k-Anonymity on Metagenomic Features in Microbiome Databases

Rudolf Mayer
TU Wien & SBA Research
Vienna, Austria
mayer@ifs.tuwien.ac.at,rmayer@sba-research.org

Alicja Karlowicz
SBA Research
Vienna, Austria
akarlowicz@sba-research.org

Markus Hittmeir
SBA Research
Vienna, Austria
mhittmeir@sba-research.org

ABSTRACT

The human microbiome is increasingly subject to extensive research, due to its relations to health, diet, exercise and illness. While ever more microbiome data is gathered and stored, recent works have demonstrated the threat of individual re-identification based on matching samples taken at different points in time, by matching metagenomic features extracted from microbiome readings. The individual and temporal stability of the microbiome varies for different body sites and is particularly pronounced for readings from the gastrointestinal tract. To meet the resulting need for privacy-protecting solutions, we adapt the well-known concept of k-anonymity and make it suitable for application to microbiome datasets. In particular, our approach for establishing k-anonymity is based on micro-aggregation.Our evaluation uses ten datasets containing samples of gut microbiomes, and analyzes the decreased privacy risk on the anonymised dataset as well as the incurred information loss. The analysis demonstrates the suitability of our approach for the protection of sensitive microbiome data.

CCS CONCEPTS

• **Security and privacy → Data anonymisation and sanitization**; *Privacy protections.*

KEYWORDS

Human Microbiome, Re-Identification, Anonymisation

ACM Reference Format:
Rudolf Mayer, Alicja Karlowicz, and Markus Hittmeir. 2023. k-Anonymity on Metagenomic Features in Microbiome Databases. In *The 18th International Conference on Availability, Reliability and Security (ARES 2023), August 29–September 01, 2023, Benevento, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600160.3600178

1 INTRODUCTION

Research on the human microbiome has been flourishing for several years, sparked by its high potential for analysis in clinical settings. The human microbiome includes bacteria, archaea, viruses, fungi and protists, living on different sites of the human body, i.e. on or within human tissues and biofluids of various anatomical sites.

ARES 2023, August 29–September 01, 2023, Benevento, Italy

ACM ISBN 979-8-4007-0772-8/23/08.
https://doi.org/10.1145/3600160.3600178

Recent works suggest that the human microbiome has a great influence on our general well-being. Variations in the human microbiome indicate details about a person's diet, exercise habits and abode. Therefore, the human microbiome is increasingly utilised for prediction, diagnosis and therapy of diseases, leading to frequent publication of new results on certain correlations and interactions. For example, changes in the microbiome of the gastrointestinal tract may be related to gastrointestinal diseases [3], obesity [12], diabetes [16], and depression [17]. Human microbiome data has been used to detect e.g. colorectal cancer [25] or type-1 diabetes in infants [5].

Besides its potential for analysis in clinical settings, previous works have shown that it is vital to consider the human microbiome as personal and sensitive medical data. Franzosa et al. ([6]) demonstrated that the individual variations in metagenomic features extracted from microbiome readings allow for the matching of multiple samples of the same individuals among populations of hundreds: for an initial microbiome sample, it was attempted to match them to follow-up samples collected 30-300 days later, with a correctness of approx. 30%. Especially the widely-considered gastrointestinal microbiome was shown to be very stable, allowing to match up to 80% of individuals. This personal microbiome identification (PMI) thus poses a privacy threat to those participating in microbiome studies. These observations have been further aggravated by a nearest-neighbour approach for the matching of microbiome samples ([10]), thus stressing the need for privacy-enhancing technologies for microbiome data While there are emerging techniques for genomic datasets ([2]), and specifically for human DNA sequence data ([13], [14]), techniques for protecting *microbiome reports* against personal microbiome identification are still lacking. Such microbiome reports are tables containing individuals (i.e., sample vectors) described by hundreds to thousands of metagenomic features. Previous work ([11]) studied the utility of synthetic microbiome data, and the experiments on several classification tasks showed only small deviations of utility scores of models trained on synthetic data compared to models trained on the original data. Our contribution is an evaluation of another privacy-protecting technique, k-anonymisation, obtained by the micro-aggregation of microbiome samples. Our ultimate goal is to ensure that PMI methods only find sets of samples of size at least k that may or may not contain the correct sample, but are no longer able to uniquely identify the correct sample. In addition, we will analyse the utility of the anonymised microbiome data on the same datasets as those used in [11].

The remainder of this paper is structured as follows: In Section 2, we give a more detailed overview of PMI methods and on the concepts used in our anonymisation method. Section 3 describes the threat model and the goal of our technique. In Section 4, we use four datasets containing samples of gut microbiomes and six additional datasets for the predictive machine learning task to evaluate anonymised data utility. Finally, we provide concluding remarks and a discussion on future work in Section 5.

2 PRELIMINARIES AND RELATED WORK

We start by discussing relevant aspects on microbiome data. Samples of human microbiome may be taken from several body sites – besides the microbiome of the gastrointestinal tract mentioned above, this might be from saliva, throat, anterior nares (the external portion of the nose) or buccal mucosa (the inside of the cheek). While gut microbiome data is particularly useful for clinical research, it is also most vulnerable to the mentioned PMI methods. We will thus focus on the gastrointestinal datasets from Franzosa et al.'s study ([6]). These were obtained from raw microbiome data that is publicly available through the Human Microbiome Project's (HMP) repository[1]. From the initial metagenomic sequences, ([6]) applied 16S ribosomal gene sequencing as well as whole metagenome shotgun sequencing, to eventually derive structured tabular datasets with different feature types. Samples from these datasets then serve as input to the PMI methods by Franzosa et al. ([6]) and in the improved approach ([10])[2]. In these two PMI approaches, the input is tabular data: the samples (individuals) are represented in rows, and the columns contain the features describing the samples. The datasets come with four different feature types: (i) operational taxonomic unit abundance ('OTU'), (ii) bacterial and archael species abundance ('Species'), (iii) species-specific marker genes ('Markers') and (iv) tiled kilobase windows ('KBW'). The units for each sample are either measured in (i) *relative abundance*, which means that the values correspond to percentages, and that thus their sum for each sample equals 1, or (ii) *reads per kilobase per million sample reads* (RPKM); these values are scaled by a different factor and may be much larger than 1.

In Franzosa et al.'s method for PMI in tabular data ([6]), a feature is considered to be binary, i.e. either present or absent, determined by feature detection limits that correspond to the general unit size; thus, the original sample vectors with continuous values are effectively transformed into binary 0, 1-vectors. The method then tries to identify so-called *metagenomic codes*, which are unique for each sample. The code comprises a small subset of features that are present in the individual sample, but are, in this combination, not present in any other sample. To obtain these codes, Franzosa et al. use a greedy algorithm; they empirically demonstrated that these codes are stable enough over time, and may thus be used to find pairs of samples that belong to the same individual.

This PMI technique was improved by [10] with a nearest-neighbour extension, consisting of three phases. In the first phase, the relative abundance and RPKM values in the sample vectors are discretised, i.e. they are transformed to integer values, by using feature abundance limits. This phase is similar to the encoding via the feature detection limits used in the metagenomic-code approach, but is more fine-granular than the binary discretisation. The second phase identifies possible matches: to match a specific sample s against a dataset D, one computes the most similar (the "nearest-neighbour") sample $\bar{s}$ of s in D[3]. Thus, one obtains $(s, \bar{s})$ as candidate for a pair of samples belonging to the same individual. A problem of this nearest-neighbour approach is that *every* sample s has *some* nearest-neighbour in D; thus, this method generally leads to a large number of false positive matches. This problem is solved, in the third phase of the algorithm, by introducing a criterion for deciding on whether to accept or reject a pair $(s, \bar{s})$. Intuitively, this criterion contrasts the similarity between the potential matching pair s and $\bar{s}$ to the similarity between $\bar{s}$ and *all other samples* in D. It has been shown empirically in a large number of experiments in [10] that this acceptance criterion, a form of *thresholding*, is able to filter most false positives from true positives. This thresholded nearest-neighbour method [10] overall shows an increased success over the earlier method [6] on most considered body sites. Especially, we can note an increased percentage of true positive matches (by up to 30%) of the widely studied gut microbiome, averaged over the four different feature types described earlier.

Privacy-preserving data publishing (PPDP) [7] deals with providing data that can be shared with other parties, without infringing the privacy of contained individuals. A wealth of different methods have been proposed, including e.g. synthetic data generation (SDG), where new, artificial data is generated from a learned representation of the original data, e.g. the distribution of the features and the correlations between them [8]. As a result, a new dataset is obtained with no 1-1 correspondence to real individuals, while still being useful for analysis. While synthetic data is used for a variety of machine learning tasks, e.g. classification [9], regression, or anomaly detection [15], current SDG methods often need to be adapted to deal with the idiosyncrasies of data from specific domains. Such an adaption was performed in [11] for microbiome tabular data for machine learning tasks. To put the results from that work into perspective and further investigate the PMI methods' effectiveness, in this paper, we adopt an anonymisation technique for privacy-preserving data publishing.

Within a dataset, we can generally distinguish different types of attributes (features). *Identifying attributes*, such as an e-mail address or a social security number, directly reveal the identity of a record. *Quasi-identifiers* (QIs) do not directly identify a record, but may do so when used in combination with other quasi-identifiers. As an example, in a dataset containing demographic information on individuals, the date of birth in combination with the sex and the residence of a person can uniquely identify a certain number of records. While the attributes in the microbiome datasets used in our evaluation are more abstract and carry less semantic meaning, they can in combination still become identifying. The privacy model k-anonymity [18] is a well-explored approach that can be used to obfuscate sensitive datasets and prevent (re-)identification of individual samples. A k-anonymous dataset means that there are

[1] https://www.hmpdacc.org

[2] There are several other methods for microbiome-based identification, e.g. GePMI by Wang et al. ([23]). However, GePMI is not based on tabular microbiome data, but extracts its features from the raw microbiome sequence data. Since it does not operate on datasets with *metagenomic* features, it is not in the scope of this paper.

[3] In principle, this computation may be performed with respect to any distance metric. In [10], the Pearson correlation coefficient has been used.

always at least k records that are *indistinguishable* with regards to the quasi-identifiers. Records that share the same quasi-identifier values are called *Q-blocks* or *equivalence groups (or classes)*.

k-anonymity can be achieved by suppression and generalisation of the values that are unique within a Q-block. Suppression means simple deletion of values, whereas generalisation decreases a value's granularity. *Global* generalisation (also called full domain generalisation) means that an attribute is put to the same generalisation level for each data record. *Local* generalisation on the other hand optimises the generalisation by finding a minimal required loss of precision for each Q-block.

Micro-aggregation ([4], [20], [1]) is related to k-anonymity and well-suited for numerical data. In micro-aggregation, Q-blocks (clusters) containing at least k similar records are created in a similar fashion to local generalisation, but the samples are made indistinguishable by replacing attribute values by Q-block (cluster) averages. This means that (i) the granularity of the input data is not decreased, (ii) continuous numerical attributes are not discretised and (iii) outliers cause less distortion (whereas in data generalisation, they may force very coarse values) [21].

As we will discuss in Section 3, the structural similarities of the two PMI methods imply that a solution for mitigating the risks resulting from the nearest-neighbour approach also protects against the metagenomic-code method. Our main idea is to establish k-anonymity on the dataset by applying an approach based on micro-aggregation, with the aim of preventing both methods from finding unambiguous matches. To obtain k-anonymity, we selected the **μ-Ant** tool as a stand-alone open source Java software [21]. The implementation scales well for high-dimensional data and is easily configurable. μ-Ant requires as input a CSV file with the data to be anonymised, and an XML configuration file describing the attributes of the dataset, along with their types, sensitivity, and desired anonymisation parameters to be obtained. The output is another CSV file with an anonymised version of the dataset.

3 THREAT MODEL AND MAIN GOAL

In this work, we consider a threat model similar to [11].

Victim: an individual who provided their microbiome samples, e.g. in the course of medical studies, diagnosis, therapy, or personal health and fitness advise. Their microbiome data, and possible analysis results or additional metadata, are electronically available.

Adversary: a party in possession of unidentified microbiome samples, wanting to link them to other samples. Their goal is to accumulate information about the underlying individuals and possibly identify them. The adversary may obtain microbiome samples via various ways, e.g. from public microbiome databases, cyberattacks against healthcare or research facilities, data exfiltration via insiders, or even directly from the human victim (e.g., saliva).

Threat: we assume that an adversary possesses a sample of a certain individual. We discuss four reasons for the adversary to match the sample against another database.

(i) To find out if a person participated in a certain study, i.e. membership disclosure. This might allow them to infer sensitive information, e.g. if the study is conducted in the context of a disease. Even if the microbiome samples connected to the disease study do not include identifying metadata, a match with a known sample will identify the person.
(ii) The attacker may be able to obtain (previously unknown) metadata that is associated with the sample identified as match in the new database (e.g., medical and personal data provided in the course of a study or treatment).
(iii) The attacker may be able to get hold of new microbiome samples from the same individual, and even in the absence of metadata could thereby learn about changes over time in the individual's human microbiome. These changes could be caused, e.g., by diseases, depression, or changes in diet.
(iv) Ongoing research increasingly associate microbiome samples with other individual traits, such as the age or geographical background [26]. Therefore, collecting and linking multiple samples from the same individual could also aid an adversary in identifying the person behind a sample.

In Section 2, we discussed currently known techniques for personal microbiome identification on datasets containing metagenomic features. While these techniques show that personal microbiome identification is possible, this paper aims at *mitigating* the threats discussed above. The ultimate goal of our anonymisation approach is to prevent PMI methods from achieving unique re-identifications on the anonymised dataset. Considering the functionality of the metagenomic-code approach and the thresholded nearest-neighbour technique, we specify the following two objectives. After applying feature abundance limits on the anonymised dataset, it should be impossible to

(1) construct a unique metagenomic code for any individual
(2) distinguish any individual sample from all others

The first objective refers to the metagenomic-code approach [6], the second to the nearest-neighbour technique [10]. From a privacy-preserving perspective, it is clear that a solution for achieving the second objective may also serve as a solution for the first objective. Hence, our anonymisation approach should produce an output dataset such that, after applying feature limits, it satisfies (at least) k-anonymity in the sense of the second goal. In addition, our strategy presented and evaluated in Section 4 will also consider the utility of the anonymised datasets, e.g. by computing measures for the information loss and comparing performance on machine learning classification tasks.

4 EVALUATION

We evaluate the proposed k-anonymity approach in three steps. First, we demonstrate the infeasibility of personal microbiome identification on the anonymised datasets, i.e. we demonstrate the privacy-preserving power of the method. Subsequently, we measure the effects of anonymisation on the data quality and utility, by two different means: we measure (i) the statistical information loss compared to the original data, and (ii) the utility of the anonymised dataset for selected downstream task, namely supervised classification tasks from knights-lab's microbiome machine learning repository[4].

For the first part on PMI, we apply our approach to four datasets containing samples from the microbiome of the gastrointestinal

[4] https://knights-lab.github.io/MLRepo

tract. Originally, they have been published in Franzosa et al.'s study [6][5], which is based on raw microbiome sequence data available through the Human Microbiome Project. As indicated in Section 2, the four datasets correspond to four different metagenomic feature types: (i) operational taxonomic unit abundance ('OTU'), (ii) bacterial and archael species abundance ('Species'), (iii) species-specific marker genes ('Markers') and (iv) tiled kilobase windows ('KBW'). Species and OTUs are on the taxon-level and measured in *relative abundance*, i.e. the sum of all components in each sample vector equals 1. Markers and KBW, on the other hand, are on the gene-level and measured in *reads per kilobase per million sample reads* ('RPKM'), and are not normalised values.

Table 1: Characteristics of the four gut microbiome datasets used in the first part of our experiment

Dataset	# Features	# Individuals
Species	317	50
OTUs	2663	87
Markers	349,779	50
KBW	263,847	45

Table 2: Feature abundance limits used in nearest-neighbour PMI ([10])

Limits	t_0	t_1	t_2	t_3	t_4
Taxon-Level	0.00005	0.0005	0.005	0.05	0.5
Gene-Level	0.005	0.05	0.5	5	50

Table 1 provides an overview on the size of the four datasets. We in fact have two datasets D and F for each of the feature-types: D and F contain exactly the same individuals represented by the same features, but D contains initial microbiome samples, while F contains follow-up samples from the same individuals at a later point in time (30-300 days later). Interpreting this along our threat model (see Section 3), D is the dataset under attack, and F contains samples available to an adversary that they want to check against D. Furthermore, we will make use of the feature abundance limits given in Table 2, encoding the dataset by applying following rule:

$$x \leftarrow \begin{cases} 0 & \text{if } x < t_0, \\ i & \text{if } t_{i-1} \leq x < t_i \text{ for some } 1 \leq i \leq 4, \\ 5 & \text{if } x \geq t_4. \end{cases}$$

To summarise, we employ the following experimental setup:

(1) Apply the anonymisation with the μ-Ant tool to D, specifying all features as quasi-identifiers, and obtain $\bar{D}$.
(2) Perform personal microbiome identification to match the samples of F with those in $\bar{D}$. Compare the results to those obtained on the original dataset D. This measures the achieved privacy.
(3) Let $\bar{D}_f$ and D_f be the datasets after application of the feature abundance limits in Table 2. Compare $\bar{D}_f$ to D_f and analyse the information loss. This measures the reduction in data utility (see Section 4.2).

[5]Code and data is available at: https://huttenhower.sph.harvard.edu/idability/

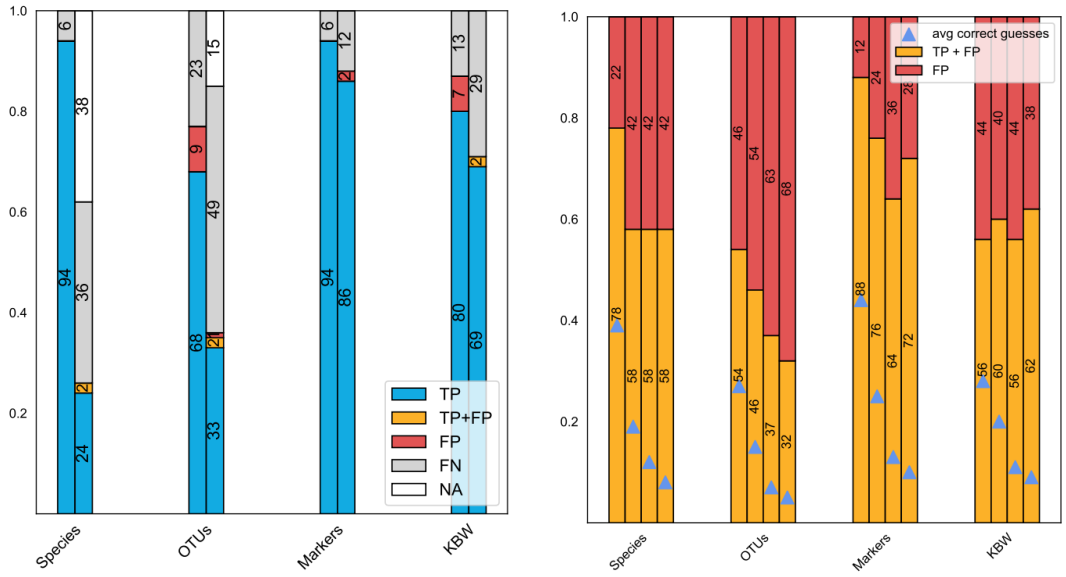
The second step demonstrates that the anonymisation algorithm holds its promise of greatly inhibiting the capabilities of the currently best techniques for PMI. The purpose of the third step is the investigation of the preserved utility of the anonymised dataset.

4.1 Empirical verification of k-anonymity

We perform this step by applying the personal microbiome identification (PMI) techniques; serving as a baseline, we first discuss the results on the original datasets. Figure 1a presents these results for the two PMI methods discussed in Section 2. The left bars show the nearest-neighbour technique ([10]), and the right bars the metagenomic-codes approach ([6]). For every sample in F, the methods construct a set of matches of samples from D. In our case, these sets contain either the nearest-neighbour ([10]) or the samples that match the metagenomic code ([6]). We count the number of sets of matches containing (i) only the correct individual (true positives, TP), (ii) only wrong individuals (false positives, FP), (iii) correct and also wrong individuals (TP+FP), as well as (iv) the number of individuals that incorrectly have not been matched (FN), and (v) the number of individuals for which the technique of [6] was not able to construct a unique metagenomic code (NA). Comparing the left to the right bars in Figure 1a, we can see that the nearest-neighbour approach outperforms the metagenomic-codes technique on each of the datasets and achieves a high number of true-positive identifications.

Let us now consider the results of anonymisation on the performance of the PMI methods. After the anonymisation of the dataset D with the procedure described in the beginning of this section, we obtain $\bar{D}$. We now match the samples of F with those in $\bar{D}$ by applying the PMI techniques to these datasets. Our observation is that the approach based on metagenomic codes is not able to construct a unique code for *any* of the samples in $\bar{D}$, hence the result for every sample in F is 'NA'. Similarly, the thresholded nearest-neighbour approach yields 'FN' as result for each sample, since it rejects every initial match. These outcomes were to be expected, as the anonymisation algorithm is designed to lead to this exact behaviour of the PMI methods with standard settings. We can further consider a variation where an attacker might try to bypass the mentioned obstacles and obtain some other output besides 'NA' or 'FN': the attacker might decide to turn off the threshold in the nearest-neighbour approach. This will certainly allow them to obtain some coincidental matches (TP + FP). However, since we have k-anonymity in the sense that always k samples in the anonymised datasets are indistinguishable, the chance to obtain the correct sample by guessing is (at most) $1/k$. With increasing levels of anonymity, the number of neighbouring samples returned by the PMI method is also increasing, but only one among them can be an actual match.

Applying our nearest-neighbour approach with the number of neighbours equal to anonymisation level k (looking for the k closest neighbours) and without a threshold for the acceptance of matches yields the results shown in Figure 1b. Bars in the presented plot correspond to consecutive anonymity levels, starting from the left with $k = 2$, increasing up to $k = 7$. The blue markers indicate the percentage of samples for which the attacker on average guesses correctly for the different levels of k-anonymity – an attacker will



(a) PMI on original datasets

(b) PMI on the anonymised datasets

Figure 1: PMI before and after anonymisation (results in percent). In (a), the left bar shows the results from the nearest-neighbour (NN) method ([10]), and the right bar shows the results from the metagenomic-codes method ([6]). In (b), we applied the nearest-neighbour approach without thresholding, setting the number k of nearest neighbours equal to the anonymity level. Bars correspond to respective anonymity levels, starting from the left with $k = 2, 3, 5, 7$. Blue triangles indicate the percentage of average correct guesses by the attacker, assuming a chance of 1/k in the TP+FP and a chance of 0 in the FP matches.

always guess wrong for the FP outcomes, while for TP+FP matches, the correct guess is on average (TP+FP)/k, as described above. The ratio of correct guesses on average is naturally decreasing with an increase in k, the anonymity level. The number of TP+FP matches varies between datasets and is similar or even lower for higher levels of k. As the Q-blocks' size of nearest neighbours increases with k, one would expect an increase in TP+FP matches for higher levels of k. However, that does not appear to be the case, as can be seen in Figure 1b.

Our assumption is that the closeness of original samples in each dataset varies and hence, while performing PMI, some Q-blocks of k-nearest neighbours are overlapping within particularly close samples. Since the PMI method only gets k samples as an outcome and determines if any sample within that subset is a match, there is a higher chance of getting only false positives if the samples are similar. This appears to be the case for the OTUs dataset, where the FP ratio increases with a higher anonymity level.

In comparison with the left bars in Figure 1a (the thresholded nearest-neighbour approach with $k = 1$), we can see that an attacker can still find a large number of correct Q-blocks on the anonymised datasets. This behaviour may be considered as a consequence of the fact that the anonymisation algorithm is designed to provide k-anonymity by simultaneously losing as little information as possible. However, an attacker now has to guess the correct individuals from the obtained TP+FP Q-blocks, and also has no means to distinguish them from the FP Q-blocks, in which all samples are incorrect. We hence conclude that the goal stated in Section 3 is achieved.

4.2 Statistical utility analysis

We proceed with a statistical analysis of the differences between original and anonymised datasets. As mentioned in the experimental setup, we compare the datasets after the application of the feature abundance limits in Table 2. Due to the large number of features in the datasets, a direct and complete comparison of the distribution of attributes and the correlations between them is less feasible. We thus first consider a general measure of information loss that computes a scalar, namely the distance measure *IL1s*, introduced in [24], and for continuous attributes as

$$\text{IL1s} = \frac{1}{dn} \sum_{j=1}^{d} \sum_{i=1}^{n} \frac{|x_{ij} - y_{ij}|}{\sqrt{2} S_j},$$

where d is the number of the features, n is the number of samples in the datasets, x_{ij} and y_{ij} are the values for feature j and individual i before and after anonymisation (respectively), and S_j is the standard deviation of the feature j in the original dataset. Note that this measure uses S_j as a common scale for all values of the same feature. IL1s returns values in the range $[0, \infty)$. Smaller values indicate a greater similarity between the compared datasets, and 0 is returned for identical datasets.

It is evident that with a greater anonymity level, the IL1 measure increases. The lowest information loss is reported for the Markers-type dataset, preserving the most statistical utility after anonymisation. However, the value of IL1s is generally low on all anonymised datasets.

Table 3: IL1s-measure on the anonymised datasets

k	Species	OTUs	Markers	KBW
2	0.126	0.074	0.022	0.155
3	0.165	0.096	0.030	0.169
5	0.191	0.114	0.034	0.224
7	0.201	0.121	0.036	0.240

4.3 Predictive machine learning tasks

The utility of anonymised data was tested with the predictive machine learning tasks defined for the datasets from the Knights lab microbiome machine learning repository [22]. We measure and compare the effectiveness of the machine learning models trained on the original data to those trained on the anonymised data, similar as in e.g. [27].

Table 4: Characteristics of the six datasets used from the Knights lab microbiome machine learning repository

Dataset	Task	# Samples	# Features
Gevers	control vs cd (crohn's disease), rectum	160	943
Gevers	control vs cd (crohn's disease), ileum	140	943
Morgan	healthy vs cd (crohn's disease)	128	829
Morgan	healthy vs uc (ulcerative colitis)	128	829
Turnbaugh	lean vs obese	142	557
Kostic	healthy vs tumor biopsy	172	908

In total, the repository contains data for 33 curated machine-learning tasks, mainly for binary classification. We utilised six medium-sized datasets: two for distinguishing healthy microbiome samples from those where the hosts suffer from Morbus Crohn (the "Gevers" datasets in the repository), two datasets with tasks on Inflammatory Bowel Disease (the "Morgan" datasets), one for distinguishing lean from obese individuals (the "Turnbaugh" dataset), and one for detecting tumours (the "Kostic" dataset). The details of each dataset can be found in Table 4. In each case, we will use data containing RefSeq-based OTU abundance counts to compare our results directly to the baselines shown in [22]. Furthermore, we applied the same preprocessing steps as in the original publication. Hence, our experimental setup is as follows:

(1) The OTU counts are converted to relative abundances, filtered at a minimum of 10% prevalence, and collapsed at a complete-linkage correlation of 95%.
(2) 5-fold cross-validation is applied in a stratified fashion and with regard to the control variable, if it is specified[6].
(3) Within each fold, the training samples are anonymised with μ-Ant for four levels of k-anonymity, namely $k = 2, 3, 5, 7$. All features are configured as quasi-identifiers.
(4) The relative abundances in the original and anonymised data are transformed by using the same feature abundance limits as in [10] and shown in Table 2.
(5) Finally, we train machine learning models on both original and anonymised datasets and compare their performances on the test data. We applied Random Forest (RF) with 500 estimators, Support Vector Machines (SVM) with radial basis and linear kernel, and Extreme Gradient Boosting. All other model parameters use the default values. A random state of 123 was set to make the results more reproducible.
(6) This entire process is repeated ten times, and the mean class probabilities are used to calculate the ROC-AUC score for Random Forest and XGBoost. For ROC-AUC calculations of SVMs, the decision function was used.[7]. The models' predictions are used to calculate the F1 score.

This experimental setup is as close as possible to the one first implemented by [22]. In addition, a similar setup has been used in the recent evaluation of synthetic microbiome data [11].

The application of feature abundance limits appears to improve the overall classification performance in [11]. Therefore, we also apply them in Step four inside the cross-validation loop. This is a slight adjustment to [11], where the feature abundance limits were used before the split. As a difference to both papers, Extreme Gradient Boosting[8] was added to the list of tested models. The experiment of this paper was implemented in Python, and we did not use the R implementations of in [22] and [11]. Because of this, some of the models' details and parameters differ from the original experiment. For example, the default value of sigma parameter for radial SVM in R is 0.1, whereas we observed extremely poor results with this value in the scikit-learn SVM implementation. For this reason, we used the default sigma[9] parameter for this library. Our code is available online[10]

To compare the performance of models trained on original and anonymised data, ROC curves were plotted with respective AUC scores, and colour-coded by the model type. Additionally, F1 score plots are presented in Figure 6.

We can observe only a small drop in effectiveness for anonymised data from the results on the first Gevers dataset "control vs cd, rectum" in Figure 2a. Especially Random Forest seems to be, across the experiments, quite robust to anonymisation strength, with a difference of 9 percentage points (pp) between the original AUC score and the score on $k = 7$ anonymised data. It is also the model with the best performance on anonymised data for $k = 7$. For SVM with linear kernel, the AUC score fluctuates, even obtaining a higher value for $k = 2$ than for the non-anonymised data. It suffers the smallest drop of 3 pp in performance for the highest level of anonymisation compared to the original. The biggest drop can be observed for SVM with radial kernel (16 pp), which is the model with the highest AUC on original data.

For the second Gevers dataset "control vs cd, ileum" (cf. Figure 2b), the observations are similar. The drop in performance is usually only a few percentage points for increasing anonymisation levels. RF is again the best model for the highest level of anonymisation, while XGBoost scores the lowest; it is also the most stable

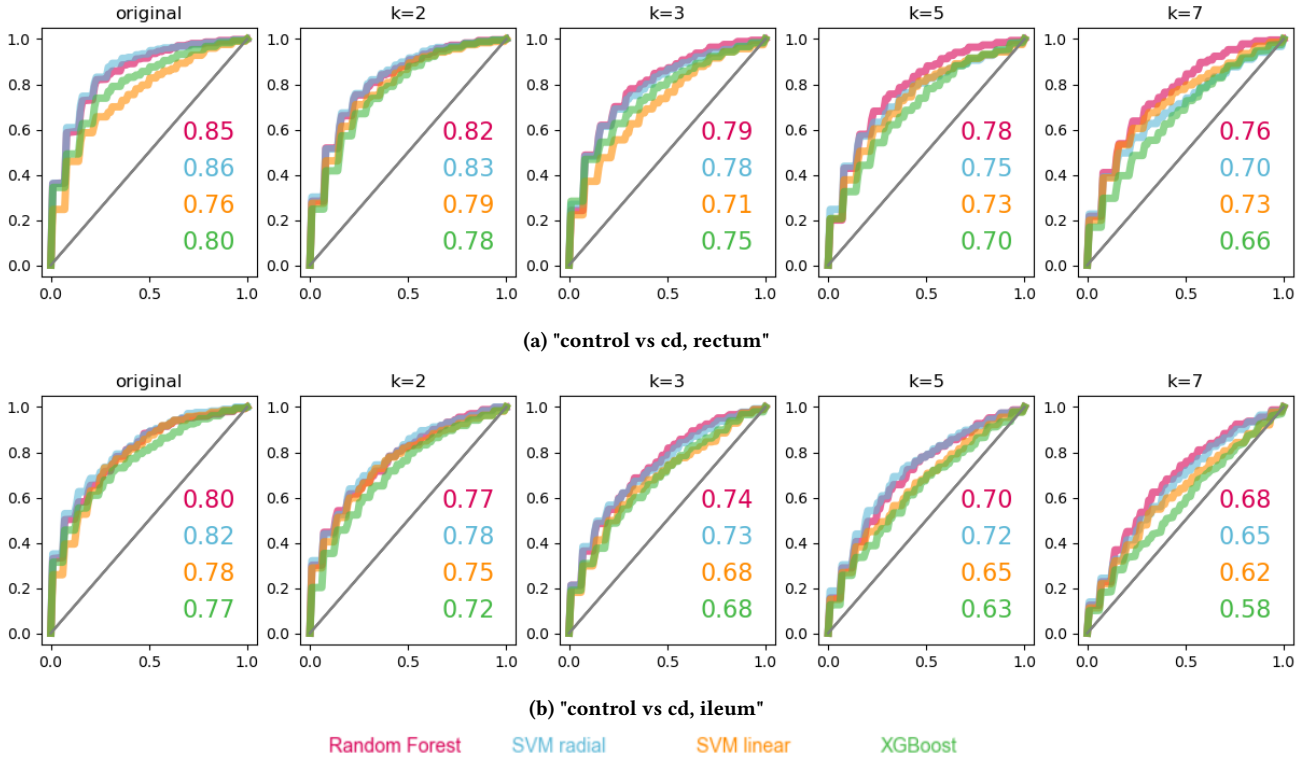
[6] In the case when the dataset contains a control variable, folds are selected such that samples with the same control variable value are contained within the same fold.

[7] The approach of using the decision function rather than probability scores to obtain confidence scores for SVM's ROC curve comes from the fact that SVMs do not output probabilities natively, but rather use Platt scaling to obtain them, which is known to have some inconsistency issues, see: https://scikit-learn.org/stable/modules/svm.html#scores-probabilities

[8] Implementation of the xgboost library for Python: https://github.com/dmlc/xgboost

[9] The sigma parameter of Support Vector Machines in sci-kit learn implementation is called gamma, see: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html?highlight=svc#sklearn.svm.SVC

[10] https://github.com/sbaresearch/microbiome-k-ano



(a) "control vs cd, rectum"

(b) "control vs cd, ileum"

Figure 2: Results on Gevers: Final ROC curves and AUC scores.

model, dropping only 12 pp in score for $k = 7$. The relative difference in AUC scores is the most striking for XGBoost (19 pp).

On the Morgan dataset ("healthy vs cd, stool" in Figure 3a), all of the models drop significantly in performance for higher levels of anonymisation. Both SVM flavours suffer the biggest falls, reaching around 50% of AUC for $k = 7$, meaning that the model is no better than randomly guessing the class. Even though the difference is bigger than for the previously described Gevers datasets, the models still obtain AUC scores of 74-80% for $k = 2$, which proves their usefulness for classification. In fact, SVM with radial kernel does not suffer any drop in score between level 2 and 3 of anonymisation. The second dataset from Morgan ("healthy vs uc, stool" in Figure 3b) results in overall similar findings as for the first Morgan dataset. Despite obtaining low AUCs on $k = 5, 7$ anonymised data, models perform reasonably well for lower levels. Both of the SVM models obtain a higher score for $k = 3$ than for $k = 2$ anonymised data.

On the Turnbaugh dataset (Figure 4a), Random Forest obtains 79% AUC on the original dataset, maintains this score for $k = 2$, and only drops by 5 pp for the highest level of anonymisation, proving that it is still a usable model. XGBoost, although still effective for lower levels of anonymisation (2 pp drop for $k = 2$), drops by 12 pp on $k = 7$, more than twice as much as RF. However, the most substantial drop for this anonymity level is observed for SVM radial basis, where 23 pp are lost compared to the original.

Results from the Kostic dataset are presented in Figure 4b. The SVM models maintain the highest AUC for $k = 7$ level of 74%. For SVM linear, this only amounts to a 5 pp drop. Random Forest is similarly robust, without substantial losses in performance, however having slightly larger, increasing pp drops than SVMs for this dataset. The AUC of XGBoost fluctuates, reaching rather low scores for anonymised data in general.

Overall, these results indicate that we can still obtain satisfying performances with the Random Forest model, even for a higher level of anonymisation. XGBoost, being the second decision tree-based algorithm, did not perform as well as Random Forest. Its effectiveness is highly dependent on tuning a large number of hyperparameters. However, since we focused on comparability and not on achieving a high baseline performance, we did not tune the parameters of any model. SVM models maintain their utility for lower levels of anonymisation, but their performance is generally less robust than the Random Forest model.

While the discussed patterns recurred through the experiments, the final effectiveness of all models certainly depends on the particular dataset and scenario. This is also reflected in the F1 scores (cf. Figure 6). For the Gevers datasets, there is a substantial decrease with higher levels of anonymisation, especially in the “control vs cd, ileum” dataset for $k = 7$. However, the bar plots show that all of the models had a particular problem on the Morgan “healthy vs cd, stool” dataset, resulting in an extremely low F1 score even on the original data. A closer look indicates that this behaviour might be caused by imbalanced data. Figure 5 shows the first two confusion matrices for Random Forest, which prove that the models predict mainly the majority class due to a heavy imbalance in

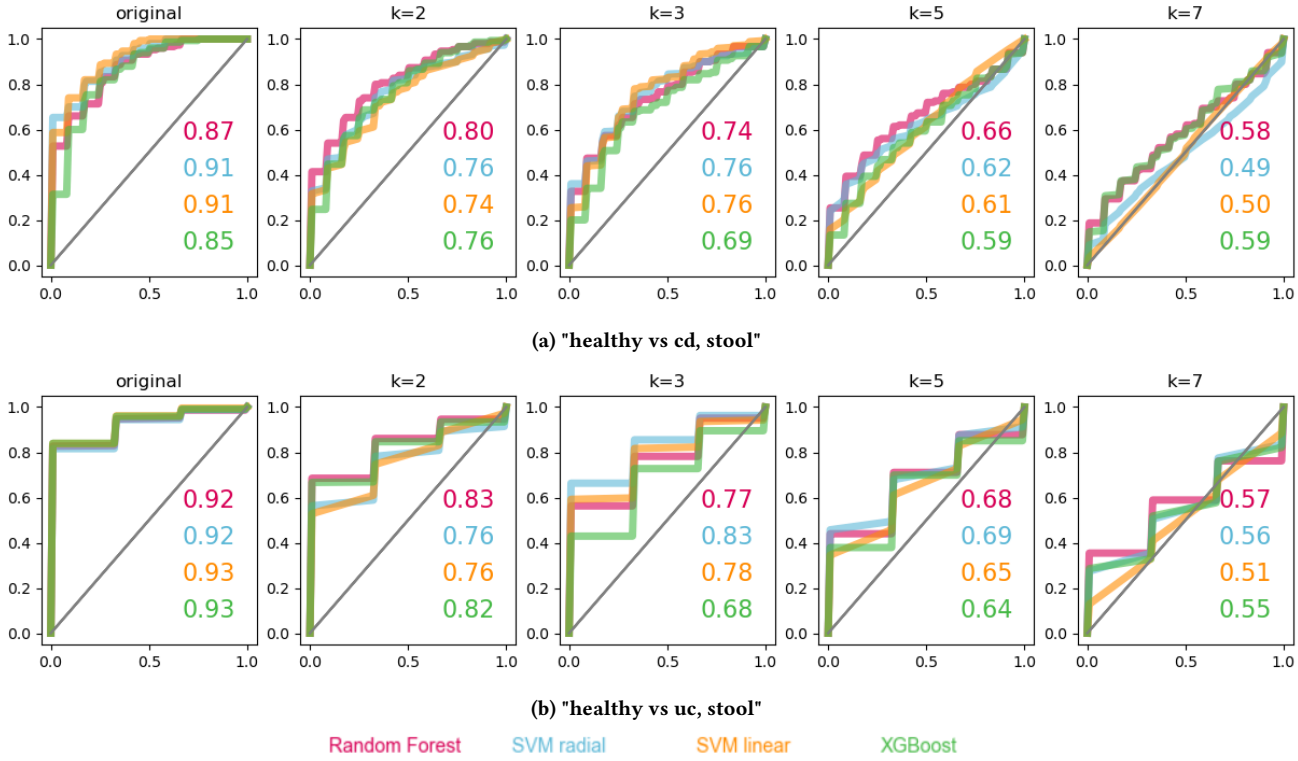


Figure 3: Results on Morgan: Final ROC curves and AUC scores.

the number of samples between classes for this dataset. For other datasets, there is only a small drop or no difference in the F1 score with anonymisation, which proves the data's utility remains high for particular tasks.

5 CONCLUSION AND FUTURE WORK

In this paper, we have discussed and evaluated k-anonymity as a privacy-preserving method for microbiome data with metagenomic features. Our experiments investigated the protection against personal microbiome identification and analysed the utility of the data after anonymisation. We have shown that most of the information is preserved from the original data while protecting individuals from direct identification via PMI methods, which is known to be a threat to paired datasets. We further evaluated the utility for predictive classification tasks. While the individual effectiveness of the models depends on the dataset and the task, we have shown that Random Forest is particularly well suited and able to maintain its performance when trained on anonymised data, even for higher levels of anonymisation.

It is certainly difficult to directly compare the utility results of the synthetization applied in [11] and our k-anonymity approach, in particular considering the varying values of k. Synthetization focuses on preserving only the global information of the original dataset and achieves privacy risk reductions by getting rid of the 1-to-1 correspondence of real individuals and synthetic samples. By contrast, anonymisation achieves privacy risk reductions by small changes to the values in the cells of the original dataset, thereby preserves more of the local information and, in principle, retains the link between individuals and samples. We therefore conclude that anonymisation is a suitable addition to the toolkit of privacy-preserving measures for microbiome data and may be particularly useful in specific practical scenarios.

Our future work will focus on the application, evaluation and comparison of both anonymisation and synthetization techniques on larger microbiome datasets, though datasets that allow for both PMI evaluation and utility evaluation are currently not available. We will study the scalability of the approaches and possible adaptations to other data formats, such as raw microbiome data in form of genetic sequences. We will further investigate how other methods commonly used in data sharing, such as watermarking and fingerprinting [19], need to be adapted to the specificities of tabular microbiome data.

ACKNOWLEDGMENTS

This work was partially funded by the Austrian Research Promotion Agency FFG under grant 877173 (GASTRIC) and by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 826078 (project 'FeatureCloud'). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

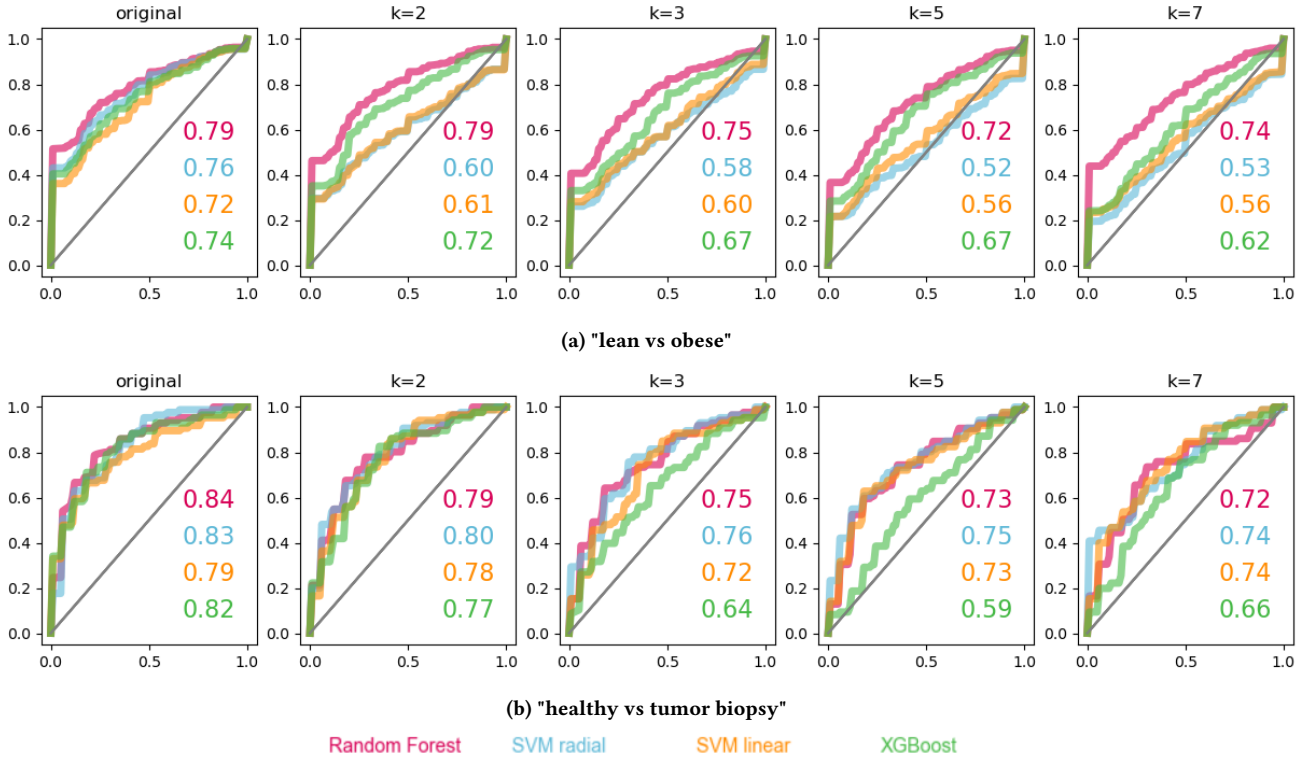


Figure 4: Results on Turnbaugh (first row) and Kostic (second row): Final ROC curves and AUC scores.

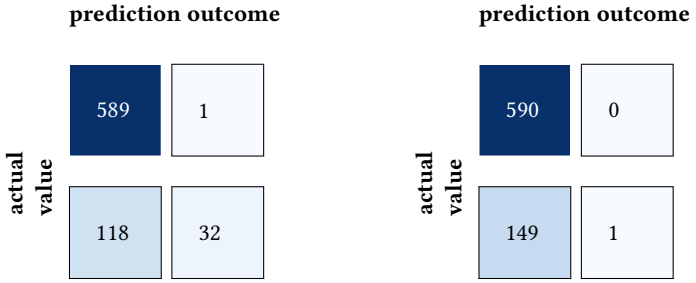
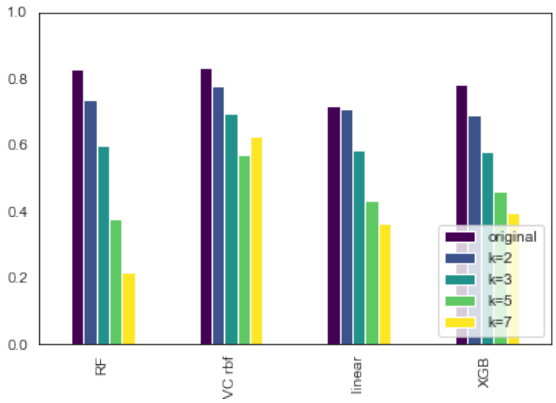
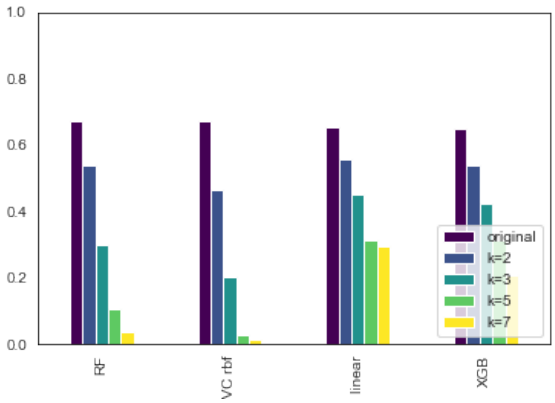


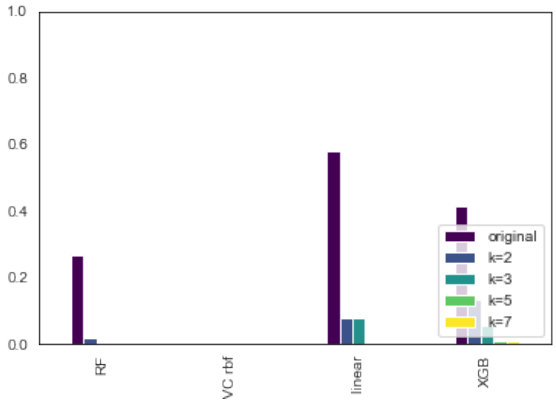
Figure 5: Confusion matrices for Morgan "healthy vs cd, stool" from Random Forest trained on original and anonymised data



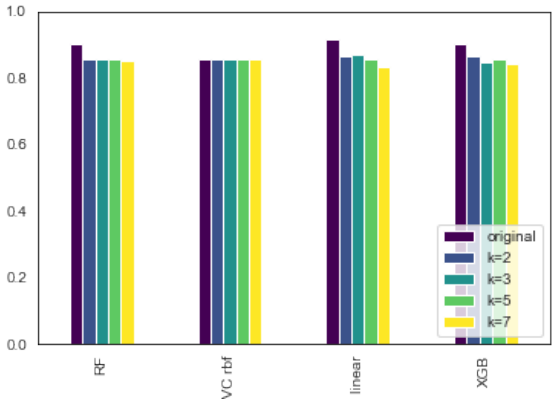
(a) Gevers "control vs cd, rectum"



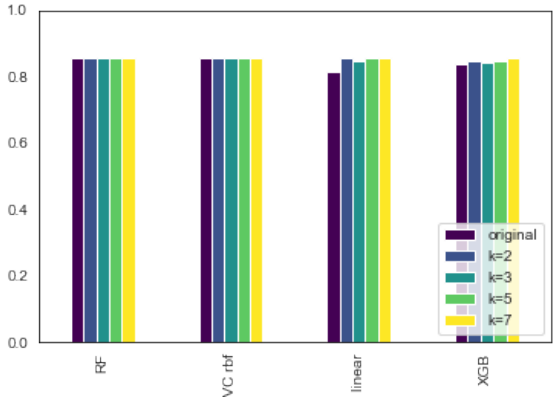
(b) Gevers "control vs cd, ileum"



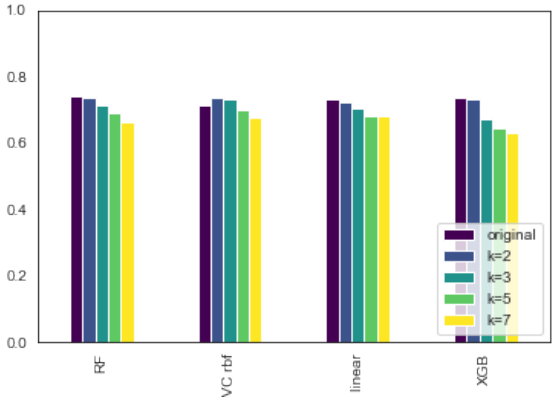
(c) Morgan "healthy vs cd, stool"



(d) Morgan "healthy vs uc, stool"



(e) Kostic "lean vs obese"



(f) Turnbaugh "healthy vs tumor biopsy"

Figure 6: Results on all datasets: F1 scores.

REFERENCES

[1] Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. 2006. Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, Chicago IL USA, 153–162. https://doi.org/10.1145/1142351.1142374

[2] Bonnie Berger and Hyunghoon Cho. 2019. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biology* 20, 1 (Dec. 2019), 128, s13059–019–1741–0. https://doi.org/10.1186/s13059-019-1741-0

[3] Eleonora Distrutti, Lorenzo Monaldi, Patrizia Ricci, and Stefano Fiorucci. 2016. Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. *World Journal of Gastroenterology* 22, 7 (Feb. 2016), 2219–2241. https://doi.org/10.3748/wjg.v22.i7.2219

[4] J. Domingo-Ferrer and J.M. Mateo-Sanz. 2002. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 14, 1 (Feb. 2002), 189–201. https://doi.org/10.1109/69.979982

[5] Diego Fernández-Edreira, Jose Liñares-Blanco, and Carlos Fernandez-Lozano. 2021. Machine Learning analysis of the human infant gut microbiome identifies influential species in type 1 diabetes. *Expert Systems with Applications* 185 (Dec. 2021), 115648. https://doi.org/10.1016/j.eswa.2021.115648

[6] Eric A. Franzosa, Katherine Huang, James F. Meadow, Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, and Curtis Huttenhower. 2015. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences* 112, 22 (June 2015). https://doi.org/10.1073/pnas.1423854112

[7] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *Comput. Surveys* 42, 4 (June 2010), 1–53. https://doi.org/10.1145/1749603.1749605

[8] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 493 (July 2022), 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

[9] Markus Hittmeir, Andreas Ekelhart, and Rudolf Mayer. 2019. On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks. In *14th International Conference on Availability, Reliability and Security (ARES 2019)*. Canterbury, United Kingdom. https://doi.org/10.1145/3339252.3339281

[10] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. 2022. Distance-based Techniques for Personal Microbiome Identification. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*. ACM, Vienna Austria, 1–13. https://doi.org/10.1145/3538969.3538985

[11] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. 2022. Utility and Privacy Assessment of Synthetic Microbiome Data. In *Data and Applications Security and Privacy XXXVI*, Vol. 13383. Springer International Publishing, Cham, 15–27. https://doi.org/10.1007/978-3-031-10684-2_2

[12] Ruth E. Ley, Peter J. Turnbaugh, Samuel Klein, and Jeffrey I. Gordon. 2006. Human gut microbes associated with obesity. *Nature* 444, 7122 (Dec. 2006), 1022–1023. https://doi.org/10.1038/4441022a

[13] Guang Li, Yadong Wang, and Xiaohong Su. 2012. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. *Computer Methods and Programs in Biomedicine* 108, 1 (Oct. 2012), 1–9. https://doi.org/10.1016/j.cmpb.2011.02.013

[14] B. A. Malin. 2005. Protecting Genomic Sequence Anonymity with Generalization Lattices. *Methods of Information in Medicine* 44, 05 (2005), 687–692. https://doi.org/10.1055/s-0038-1634025

[15] Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. 2020. Privacy-preserving Anomaly Detection using Synthetic Data. In *34th Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSec)*. Springer International Publishing, Regensburg, Germany, 195 – 207. https://doi.org/10.1007/978-3-030-49669-2_11

[16] Giovanni Musso, Roberto Gambino, and Maurizio Cassader. 2010. Obesity, Diabetes, and Gut Microbiota. *Diabetes Care* 33, 10 (Oct. 2010), 2277–2284. https://doi.org/10.2337/dc10-0556

[17] G B Rogers, D J Keating, R L Young, M-L Wong, J Licinio, and S Wesselingh. 2016. From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. *Molecular Psychiatry* 21, 6 (June 2016), 738–748. https://doi.org/10.1038/mp.2016.50

[18] P. Samarati. 2001. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering* 13, 6 (Dec. 2001), 1010–1027. https://doi.org/10.1109/69.971193

[19] Tanja Sarcevic, Rudolf Mayer, and Andreas Rauber. 2022. Adaptive Attacks and Targeted Fingerprinting of Relational Data. In *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, Osaka, Japan, 5792–5801. https://doi.org/10.1109/BigData55660.2022.10020266

[20] Yancheng Shi, Zhenjiang Zhang, Han-Chieh Chao, and Bo Shen. 2018. Data Privacy Protection Based on Micro Aggregation with Dynamic Sensitive Attribute Updating. *Sensors* 18, 7 (July 2018), 2307. https://doi.org/10.3390/s18072307

[21] David Sánchez, Sergio Martínez, Josep Domingo-Ferrer, Jordi Soria-Comas, and Montserrat Batet. 2020. μ *-ANT* : semantic microaggregation-based anonymization tool. *Bioinformatics* 36, 5 (March 2020), 1652–1653. https://doi.org/10.1093/bioinformatics/btz792

[22] Pajau Vangay, Benjamin M Hillmann, and Dan Knights. 2019. Microbiome Learning Repo (ML Repo): A public repository of microbiome regression and classification tasks. *GigaScience* 8, 5 (May 2019). https://doi.org/10.1093/gigascience/giz042

[23] Zicheng Wang, Huazhe Lou, Ying Wang, Ron Shamir, Rui Jiang, and Ting Chen. 2018. GePMI: A statistical model for personal intestinal microbiome identification. *npj Biofilms and Microbiomes* 4, 1 (Sept. 2018), 20. https://doi.org/10.1038/s41522-018-0065-2

[24] William E. Yancey, William E. Winkler, and Robert H. Creecy. 2002. Disclosure Risk Assessment in Perturbative Microdata Protection. In *Inference Control in Statistical Databases*. Vol. 2316. Springer Berlin Heidelberg, Berlin, Heidelberg, 135–152. https://doi.org/10.1007/3-540-47804-3_11 Series Title: Lecture Notes in Computer Science.

[25] Yongzhi Yang, Lutao Du, Debing Shi, Cheng Kong, Jianqiang Liu, Guang Liu, Xinxiang Li, and Yanlei Ma. 2021. Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications* 12, 1 (Nov. 2021), 6757. https://doi.org/10.1038/s41467-021-27112-y

[26] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. 2012. Human gut microbiome viewed across age and geography. *Nature* 486, 7402 (June 2012), 222–227. https://doi.org/10.1038/nature11053

[27] Tanja Šarčević, David Molnar, and Rudolf Mayer. 2020. An Analysis of Different Notions of Effectiveness in k-Anonymity. In *Privacy in Statistical Databases*, Vol. 12276. Springer International Publishing, Cham, 121–135. https://doi.org/10.1007/978-3-030-57521-2_9