 3. The two anonymized tables were the merged together. The research applied the k-anonymity and the L diversity model. Local transformation with iterations at a 100 and a suppression rate of 4.9% were employed. The sensitive attribute occupation was the target attribute.

V. RESULTS & DISCUSSIONS

In this section, the study presents the results. The performance of different classifiers build from the anonymized dataset is compared. Three classifiers; Random Forest, Logistic regression and, Naïve bayes were used. The performance of these classifiers was compared with the performance of classifiers built from the initial dataset which had not been anonymized. The results are as presented in Table 2.

Table 2: Performance of Classifiers			
	Random Forest	Logistic Regression	Naïve Bayes
Initial Dataset	30.076	32.272	30.654
Anonymized Initial Dataset	26.904	29.8	28.141
T* k=5 L=3	19.541	19.054	19.946
T*(T1 k=5 L=3, T2 K=5 L=5)	27.846	28.695	26.55

The results in Table 2 show that for the three classifiers built from the initial dataset had the highest utility. When the initial dataset was anonymized the accuracy of the classifiers dropped to, Random Forest 26.904, Logistic regression 29.8 and, Naïve Bayes 28.141. Comparing the results of the approach prosed in the study for dataset T* with k=5 and L=3, the performance of the classifiers was as follows; Random Forest 19.541, Logistic regression 19.054 and Naïve Bayes 19.946. When the value of k=5 and L=5, the performance of the classifiers was as follows; Random Forest 27.846, Logistic regression 28.695 and Naïve Bayes 26.55. It was noted that the classifiers for Table T* k=5 L=3 were less accurate compared to that of the initial anonymized dataset with the same parameters. However when the value of L=5 for table T2 in Table T*, the performance of classifiers improved by; Random Forest 8.035, Logistic regression 9.641 and Naïve Bayes 6.604. The Random Forest classifier performed better than the anonymized initial dataset.

Attack model was described in section 2.5. The results for the prosecutor and journalist model were presented in table 3. For the initial dataset the records at risk were 41.207% while the highest risk 100% and the Success rate was 32.249%. When the initial dataset was anonymized the risk reduced as follows; for the records at risk to 0% while the highest risk to 20% and for the Success rate to7.533%. For the dataset anonymized using the proposed approach, the risk reduced to; for the records at risk to 0% while the highest risk to 20% and for the Success rate to 0.316% for Table T* k=5 L=3. When the value of L=5 for Table T2 in T*, the records at risk was 0% while the highest risk was 16.667% and for the Success rate was 0.316%.

Table 3: Prosecutor and Journalist model

	Records at Risk %	Highest Risk %	Success Rate %
Initial dataset	41.207	100	32.249
Anonymized Initial Dataset	0	20	7.533
T* k=5 L=3	0	20	0.316
T*(T1 k=5 L=3, T2 K=5 L=5)	0	16.667	0.316

From Table 3, with the proposed approach, the records at risk, the risk remained the same, while for the highest risk was the same when the value of the parameter L=3 and dropped by 3.333% when the value of L=5. The success rate for the proposed approach was significantly reduced from 7.533% to 0.316% both when the value of L=3 and L=5.

Table 4: The Marketer attack model

	Success Rate
Initial dataset	32.249
Anonymized Initial Dataset	7.533
T* k=5 L=3	0.316
T*(T1 k=5 L=3, T2 K=5 L=5)	0.316

Table 4 presents the results for the marketer attack model. The risk was highest for the initial dataset at 32.249%, while that for the anonymized initial dataset the risk was 7.533%. The study observed that the marketer risk was lowest while implementing the proposed approach at 0.316%.

From the experiments it seems that the approach adopted in the study led to poor performance of the classifiers compared to the existing approaches in all instances except with the Naïve Bayes classifier when the value of the parameter L is increased. Nevertheless, the attack models performed better.

VI. CONCLUSION

This paper looked at sensitivity based information data anonymization with information gain and demonstrated that data anonymization is one of the most commonly used approaches by the data publishers to achieve data privacy. Information gain was used to determine attributes with redundant features. A major setback in privacy-preserving model is the privacy utility trade off as the amount of data anonymization performed on a given dataset significantly influences both the quality and the utility of data. Data publishers must work towards developing models and algorithms that maintain a high level utility of data while preserving privacy.

REFERENCES

[1] O. Temuujin and J. Ahn, “Efficient L-Diversity Algorithm for Preserving Privacy of Dynamically Published Datasets,” *IEEE Access*, vol. 7, pp. 122878–122888, 2019, doi: 10.1109/ACCESS.2019.2936301.

[2] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, “A flexible approach to distributed data anonymization,” *J. Biomed. Inform.*, vol. 50, pp. 62–76, 2014, doi: 10.1016/j.jbi.2013.12.002.

[3] M. H. A. Ibrahim, K. Zhou, and J. Ren, “Privacy Characterization and Quantification in Data Publishing,” *IEEE Trans. Knowl. Data Eng.*, vol. 4347, no. c, pp. 1–14, 2018, doi: 10.1109/TKDE.2018.2797092.

[4] H. Lee, S. Kim, J. W. Kim, and Y. D. Chung, “Utility-preserving anonymization for health data publishing,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, pp. 1–12, 2017, doi: 10.1186/s12911-017-0499-0.

[5] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, “T-closeness through microaggregation: Strict privacy with enhanced utility preservation,” *2016 IEEE 32nd Int. Conf. Data Eng. ICDE 2016*, vol. 27, no. 11, pp. 1464–1465, 2016, doi: 10.1109/ICDE.2016.7498376.

[6] T. Basso, R. Matsunaga, R. Moraes, and N. Antunes, “Challenges on anonymity, privacy, and big data,” in *Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016*, 2016, pp. 164–171, doi: 10.1109/LADC.2016.34.

[7] G. Yang, J. Li, S. Zhang, and L. Yu, “An enhanced l-diversity privacy preservation,” in *Proceedings - 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2013*, 2013, no. 61170060, pp. 1115–1120, doi: 10.1109/FSKD.2013.6816364.

[8] P. Samarati and L. Sweeney, “Protecting Privacy when Disclosing Information: k-Anonymity and

its Enforcement Through Generalization and Suppresion.," *Proc IEEE Symp. Res. Secur. Priv.*, pp. 384–393, 1998, doi: http://dx.doi.org/10.1145/1150402.1150499.

[9] C. Dwork, "Differential Privacy," *Proc. Int. Colloq. Autom. Lang. Program. Part II*, pp. 1–12, 2006, doi: 10.1007/11787006.

[10] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "ℓ-Diversity: Privacy Beyond k-Anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3-es, 2007, doi: 10.1145/1217299.1217302.

[11] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and ℓ-diversity," in *Proceedings - International Conference on Data Engineering*, 2007, no. 2, pp. 106–115, doi: 10.1109/ICDE.2007.367856.

[12] L. Sweeney, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.

[13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," *Proc. - Int. Conf. Data Eng.*, vol. 2006, p. 25, 2006, doi: 10.1109/ICDE.2006.101.

[14] A. M. Olawoyin, C. K. Leung, and A. Cuzzocrea, "Preserving Privacy of Temporal Big Data," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 4042–4051, 2020, doi: 10.1109/BigData50022.2020.9378040.

[15] C. Liu, S. Chakraborty, and P. Mittal, "Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples," *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2016, doi: 10.14722/ndss.2016.23279.

[16] Y. Yang, Z. Zhang, G. Miklau, M. Winslett, and X. Xiao, "Differential privacy in data publication and analysis," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of data (SIGMOD)*, 2012, pp. 601–605, doi: 10.1145/2213836.2213910.

[17] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," *Proc. 32nd Int. Conf. Very Large Database, ACM*, pp. 150, 139, 2006, [Online]. Available: http://portal.acm.org/citation.cfm?id=1164127.1164141.

[18] P. Canbay and H. Sever, "The Effect of Clustering on Data Privacy," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 277–282, doi: 10.1109/ICMLA.2015.198.

[19] S. M. T. Hasan, Q. Jiang, and C. Li, "An effective grouping method for privacy-preserving bike sharing data publishing," *Futur. Internet*, vol. 9, no. 4, pp. 1–18, 2017, doi: 10.3390/fi9040065.

[20] E. M. K. O. and C. T. Mark Elliot, "The Anonymisation Decision-Making Framework," *Univ. Manchester*, 2016.

[21] A. Ghosh and A. Roth, "Selling Privacy at Auction Categories and Subject Descriptors," *ACM*, 2011.

[22] T. S. Gal, T. C. Tucker, A. Gangopadhyay, and Z. Chen, "A data recipient centered de-identification method to retain statistical attributes," *J. Biomed. Inform.*, vol. 50, pp. 32–45, 2014, doi: 10.1016/j.jbi.2014.01.001.

[23] Y. Wu, X. Ruan, S. Liao, and X. Wang, "P-cover k-anonymity model for protecting multiple sensitive attributes," *ICCSE 2010 - 5th Int. Conf. Comput. Sci. Educ. Final Progr. B. Abstr.*, pp. 179–183, 2010, doi: 10.1109/ICCSE.2010.5593663.

[24] G. Saranya and A. Pravin, "An Efficient Feature Selection Approach using Sensitivity Analysis for Machine Learning based Heart Desease Classification," *Proc. - 2021 IEEE 10th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2021*, pp. 539–542, 2021, doi: 10.1109/CSNT51715.2021.9509673.

[25] A. Nugroho, A. Z. Fanani, and G. F. Shidik, "Evaluation of Feature Selection Using Wrapper for Numeric Dataset with Random Forest Algorithm," *Proc. - 2021 Int. Semin. Appl. Technol. Inf. Commun. IT Oppor. Creat. Digit. Innov. Commun. within Glob. Pandemic, iSemantic 2021*, pp. 179–183, 2021, doi: 10.1109/iSemantic52711.2021.9573249.

[26] N. A. N. Shaltout, M. El-Hefnawi, A. Rafea, A. Moustafa, and M. El-Hefnawi, "Information gain as a feature selection method for the efficient classification of Influenza-A based on viral hosts," *Proc. World Congr. Eng.*, vol. 1, pp. 625–631, 2014, [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84907414411&partnerID=tZOtx3y1.

[27] F. He, H. Yang, Y. Miao, and R. Louis, "A Hybrid Feature Selection Method Based on Genetic Algorithm and Information Gain," pp. 320–323, 2016.

[28] F. Prasser and F. Kohlmayer, "Medical Data Privacy Handbook," *Springer Int. Publ. Switz. 2015*, 2015, doi: 10.1007/978-3-319-23633-9.

[29] F. K. Dankar and K. El Emam, "A method for evaluating marketer re-identification risk," *Proc. 1st Int. Work. Data Semant. - DataSem '10*, p. 1, 2010, doi: 10.1145/1754239.1754271.

[30] F. Prasser, J. Eicher, R. Bild, H. Spengler, and K.

A. Kuhn, “A Tool for Optimizing De-identified Health Data for Use in Statistical Classification,” *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2017-June, 2017, doi: 10.1109/CBMS.2017.105.

[31] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, “Secure Anonymization for Incremental Datasets,” 2006, pp. 48–63.

[32] A. S. M. T. Hasan and Q. Jiang, “A General Framework for Privacy Preserving Sequential Data Publishing,” *2017 31st Int. Conf. Adv. Inf. Netw. Appl. Work.*, pp. 519–524, 2017, doi: 10.1109/WAINA.2017.18.

[33] S. A. Abdelhameed, “Enhanced Additive Noise Approach For Privacy- Preserving Tabular Data Publishing,” no. Icicis, pp. 284–291, 2017.