2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Diversity-Aware Anonymization for Structured Health Data

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

Abstract— **Patients' health data are captured by local hospital facilities, which has the potential for data analysis. However, due to privacy and legal concerns, local hospital facilities are unable to share the data with others which makes it difficult to apply data analysis and machine learning techniques over the health data. Analysis of such data across hospitals can provide valuable information to health professionals. Anonymization methods offer privacy-preserving solutions for sharing data for analysis purposes. In this paper, we propose a novel method for anonymizing and sharing data that addresses the record-linkage and attribute-linkage attack models. Our proposed method achieves anonymity by formulating and solving this problem as a constrained optimization problem which is based on the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models. The proposed method has been evaluated with respect to the utility and privacy of data after anonymization in comparison to the original data.**

I. Introduction

Patients' data is private and may contain sensitive information, e.g., information about a health condition. Such data may not be shared with other parties in their raw format due to privacy and legal concerns [1], [2]. However, such data may be required for analysis purposes to provide value to medical experts and utilized for analysis by adopting privacy-preserving data mining or privacy-preserving data sharing approaches depending on the particular application and scenario.

Privacy-preserving data mining techniques perform the analysis without direct access to the data. Several approaches adopt homomorphic encryption techniques for learning tasks [3], [4]. However, such methods suffer from communication and computation overhead and are not always practical [5]. Several state-of-the-art techniques modify and adapt algorithms for learning from distributed data without sharing data and sacrificing privacy [6], [7], [8], [9], [10]. Nevertheless, each algorithm should be extended to support privacy-preserving distributed learning. Moreover, learning a classification model from data is not the only objective in particular scenarios, and a version of data may be required to be published, e.g., for medical expert inspection and visualization.

Privacy-preserving data sharing techniques share an altered version of data for analysis. Several studies add noise to data and perturb it before sharing [11], [12], [13]. However, the utility of data will be negatively affected by the perturbation of data. On the other hand, privacy will not be preserved if the noise added is not sufficient. Moreover, noise removal approaches pose a threat to the privacy of such methods [14], [15]. Several studies adopt neural networks and generative adversarial networks (GAN) [16] for altering the data before sharing [2], [17], [18], [19]. Such approaches mainly focus on particular time-series data and data in wearable devices' applications [20], [21], [22], [23].

Anonymization methods also alter the data to avoid identifying data subjects in such datasets [24]. Previous studies proposed several privacy models for anonymization, e.g., *k*-anonymization [25], *l*-diversity [26], *t*-closeness [27], *LKC*-privacy [28]. The data holder selects a model based on the scenario, utility, and privacy requirements. Several methods have been proposed to comply with such privacy models and avoid the associated attacks, i.e., record-linkage and attribute-linkage attacks, e.g., using genetic algorithms to kd-trees algorithms for generalization and achieving anonymity [29], [30], [31], [32], [33].

In particular, [34] proposes the utilization of Mixed-Integer Programming for achieving *k*-anonymity. Similarly, [35] formulates the anonymization problem in a Mixed-Integer Linear Programming (MILP) framework and achieves *k*-anonymity based on optimization. This approach uses generalization for anonymization and optimizes the lower and upper bound for each value of quasi-identifiers, which are the attributes that the adversary may have information about for identification. However, these anonymization methods [34], [35] merely consider *k*-anonymity and does not prevent the attribute-linkage attack, which is the issue addressed by the *l*-diversity and *t*-closeness privacy models. Therefore, the joint consideration of the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in such frameworks have not been considered to date.

In this paper, we propose a method to anonymize data to ensure that each record is indistinguishable from, at least, *k-1* other records in the shared data while taking the diversity and frequency of values in the sensitive attribute into consideration. In other words, we propose a method for anonymization of data considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in a unified framework. We formulate the anonymization problem in a constrained optimization framework as a clustering problem, where the diversity and frequency of sensitive values are captured and enforced by constraints. We refer to our proposed method as diversity-aware anonymization, where diversity captures both the diversity concept in the *l*-diversity privacy model and the frequency and distribution of sensitive values in the *t*-closeness privacy model. The experimental results show the preservation of utility of data for classification tasks and the privacy properties noted in the discussed models.

The rest of this paper is organized as follows: Section II covers the background with respect to *k*-anonymity,

[1] Western Norway University of Applied Sciences, Bergen, Norway `firstname.lastname@hvl.no`

[2] University of Bergen, Bergen, Norway

[3] University of Oslo, Oslo, Norway

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

l-diversity, *t*-closeness, and their corresponding attack models. We formulate our proposed anonymization method in the constrained optimization framework in Section III. Section IV provides the experimental results for evaluation of our method. Section V concludes our paper.

II. Background

In this section, we briefly review the record-linkage and attribute-linkage attack models. In addition, we discuss three popular privacy models addressing such attacks, i.e., *k*-anonymity, *l*-diversity, and *t*-closeness.

In the record-linkage and attribute-linkage attack models, we suppose that a version of data after removing the identifier attributes of patients, e.g., name and address, is shared with a data recipient. At the same time, the adversary has access to the data shared with the data recipient. This data contains several attributes through which a patient (record owner) can be identified, i.e., quasi-identifiers, and it is assumed that the adversary has the exact value of these attributes for the victim patient. Finally, there is a sensitive attribute in the data, e.g., family history for a health pathology, that the adversary is interested in knowing about.

To explain this attack models, we use Tables Ia and Ib as an example. The 2nd-4th columns are considered as quasi-identifiers and refer to age, the number of children, and the smoking state of the patient (*Yes*/*No*). The 5th column is a sensitive attribute capturing the state of the HIV disease for the patient (*Positive*/*Negative*). Table Ia represents shared data after removing the identifier features. Suppose that Table Ia is shared with the data recipient. If the adversary knows that the victim is 37 years old, has two children, and smokes, he/she can easily match his/her information to one of the records (record one in Table Ia) and identify that the victim is diagnosed with HIV. The record-linkage attack occurs by matching the adversary's information (quasi-identifiers) with published data for identifying the patient's (record owner) sensitive information [36].

The *k*-anonymity privacy model was proposed to address the record-linkage attack model. A dataset is *k*-anonymous when the values of quasi-identifiers for each record are the same as the values for at least *k-1* other records in the data. In this way, the adversary can only match his/her information with at least *k* records. Table Ib shows a 3-anonymous version of the same data in Table Ia. For instance, in our example in Table Ib, if the adversary knows that the victim is 37, has two children, and smokes, he/she can merely match his/her information with a qid group containing the records of three patients, records *1-3*.

While the *k*-anonymity model guarantees that a patient is only matched with a qid group, however, this model does not guarantee the protection of patients' privacy against attribute-linkage attacks. That is, *k*-anonymity does not consider the diversity of values for the sensitive attribute in each qid group. In this example, in the first qid group, all the values for the sensitive attribute are *Positive*. Therefore, in the first qid group, the adversary can infer that the victim patient is diagnosed with HIV by matching quasi-identifiers' information. The attribute-linkage attack model occurs in situations where the diversity of values for the sensitive attribute is low. As a result, the adversary may infer the sensitive attribute with high confidence.

To address the attribute-linkage attack, the *l*-diversity model proposes that every qid group should have a least *l* distinct values for the sensitive attribute. For instance, in Table Ib, if the adversary matches his/her information with the third qid group, he/she can not identify that the patient was diagnosed with HIV for sure because both *Negative* and *Positive* values are in that qid group. However, this does not consider the confidence of the adversary's inference properly. For example, if we have both *Negative* and *Positive* values in all qid groups, we have 2-divers data, but if the proportion of *Positive* values in one qid group is high, the adversary can infer that the patient is diagnosed with HIV with high confidence. The entropy *l*-diversity and recursive *(c,l)*-diversity are proposed to address such issues [26].

Entropy *l*-diversity is one of the existing privacy models to address the distribution of values in the sensitive attribute. A data table meeting the following condition for each qid group is entropy *l*-diverse:

$$-\sum_{s\in S} P(qid, s)\log(P(qid, s)) \geq \log(l), \tag{1}$$

where *S* is the set of values for sensitive attribute, and *P(qid, s)* is the probability/proportion of value s for the sensitive attribute in the qid group.

The entropy *l*-diversity still has several limitations. For instance, if the entropy of values for the sensitive attribute in qid groups is high, the *l* will be high. The entropy is highest when the distribution of values is a uniform distribution. Nevertheless, we prefer the minimum probability for the sensitive value (*Positive* in our example) in the qid group. In our example, we favor as few *Positive*s in the qid groups as possible to lower the confidence of inferring HIV positive for the victim patient. Still, entropy *l*-diversity encourages an equal number of *Positive*s and *Negative*s in the qid groups.

Recursive *(c,l)*-diversity controls the frequency of values for the sensitive attribute in the qid group. In this model, c is a constant greater than zero, $c > 0$. The values for the sensitive attribute *S* are: $s_1, s_2, \ldots, s_m$. The number of occurrence for each value (for the sensitive attribute) in the qid group are: $n_1, n_2, \ldots, n_m$. The number of occurrence for values sorted in a decreasing order are: $r_1, r_2, \ldots, r_m$. If a data table meets $r_1 \leq c\sum_{i=l}^{m} r_i$ for each qid group, then the data is recursive *(c,l)*-diverse.

The recursive *(c,l)*-diversity can relax the restrictiveness compared to entropy *l*-diversity. When we have a larger c, we can have a larger l. Therefore, we can relax the restrictiveness by increasing c. This privacy model avoids having a high frequency of highly repeated values (in the dataset for sensitive value) in the qid group. It also forces the less frequent values (in the dataset for sensitive value) to be more frequent in the qid group. However, this may not be desirable in certain scenarios. Many healthcare datasets have sensitive attributes with highly imbalanced values. For

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	37	2	Yes	Positive
2	36	0	Yes	Positive
3	40	0	Yes	Positive
4	35	3	Yes	Negative
5	32	1	Yes	Negative
6	34	1	Yes	Negative
7	30	2	No	Positive
8	34	2	No	Negative
9	28	1	No	Negative
10	31	1	No	Negative

(a) Original data

Index	Quasi Identifier			Sensitive
	Age	Number of Children	Smoke	HIV
1	[36-40]	[0-2]	Yes	Positive
2	[36-40]	[0-2]	Yes	Positive
3	[36-40]	[0-2]	Yes	Positive
4	[32-35]	[1-3]	Yes	Negative
5	[32-35]	[1-3]	Yes	Negative
6	[32-35]	[1-3]	Yes	Negative
7	[28-34]	[1-2]	No	Positive
8	[28-34]	[1-2]	No	Negative
9	[28-34]	[1-2]	No	Negative
10	[28-34]	[1-2]	No	Negative

(b) 3-anonymous data

TABLE I: Patient data tables in original and 3-anonymous formats

instance, in a table of data with 1000 records, we may have merely 20 patients diagnosed with HIV. In our example, by increasing the frequency of a sensitive value (with low frequency in the dataset) in a qid group, the adversary can more confidently infer that the patient is diagnosed with HIV.

The *t*-closeness privacy model proposes having a more similar distribution of values in the sensitive attribute among the qid groups and the whole dataset. In the *t*-closeness model, the maximum distance between these two distributions may not be greater than the threshold *t*. For measuring the distance between probabilistic distributions, one possible metric is as follows:

$$D[P,Q] = \sum_{i=1}^{m} |p_i - q_i|, \tag{2}$$

where *m* is the number of values for the sensitive attribute. $P = \{p_1, p_2, \ldots, p_m\}$ and $Q = \{q_1, q_2, \ldots, q_m\}$ are the distributions of sensitive attribute in the entire dataset and in a particular qid group, respectively. This distance metric (variational distance) does not consider the semantic distance between values. In scenarios where the semantic distance of values is important, we may use other distance measures.

In this paper, we propose a method for anonymization of data by jointly considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models in a unified framework.

III. Approach

In this section, we describe our method for addressing the attack models discussed in Section II. In our method, we consider the indistinguishability of samples in a qid group, proposed in *k*-anonymity, diversity of values in sensitive attributes in qid group, discussed in *l*-diversity, and frequency of sensitive values in qid group in *t*-closeness.

In this method, we suppose that the values for the sensitive attribute are either sensitive or not. In our example, the *Positive* value shows that the patient (record owner) is diagnosed with HIV and is sensitive, while the value *Negative* if known to the adversary causes no consequence to the patient. Therefore, we consider a binary state for the values in the sensitive attribute and distribute them in the qid groups evenly.

Our method clusters the points in the space of quasi-identifiers and shares the center of each cluster (qid group) as the quasi-identifiers' values for each qid group. Each cluster contains *k* samples and is clustered based on the distance of instances to the cluster center and the number of samples with sensitive values in each cluster.

We adopt the constrained optimization framework to solve the described clustering problem. The classical clustering techniques do not fulfill our requirements. First, we need to introduce the constraints to have *k* samples in each cluster to ensure the indistinguishability property of the *k*-anonymity model. Second, we need to introduce a constraint for distributing instances with sensitive values evenly among qid groups (clusters) to ensure diversity in the *l*-diversity and *t*-closeness models.

The described anonymization problem is formulated in the Mixed-Integer Linear Programming (MILP) framework, as follows:

$$\min_{B,C} \quad \sum_{i=0}^{n_C} \sum_{j=0}^{n_S} |B_{ij} \cdot (X_j - Center_i)| \tag{3}$$

$$\text{s.t.} \quad \sum_{i=0}^{n_C} B_{ij} = 1, \quad \forall j \in \{0, \ldots, n_S\} \tag{4}$$

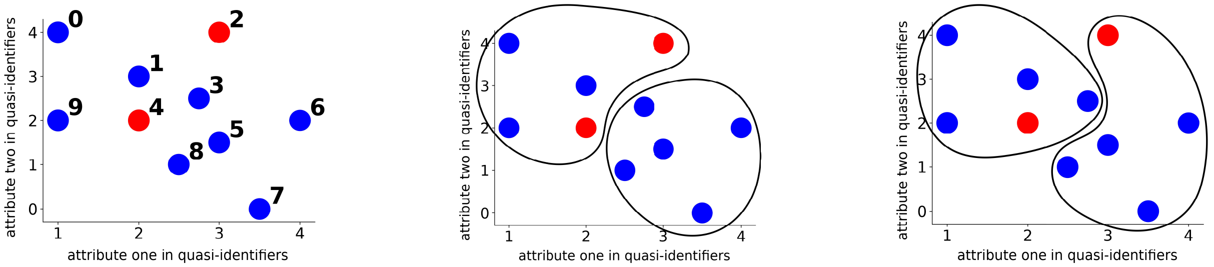
$$\sum_{j=0}^{n_S} B_{ij} = \frac{n_S}{n_C} = k, \quad \forall i \in \{0, \ldots, n_C\} \tag{5}$$

$$\frac{\left(\sum_{j=0}^{n_S} B_{ij} \cdot X_j\right)}{k} = C_i, \quad \forall i \in \{0, \ldots, n_C\} \tag{6}$$

$$\sum_{j=0}^{n_S} B_{ij} \cdot S_j \le \alpha \cdot \frac{\sum_{j=0}^{n_S} S_j}{n_C}, \quad \forall i \in \{0, \ldots, n_C\}, \tag{7}$$

where n_C is the number of clusters (qid groups), and n_S is the number of samples to be anonymized. X_j is the vector of quasi-identifiers' values for sample *j*. B_{ij} indicates if sample *j* belongs to cluster (qid group) *i* and it is a Boolean optimization variable. *Center*$_i$ is the *i*-th cluster center calculated by k-means algorithm to be used as an initial solution in our method to reduce the complexity of our optimization problem.

The parameter *k* is the number of samples in each cluster and is equal to $\frac{n_S}{n_C}$. C_i is the center of cluster *i* which will be optimized during solving this problem. The values of



(a) Data samples before optimization to find qid groups

(b) Optimization without considering the diversity constraint

(c) Optimization with considering the diversity constraint

Fig. 1: Illustrative example for our anonymization method

vector C_i will be shared with data recipients, i.e., instead of raw quasi-identifiers' values for i-th qid group. S_j is a Boolean parameter, $S_j \in \{0, 1\}$, that identifies if sample j has a sensitive value. Finally, α is a parameter that controls the restrictiveness of the constraint, i.e., the higher the value of α, the less the restrictions in solving this optimization problem. This parameter is introduced to be able to tune the restriction with respect to diversity in each qid group.

Let us discuss the proposed formulated optimization problem. The $|B_{ij} \cdot (X_j - Center_i)|$ expression in Eq. (3) is the Manhattan distance of sample *j*, X_j, and cluster center *i*, $Center_i$, when the Boolean variable B_{ij} is equal to one. B_{ij} will be equal to one, $B_{ij} = 1$, if sample X_j belongs to cluster *i*, and it will be zero otherwise. The objective function in Eq. 3 intends to optimize B_{ij}s to minimize the distance between samples in cluster *i* and $Center_i$, for all clusters and samples.

Eqs. (4)-(7) are the constraints of our proposed optimization problem:

- The first constraint, in Eq. (4), forces each sample to belong to only one cluster. This is done by ensuring that B_{ij} is one exactly once for all *i*.
- The second constraint, in Eq. (5), forces the number of samples in each cluster to be equal to *k*. The summation of the number of samples must be equal to *k* for cluster *i*. This condition can readily be relaxed to: at least *k* samples in each cluster.
- The third constraint, in Eq. (6), finds the optimized cluster centers, i.e., C_is. The optimized center for cluster *i* is the average of all *k* samples that belong to cluster *i*.
- Finally, the last constraint, in Eq. (7), forces the optimization to distribute the samples with sensitive values ($S_j = 1$) into all clusters. The left-hand side of the constraint is equal to the number of sensitive values in cluster *i*. The right-hand side is the number of samples with sensitive value divided evenly among the clusters (multiplied by α, which is the parameter for relaxing the hard constraint in our optimization problem).

After the optimization, we know which sample belongs to which qid group or cluster, based on *B* matrix. We also know the optimized cluster centers, identified based on the values of C_is. Therefore, the values of sample quasi-identifiers will be replaced by their respective cluster center values. In this way, we obtain a solution that addresses record-linkage and attribute-linkage attack models. We force the samples in the anonymized data to be indistinguishable from *k-1* other samples while considering the diversity of values in the sensitive attribute.

Fig. 1 presents an example in which the solution in Fig. 1b merely considers *k*-anonymity property, while Fig. 1c considers the diversity of values in the sensitive attribute addressed in *l*-diversity and *t*-closeness. The color of the circles shows if the samples contain a sensitive value. If the color is blue, the sample does not have a sensitive value, $S_j = 0$, while a red circle shows having a sensitive value $S_j = 1$.

In Fig. 1b, samples *0, 1, 2, 4, 9* fall in the same qid group. The rest of the samples fall in the second group. By sharing the cluster centers for each group, we achieve 5-anonymous data. However, in such a solution, the samples with sensitive values are not evenly distributed. By considering the constraint introduced for the diversity of values in the sensitive attribute, we obtain the solution presented in Fig. 1c. In this solution, the data is still 5-anonymous, i.e., it has five samples in each cluster. Nevertheless, in this case, sample *2*, falls in the same cluster with *5, 6, 7, 8* to evenly distribute samples with sensitive values.

IV. Evaluation and Discussion

In this section, we evaluate the proposed method experimentally and discuss the experimental results. For evaluation, we consider data utility and data privacy criteria and demonstrate their trade-off [39]. Then, we present and discuss the experimental results.

In this paper, the data analysis task that is going to be performed on the anonymized data is classification. Therefore, the anonymization method should alter the data to the extent that learning high-performance classification models are possible. We train the learning algorithms on both original and anonymized data to evaluate the anonymization method in terms of data utility preservation. Our method preserves the data utility if the classification model learned from altered data has similar performance compared to the one learned from original data.

TABLE II: Classification performance for trained models on three different versions of Heart Disease dataset (Cleveland) [37], [38]

Algorithm	Original Data			Anonymized Data Without Diversity			Anonymized Data by Our Method		
	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC	F1-score	Accuracy	MCC
ERT	81.0%	81.0%	0.615	81.1%	81.4%	0.625	81.0%	81.4%	0.625
Random Forest	82.5%	82.6%	0.647	80.1%	80.4%	0.603	80.0%	80.3%	0.602
XGBoost	78.9%	79.0%	0.573	74.7%	75.1%	0.493	74.7%	75.1%	0.495
Decision Tree	73.8%	73.8%	0.470	68.9%	69.3%	0.372	69.2%	69.8%	0.382
SVM	83.0%	83.1%	0.656	73.3%	73.3%	0.459	72.8%	72.9%	0.449

On the other hand, the anonymized data should be sufficiently altered to avoid the identification of record owners. In this paper, we address the record-linkage and attribute-linkage attack models. We consider the property for making samples indistinguishable in the qid group, discussed in *k*-anonymity privacy model, the diversity of values in sensitive attribute, in *l*-diversity, and the frequency of sensitive values, in *t*-closeness.

There is a trade-off between the utility of data and privacy of data in anonymization methods. On the one hand, we can share no data to preserve patients' privacy, but there will be no utility for the data. On the other hand, we can publish the data in its original format to maximize the data utility, but the privacy of data subjects is going to be violated. Therefore, in anonymization techniques, we require altering the data to the extent that we establish a trade-off between data utility and privacy [39].

A. Experimental Setup

In our experiments, we use the Heart Disease dataset [37], which is one of the popular datasets publicly available on the UCI repository. We utilize Cleveland's processed dataset [38] to predict the presence of heart disease (presence/absence). The dataset contains 282 complete records, and each belongs to one patient. The data includes 13 attributes which we consider in this work.

Quasi-identifiers are the attributes that the adversary can potentially obtain information about them from other sources. In addition to quasi-identifiers, the sensitive attribute should also be identified. In our experiments, we suppose all 13 attributes are quasi-identifiers. Moreover, we select the Boolean attribute for family history of coronary artery disease as the sensitive attribute.

For evaluation of preservation of utility, we split the dataset into train and test sets. We anonymize the training set using our method with soft constraints and train several classification algorithms based on the resulting data. Then, we measure the classification performance on the test set. We also train the same algorithms on the original data and the data anonymized without considering the diversity constraint and measure the performance of the trained classification models on the test set. The comparison of the classification performance results indicate the utility of anonymized data in our method.

In our experiments, we randomly select 200 samples as the train set and the rest as the test set at each round. We repeat the same process for 1000 rounds and report the average results for classification performance. The algorithms used for learning classification models are Extremely Randomized Trees (ERT), Random Forest, XGBoost, Decision Tree, and linear SVM. The measures used for classification performance are F1-score, Accuracy, and Matthews Correlation Coefficient (MCC).

B. Experimental Results

Table II shows the classification performance results for three different training sets, i.e., original data, anonymized using our method, and anonymized without considering the diversity constraint. For both anonymization methods *k* is set to 10.

The classification results for the original data are at a similar level ($\pm 0.5\%$ due to randomness in the algorithms) or higher than the anonymized data. However, since there is a trade-off between privacy and utility in anonymization [39], we may accept a loss in the utility to obtain privacy. The results in Table II show that our method preserves the information in data that leads to learning high-performance models. Moreover, the classification performance difference between our method and the approach without considering the diversity is negligible. This indicates that introducing the diversity constraint in our method does not significantly affect the data utility.

We now evaluate the privacy preservation of our method in Table III. Here, we set the value of *k* to 10. This means that if the adversary has the values for quasi-identifiers for one patient, he/she can only map his/her information to 10 records. Therefore, through our method, we avoid record-linkage attacks. Second, our method evenly distributes the samples with sensitive value, i.e., having a family history of coronary artery disease, to qid groups. This weakens the confidence of the adversary's inference for identifying a patient with sensitive value.

The number of patients with the sensitive value can be different at each round. In our method, in the worst qid group with respect to *l*-diversity, entropy *l*-diversity, and recursive *(c,l)*-diversity, we have two samples with non-sensitive value and eight with the sensitive value. In other words, the proportion of patients with a family history of coronary artery disease in the qid group is 80.0%, which is optimal since the proportion of samples with the sensitive value in the training set at this round was 70.5%. This leads to $l = 2$ in *l*-diversity, $l = 1.64$ in entropy *l*-diversity, and $l = 2$ and $c \geq 4$ in recursive *(c,l)*-diversity in Table III. In the worst qid group with respect to the variational distance D in *t*-closeness, we have six with non-sensitive value and four with the sensitive value, while the proportion of samples with

TABLE III: Privacy properties of the anonymized data by our method and the approach without diversity

	No Diversity	Our Method
k in *k*-anonymity	10	10
l in *l*-diversity	1	2
l in entropy *l*-diversity	1	1.64
l and *c* in recursive *(c,l)*-diversity	*l=1*, $c \geq 1$	*l=2*, $c \geq 4$
D in *t*-closeness	1.06	0.38

the sensitive value in the dataset at this round was 59.0%. This leads to variational distance $D = 0.38$ in *t*-closeness.

For the approach without diversity constraint, in the worst qid group with respect to *l*-diversity, entropy *l*-diversity, and recursive *(c,l)*-diversity, we have ten patients with the sensitive value. This leads to $l = 1$ in *l*-diversity, $l = 1$ in entropy *l*-diversity, and $l = 1$ and $c \geq 1$ in recursive *(c,l)*-diversity in Table III. This allows the adversary to infer that the patient had a family history of coronary artery disease with 100% confidence. Moreover, in the worst qid group with respect to the variational distance *D* in *t*-closeness, we have nine records with the non-sensitive value and one with the sensitive value. The proportion of samples with the sensitive value in the dataset at this round was 63.0%. This increases the variational distance between the distributions of values in the sensitive attribute in the qid group and the whole dataset to $D = 1.06$ in Table III.

The results in Table III demonstrates that by adopting our method, we will have higher *l* in *l*-diversity, entropy *l*-diversity, and recursive *(c,l)*-diversity. Moreover, the variational distance between the distributions of values in the sensitive attribute for the train set and the qid group is lower in our method. Therefore, regarding the diversity of values in sensitive attributes and the attribute-linkage attack, we observe that introducing the diversity constraint improves patients' privacy.

We also investigate the data privacy and data utility based on different values of *k*, size of qid groups. For each *k*, we have 100 rounds that in each we randomly split the data into the train and test sets. The classification performance results are the average results for all rounds. The privacy results are the worst results in all rounds and qid groups. We perform these experiments based on our method and the anonymization approach without the diversity constraint and show the results in Figs. 2 and 3 for comparison.

Figs. 2a-2c show the results based on F1-score, Accuracy, and MCC metrics. The patterns in the results show that the higher the qid group size (*k*), the lower the classification performance. On the other hand, increasing the value of *k* improves the privacy with respect to the record-linkage attack model. These figures illustrate the trade-off between the privacy and data utility.

The results in Figs. 3a-3d exhibit the privacy properties of the anonymized data. Regarding the attribute-linkage attack model, the results display that the data anonymized by our method has higher privacy properties than the anonymized data without diversity constraint. Increasing the value of *k* significantly improves the diversity and frequency of values in the sensitive attribute, compared to the approach without considering the diversity constraint, but without any major loss in terms of classification performance.

The experimental results show that our method provides privacy against record-linkage and attribute-linkage attacks. Furthermore, the utility of the data is retained after anonymization, allowing learning of high-performance classification models. The slight degradation of utility is the cost for providing patients privacy, which is a common phenomenon in anonymization approaches [39].

V. Conclusion

In this paper, we have proposed a method for obtaining anonymized data by ensuring that data samples are indistinguishable in qid groups while considering the diversity and frequency of values in the sensitive attribute. Our method is based on constrained optimization and clustering of the samples into qid groups by jointly considering the *k*-anonymity, *l*-diversity, and *t*-closeness privacy models. The evaluation results show that the proposed method retains data utility while reducing the privacy concerns related to data sharing.

ACKNOWLEDGMENT

This research is supported by INTROducing Mental health through Adaptive Technology (INTROMAT) project. The paper is partially supported by SIRIUS: Centre for Scalable Data Access.

References

[1] S. D. Lustgarten, Y. L. Garrison, M. T. Sinnard, and A. W. Flynn, "Digital privacy in mental healthcare: current issues and recommendations for technology use," *Current Opinion in Psychology*, 2020.

[2] D. Pascual, A. Amirshahi, A. Aminifar, D. Atienza, P. Ryvlin, and R. Wattenhofer, "Epilepsygan: Synthetic epileptic brain activities with privacy preservation," *IEEE Transactions on Biomedical Engineering*, 2020.

[3] M. Kantarcioglu, "A survey of privacy-preserving methods across horizontally partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[4] J. Vaidya, "A survey of privacy-preserving methods across vertically partitioned data," in *Privacy-preserving data mining*. Springer, 2008.

[5] J. Vaidya, B. Shafiq, W. Fan, D. Mehmood, and D. Lorenzi, "A random decision tree framework for privacy-preserving data mining," *IEEE transactions on dependable and secure computing*, 2013.

[6] A. Aminifar, F. Rabbi, K. I. Pun, and Y. Lamo, "Privacy preserving distributed extremely randomized trees," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 2021.

[7] A. Aminifar, F. Rabbi, and Y. Lamo, "Scalable privacy-preserving distributed extremely randomized trees for structured data with multiple colluding parties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

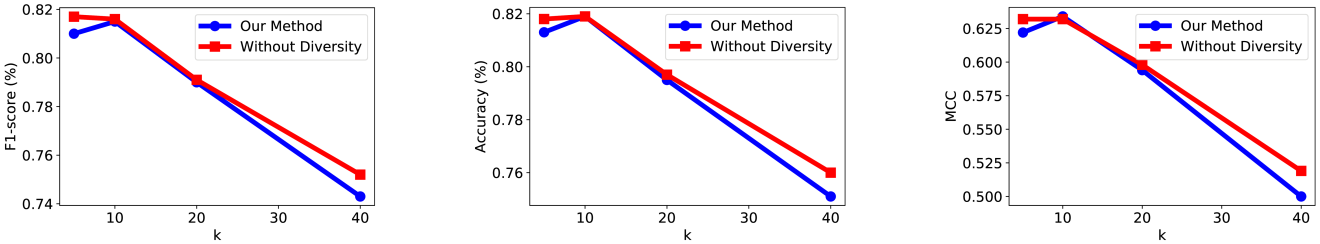
[8] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[9] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[10] S. Baghersalimi, T. Teijeiro, D. Atienza, and A. Aminifar, "Personalized real-time federated learning for epileptic seizure detection," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[11] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000.

[12] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in *21st International Conference on Data Engineering (ICDE'05)*. IEEE, 2005.

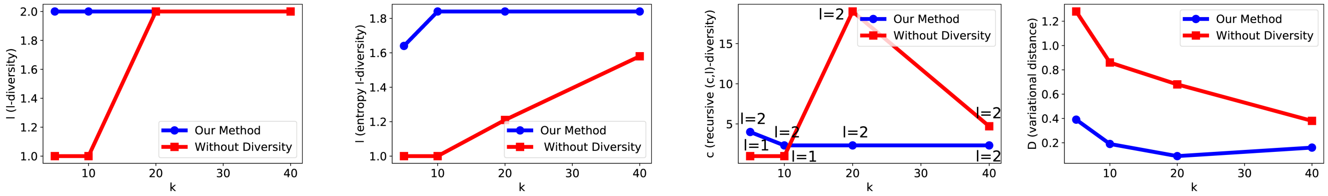


(a) F1-score

(b) Accuracy

(c) MCC

Fig. 2: The classification performance for anonymized data based on F1-score, Accuracy, and MCC measures for different values of k



(a) l in l-diversity

(b) l in entropy l-diversity

(c) c in (c,l)-diversity

(d) D in t-closeness

Fig. 3: The privacy properties of the data anonymized by our method and the approach without considering the diversity constraint for different values of k

[13] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 2002.

[14] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Third IEEE international conference on data mining*. IEEE, 2003.

[15] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 2005.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[17] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proceedings of the international conference on internet of things design and implementation*, 2019.

[18] M. Alzantot, S. Chakraborty, and M. Srivastava, "Sensegen: A deep learning architecture for synthetic sensor data generation," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2017.

[19] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *arXiv preprint arXiv:1806.03384*, 2018.

[20] D. Sopic, A. Aminifar, A. Aminifar, and D. Atienza, "Real-time event-driven classification technique for early detection and prevention of myocardial infarction on wearable systems," *IEEE transactions on biomedical circuits and systems*, 2018.

[21] ——, "Real-time classification technique for early detection and prevention of myocardial infarction on wearable devices," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017.

[22] A. Aminifar, P. Eles, and Z. Peng, "Optimization of message encryption for real-time applications in embedded systems," *IEEE Transactions on Computers*, 2017.

[23] F. Forooghifar, A. Aminifar, and D. Atienza, "Resource-aware distributed epilepsy monitoring using self-awareness from edge to cloud," *IEEE transactions on biomedical circuits and systems*, 2019.

[24] "Health informatics — Pseudonymization," International Organization for Standardization," Standard, 2017.

[25] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[26] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007.

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[28] N. Mohammed, B. C. Fung, P. C. Hung, and C.-k. Lee, "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*, 2009, pp. 1285–1294.

[29] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

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

[31] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *22nd International conference on data engineering (ICDE'06)*. IEEE, 2006.

[32] A. Majeed, F. Ullah, and S. Lee, "Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data," *Sensors*, 2017.

[33] A. Aminifar, Y. Lamo, K. I. Pun, and F. Rabbi, "A practical methodology for anonymization of structured health data," in *Proceedings of the 17th Scandinavian Conference on Health Informatics*, 2019.

[34] K. Doka, M. Xue, D. Tsoumakos, and P. Karras, "k-anonymization by freeform generalization," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 2015.

[35] Y. Liang and R. Samavi, "Optimization-based k-anonymity algorithms," *Computers & Security*, 2020.

[36] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, 2010.

[37] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, 1989.

[38] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[39] J. S. Davis and O. Osoba, "Improving privacy preservation policy in the modern information age," *Health and Technology*, 2019.