Secure Collaborative Data Sharing for Enhanced Machine Learning Insights

Sam Evans
sam.evans@os3.nl

February 2024

Abstract

This paper explores how Local Differential Privacy can be used as a privacy-preserving mechanism for Feature Selection - and how this interacts with Data Generalisation to produce private, high-utility data from distributed individual data-holding sources. A novel framework is proposed, implemented and tested which integrates the aforementioned mechanisms with a Machine Learning Classification Model. The proposed framework is shown to be capable of handling varying privacy requirements, and is shown to preserve the utility of privatised data by evaluating classifier performance on privatised and unprivatised training and test data sets. It is shown that there are minor (maximum 2.5%) degredations in the utility of processed data when compared with the utility of unprocessed data, and in some cases the processed data exceeds the performance of the unprocessed data.

A complete repository of code and all data collected can be found at https://gitlab.os3.nl/sevans/ldp-fs-dg-ai.

***Index Terms* - Local Differential Privacy, Distributed Data, Data Generalization, Classification Model**

1 Introduction

In our current world, data is gold. Organisations and individuals alike produce and collect data to evaluate and advance their practices. Today, much of data collected in bulk is put towards the continued improvement of artificial intelligence models [17] - as the increase in volume of accurate training data in turn allows for more effective training of the models [19].

Where there is demand for data, there is a requirement for said data to be shared between both individuals and organisations [8]. Likewise, where there is a demand for data to be shared, there will exist an adversarial demand for privacy. Such adversarial demands may be due to fears of organisational overreach when collecting bulk data, or due to fears of data being stolen in one of the many ever-present data breaches and cyberattacks organisations and governments suffer in current times [7]. The result of this dynamic is the ongoing development of privacy-preserving techniques - designed to mitigate the amount of information shared between parties while maximising the utility of the resultant data. These techniques are important to enable organisations and individuals the opportunity to collect and share data without compromising varying individual privacy requirements [5].

Two of such privacy-preserving techniques which will be investigated in detail in this paper are *Data Generalisation* [21] and *Local Differential Privacy (LDP)* [9]. Data Generalization is a process where a point of data is replaced by a more general value of the same data. The result is that the completed record is now far more difficult to uniquely identify - thus increasing the privacy of the data. For example, the feature *age* recorded as *21* could be generalised to $20 < x \leq 25$ and then again to *any.* The result is that an entry which could otherwise be distinguishing i.e. '21 year old electrician' becomes one of many similar records i.e. '21-25 year old electricians' which inherently makes the initial data point less distinguishable from the remaining data. This technique can again be used to turn the data into 'electricians of any age'.

Local Differential Privacy (LDP) mechanisms leverage randomized response techniques developed by Warner (1965) [22] where individual respondent's data cannot be attributed to themselves - for example in cases where individual confidentially is important. Instead, LDP permits an aggregating party to study the overall trends of a dataset without being able to correctly identify individual record. A simplified example of this could be that an untrusted aggregating party asks "Did you eat the cookie?", the respondent may then flip a fair coin secretly and must answer truthfully if heads is flipped- and must simply answer "Yes" if tails is flipped. In this case, the number of true "Yes" responses can be estimated and a general picture of the overall data can be acquired while still providing plausible deniability for any one individual's response.

To the best of the author's understanding, the two aforementioned privacy-preserving mechanisms do not exist together in one framework. This paper intends to combine the two mechanisms together, introducing a novel framework combining the characteristics of LDP, Feature Selection and Data Generalisation.

1.1 Motivating Scenario

This paper considers a scenario in which any number of individuals have a shared desire to aggregate their data without compromising the data which they themselves hold. The aggregated data would then be published and evaluated in an artificial intelligence classification model. The created framework would run on the individual's machines as well as on an aggregating node. The framework is then able to facilitate the private collection of the data, as well as the aggregation and subsequent training of AI models. It is assumed that data is distributed horizontally - meaning that each data-holding individual holds data on each of the features. It is also assumed that each individual distributes just one record to the aggregating node. Each data holder would only agree to publish their data if a certain privacy threshold is met, and every data holder must agree for the data to be published at all. This means that every data holder's privacy threshold must be met for the data to be published. Finally, it is assumed that the data-holding-individuals are honest-but-curious, meaning that they will act according to the rules of the system but wish to learn as much information about other individuals in the system as possible.

1.2 Contributions

This paper intends to contribute to the current academic understanding of the implications of LDP on the utility of collected data. It will do this by applying LDP-processed and Data Generalisation data to a common modern-day use-case, being machine learning. The proposed model incorporates Local Differential Privacy, Data Generalization and Feature Selection to optimise the privacy and utility of a processed dataset to be processed by an artificial intelligence. Performance of the model will be measured through the performance of the artificial intelligence classification model when classifying novel data. Furthermore, the privacy of the resultant features will be evaluated against the initial dataset, to determine the overall gain in privacy of the system. The aforementioned investigatory avenues are explored through this paper's answering of the following research questions:

RQ1: To what extent does the proposed framework ensure the accuracy of classifiers trained on the consolidated dataset generated through collaboration among distributed entities?

RQ2: In what ways does the framework contribute to privacy improvement for each participating entity?

1.3 Paper Outline

The remainder of this paper is organised as follows. Section 2 introduces and explains preliminary concepts which are investigated in further detail later in the study. In the next Sections 3 and 4, the experimentation method and setup are presented. Section 5 presents the results, evaluates them, Section 6 discusses related work in this field, finally Section 7 outlines this paper's conclusions and Section 8 outlines potential future work according to this paper's findings.

2 Preliminary

This section will introduce and explain the key concepts which will be explored in more detail in the rest of this paper. It will explain Local Differential Privacy and frequency estimation, Data Generalisation, privacy/utility evaluation techniques and machine learning classification models.

2.1 Local Differential Privacy

Local Differential Privacy (LDP) [9] is a privacy-preserving mechanism. That is to say, where data is inserted an LDP mechanism, the mechanism is able to create a dataset which preserves the privacy of each individual record, without excessively sacrificing the utility of the resultant dataset.

LDP allows data-holding individual clients to perturb their information locally, before sending it to an aggregating party, this perturbation mechanism is governed by a specific algorithm which the aggregating party knows. Using the example outlined in section 1, provided that the aggregating party knows that a respondent must flip a fair coin to determine their response they are able to immediately calculate the the number of true responses by removing half of the total responses from the "yes"response count. Similarly, in a more complex system an aggregating party would still be able to estimate the frequency to which responses are within given

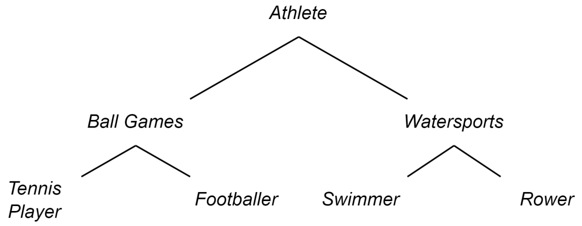


Figure 1: Taxonomy Example for "Athlete"

ranges for each response without being able to accurately guess the responses of any one individual. This is called "frequency estimation".

LDP systems use perturbation mechanisms coupled with frequency estimation in order to preserve the privacy of individual record-holders, while still gaining utility from analysis on the whole dataset.

2.2 Data Utility

The "utility" of a dataset can be measured as one's ability to gather usable information from a dataset. This process is simplified with the introduction of artificial intelligence. It is straightforward to train an artificial intelligence on data - and to then instruct it to classify otherwise unseen data. The results of the classifications performed by the artificial intelligence can be used to determine the utility [3] of the training data provided to the dataset- this could be done through statistical evaluation of the resulting estimated classifications. For example, the accuracy of the predicted classification labels - see section 5 - can be used to evaluate the utility of the training dataset.

2.3 Data Generalisation

Data Generalisation [21] is a privacy-preserving mechanism. It is the process by which data is simplified in accordance to the data's taxonomy. The objective of data generalisation is to increase the total records which contain the same information - while still retaining some detail and therefore utility, this process increases the overall privacy of a dataset. For example, using the taxonomy outlined in figure 1, a data point distinguishing a "Swimmer" could be generalised to "Watersports", and then again to "Athlete". If more generalisation were necessary, the record could be further generalised to "Any". This process increases the likelihood that there are records containing identical data, for example, an adult swimmer and an adult rower could both be generalised into adult athletes. The process increases the privacy of their data by reducing the number of features which would uniquely identify individual data points - but still maintains some utility of any individual record.

2.4 Privacy and Utility

This section defines equations which are defined in [12] and [3], where - similar to this paper - features are evaluated by privacy and utility to determine data generalisation. As such, similar - if not identical - mathematics must be completed to properly evaluate the performance of features. In order to determine which features are to be generalised in a given dataset, each feature in the dataset must be evaluated for both their privacy and their utility. Once the privacy and utility of a feature is calculated, a privacy/utility trade-off score can be found for each feature. This enables a system to determine a candidate feature whose generalisation would gain the most privacy whilst losing the least utility. The trade-off score is determined by the weighted sum of the feature's privacy and utility scores as defined by:

$$t(f) = w_p \cdot \rho(f) + w_s \cdot g(f) \tag{1}$$

Where $w_p + w_s = 1$. This can be tuned for results which favour privacy over utility, or vice-versa - but for the purposes of this project as in [12] the weights will be kept at $w_p = w_s = 0.5$.

The privacy score can be defined by the entropy of a feature given by:

$$\rho(A_j) = -\frac{1}{M(A_j)} \sum_{k=1}^{|A_j|} p(v_k \cdot A_j) \cdot \log(p(v_k \cdot A_j)) \tag{2}$$

Where $|A_j|$ is the number of values in the feature A_j, $p(v_k \cdot A_j)$ represents the proportion of the number of records which match the k'th value of the j'th attribute. $M(A_j)$ is the maximum entropy that any feature j can have - where the uncertainty of the feature is at its highest. By dividing by $M(A_j)$ this normalises the privacy score as being within the range [0, 1].

The utility score of a feature can be defined as the information gain of a feature, given by:

$$g(A) = H(C) - H(C|A) \tag{3}$$

Where $H(C)$ is the *Shannon Entropy* and $H(C|A)$ is the *Conditional Entropy*. The conditional entropy of a feature A with the set of values $A = \{v_1, v_2, ..., v_{|A|}\}$ and class labels $C = \{C_1, C_2, ..., C_m\}$ is defined as:

$$H(C|A) = -\sum_{v_k \in A} \sum_{C_i \in C} p(v_k \cdot A, C_i) \cdot \log(p(C_i|v_k \cdot A)) \tag{4}$$

Where $p(v_k \cdot A, C_i)$ is the proportion of elements of the k'th value of A with class label C_i.

The *Shannon Entropy* is defined as:

$$H(C) = -\sum_{C_i \in C} p(C_i) \log(p(C_i)) \tag{5}$$

Where $p(C_i)$ is the proportion of records in D labeled C_i.

The privacy score of the entire dataset D can be described by:

$$\rho = \frac{1}{t} \sum_{j=1}^{t} \rho(A_j) \tag{6}$$

Where feature $A_j \in D, (1 \leq j \leq t)$. This represents the average of all of the privacy scores of all of the features in the dataset D.

2.5 Machine Learning Classification

The measure of a whole dataset's utility is simplified through the use of Machine Learning (ML). This is because the overall performance of a ML model can be measured, as opposed to statistical measures of the entire dataset like information gain or entropy.

A classification ML model accepts training data which consists of features, and targets. Features can be considered the data which is collected into the dataset, the target being the measured outcome of the features. For example, the features could be a measure of weight, BMI, income, gender, etc. where the target is whether an individual has been diagnosed with Diabetes or not. The classification model is trained on feature/target pairs to be able to classify a previously unseen feature set as being a target value.

In this case, generalised values are passed into the model, as well as ungeneralised values - the goal being to understand how the performance of the model differs between the LDP generalised data, and ungeneralised raw data. These metrics proportional to each other will act as a heuristic for the proportional loss in utility of the output data which the process causes. The performance of the model trained on generalised data, compared with the performance of the model trained on ungeneralised data allows calculation of the proportional loss in utility that the framework causes to the dataset.

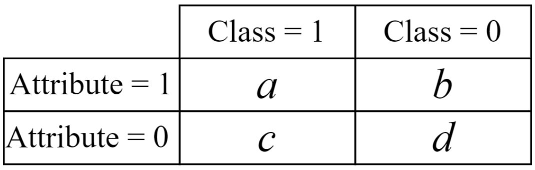
3 Method and Architecture

This section describes the experimentation setup and organisation, it will be split into two sections, denoting whether the operation would be conducted client-side (the data-holding individuals) or server-side (the aggregator) in a real-life situation.

3.1 Client

A client holds one record in a distributed system. It is assumed that a single client holds data on all features in the record, and that the clients are honest-but-curious - which means that the clients will follow any rules they are given however will attempt to learn information from any system-shared data. It is further assumed that the clients are perfectly reliable and accurate in their operations, and can accurately pass data instantly to an aggregating node.

Each client has a shared interest in aggregating their data, but wishes for their privacy requirements to be met before aggregation can take place. They wish to do this through generalisation but realise that if each client individually generalises their data, they cannot guarantee that the resulting aggregated data will be homogeneous. In order to coordinate generalisation, clients need an aggregating node to determine where the generalisation will take place. For this coordination, the aggregating node (which is untrusted by the clients) needs the client's data but they are unwilling to share. Before sharing, clients can firstly guarantee their own privacy by performing LDP - and perturbing their data before sending it to an aggregating node.



	Class = 1	Class = 0
Attribute = 1	a	b
Attribute = 0	c	d

Figure 2: Feature Selection Estimations

LDP-perturbed data lends itself specifically to frequency estimation - which means that it is usable for feature selection as outlined by Alishahi et al. (2022) [1]. LDP removes the overall structure of the data, only allowing for an estimation of the number of records of any one value in the whole dataset - with highly inaccurate results if only looking at one record. If classifiers were tained on this type of data, the classifier would be able to perform classifications according to this one dataset. However, created classifiers would only be useful for the dataset for which they have been produced - and the framework would therefore only be producing data with utility for this one specific purpose. The proposed framework's goal is to preserve the data's utility as much as possible, and create a general solution for all classification problems. For accurate machine learning classification of the type investigated in this paper to take place, the structure of the data must be preserved. Because clients aren't willing to distribute their structured data until they can be guaranteed that their privacy requirements have been met by data generalisation first, the framework needs to facilitate the privatisation of their data first.

Once the server has evaluated the features and calculated their overall privacy/utility trade-off scores, the clients will receive instructions on which data to generalise next. Each client generalises their data, in this experiment's case the data will be only binary and will be subsequently generalised to "any". The clients will then perturb their newly generalised data and send the perturbed data to the aggregating node for another round of frequency estimation and subsequent generalisation instructions to be received.

The aggregating node will determine the overall privacy of the system, and once it has exceeded a given value, it will instruct the clients to publish their generalised data publicly - and forward it to an AI. At this point, the clients can be confident that their data is private - given that it is sufficiently generalised, and that each of their data points are homogeneous because the process has been facilitated by an aggregating node.

3.2 Server

The aggregating node's primary responsibility is to determine which features to generalise. In order to do this it needs to be able to assess the utility and privacy of any given feature. The minimum amount of data necessary for this to be possible is the counts of each feature-attribute pair and which class the feature fits into. Fortunately, this can be gathered through feature estimation, which means that client nodes can simply send LDP-perturbed data to the untrusted aggregator, and the aggregator can evaluate all the information it needs from just this.

Once the aggregator receives each client's LDP-perturbed data, it can perform feature estimation. For every feature it will estimate four values, as outlined in Figure 2. This will allow the aggregator to evaluate the utility and privacy of each of the features as outlined in section 2. The aggregator calculates the privacy/utility trade-off score as in equation 1, and finds the worst-performing candidate feature, it then declares this feature to the clients as the feature to be generalised. The process then restarts, with the newly-generalised perturbed data being perturbed and sent to the aggregator to be estimated once more. During the privacy/trade-off score, the aggregator also calculates the overall privacy score, which is the average of all of the privacy scores of the features in the dataset, as given by equation 6, if this overall privacy score exceeds a certain value, then the dataset's current level of privacy is deemed sufficient, and this information is declared to each of the clients.

Once the privacy fulfillment has been declared to the clients, they can each send their generalised data to the aggregator, who is able to freely use it. In this experiment's case, the aggregator will use the newly received data as training data in a classification model in order to estimate the overall utility of the generalised data. The accuracy of the models as well as their sensitivity and specificity will be collected and evaluated to determine the performance of the models when compared with ungeneralised models, this is explained in detail in section 5. A full diagram of the process flow can be seen in Figure 3.

4 Experiments

In order to effectively evaluate the proposed architecture in practice, an appropriate example use-case for the model has been selected - and will be used to determine the overall strength of the proposed system's performance.

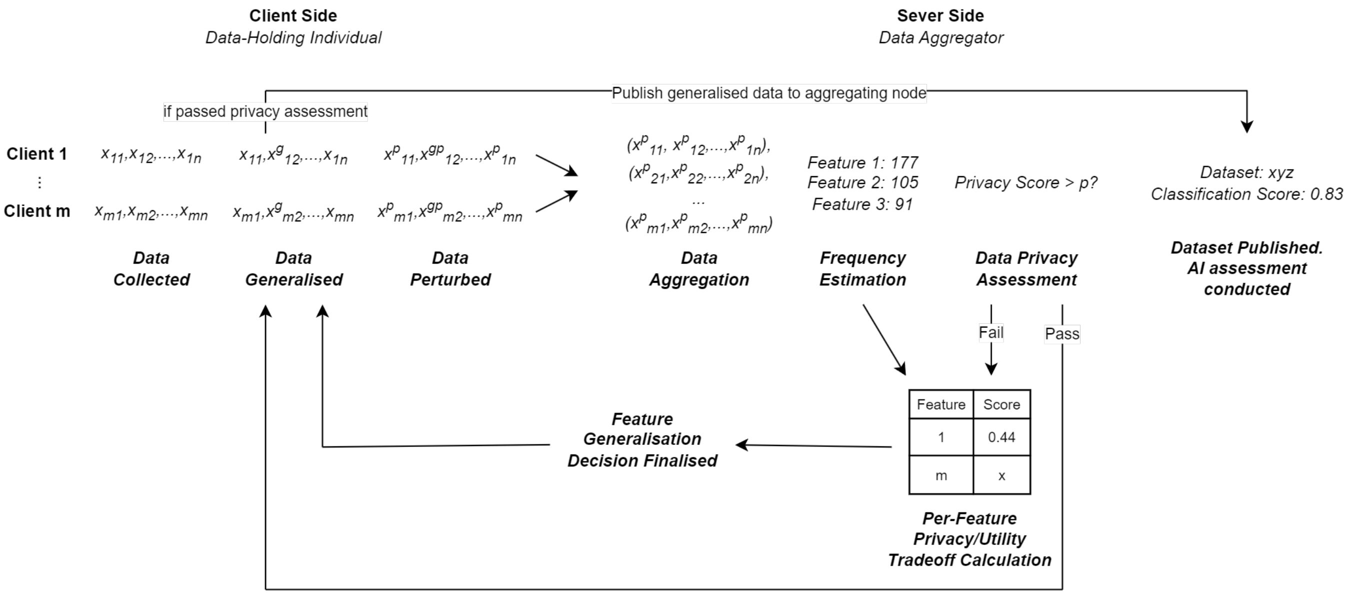

Client Side
Data-Holding Individual
Sever Side
Data Aggregator
Publish generalised data to aggregating node
if passed privacy assessment
Client 1
⋮
Client m
$x_{11}, x_{12}, \ldots, x_{1n}$
$x_{m1}, x_{m2}, \ldots, x_{mn}$
$x_{11}, x^{g}_{12}, \ldots, x_{1n}$
$x_{m1}, x^{g}_{m2}, \ldots, x_{mn}$
$x^{p}_{11}, x^{gp}_{12}, \ldots, x^{p}_{1n}$
$x^{p}_{m1}, x^{gp}_{m2}, \ldots, x^{p}_{mn}$
$(x^{p}_{11}, x^{p}_{12}, \ldots, x^{p}_{1n}),$
$(x^{p}_{21}, x^{p}_{22}, \ldots, x^{p}_{2n}),$
...
$(x^{p}_{m1}, x^{p}_{m2}, \ldots, x^{p}_{mn})$
Feature 1: 177
Feature 2: 105
Feature 3: 91
Privacy Score > p?
Dataset: xyz
Classification Score: 0.83
Data Collected
Data Generalised
Data Perturbed
Data Aggregation
Frequency Estimation
Data Privacy Assessment
Dataset Published. AI assessment conducted
Fail
Pass
Feature
Score
1
0.44
m
x
Feature Generalisation Decision Finalised
Per-Feature Privacy/Utility Tradeoff Calculation

Figure 3: Process Flow

In the use-case selected for this experiment, 30,000 individuals each hold one complete record from the *CDC Diabetes Health Indicators* [2] dataset containing personal information about ”their” health and lives and hold an identifier to declare whether or not they have been diagnosed as diabetic (or prediabetic), or not. The individuals have a shared interest in this information being aggregated but are unwilling to compromise their own privacy - in this experiment this is simulated by individual privacy requirements “p”, which will be tuned to observe the difference in the utility of the resultant data. It is similarly simulated by the individual ε values, which are used to tune how private the LDP output is. The values selected for this experiment and the results collected are: p *= 0.6, 0.7, 0.8, 0.9*, ε *= 1, 2, 3, 4, 5.*

The LDP model selected for the experiments is the RAPPOR model proposed by Google [4]. It provides a strong balance between frequency estimation performance and privacy retention, as well as security and performance in transmission of data in practical applications. The model was selected on the grounds that in practical implementations of the model proposed in this paper, many nodes would have to run client-side RAPPOR operations (in much the same way Google must with their ”chrome” browser’s data collection methods). As such, RAPPOR is optimised for this purpose.

For the purposes of experimentation, the 30,000 records are already on the same local machine, and the process of perturbing them is completed per-record locally. The resultant data of this process remains the same as if the process was completed on a distributed set of nodes, and the perturbed data sent to the aggregating node. Once the data is aggregated, the relevant 4 counts are estimated for each feature - and a privacy/utility tradeoff score for each value is calculated. If the overall privacy score exceeds the predefined score, the generalised features along with their relevant targets are forwarded to a set of AI classification models.

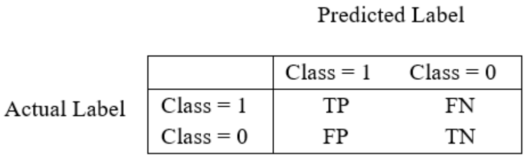
The models selected for this project are Random Forest, Decision Tree and K-Nearest-Neighbour. This is because all three are shown to be very effective at classifying binary data, such as the data which will be given as an input to these models. Furthermore, under normal use cases the models are very high-performance, meaning that multiple experiments and passes can be performed to ensure both accurate results and that a wide range of privacy parameters can be tested.

Finally, the performance of the AIs is evaluated through their ability to accurately classify unseen records from the 30,000 - these will be split into testing and training datasets (25% to 75%), where the training feature/target pairs will be fed to the model to fit itself to - and then unseen testing features will be fed to the fitted model, to be estimated - the estimates compared against the testing targets. The accuracy, sensitivity and specificity of the models will be collected and collated for analysis.

The experiments were run using Scikit-learn’s machine learning models [15] on Pycharm [16], on a Windows 11 laptop, with a *Ryzen 9 6900HS Creator Edition* CPU, 32 GB RAM @6400MHz, and an *RTX 3050 Laptop GPU.*

5 Results

One important evaluation data structure used in AI is a confusion matrix given which can be seen in figure 4. Each value in the confusion matrix represents a classification model’s guess and the true value for each of the



Predicted Label

Actual Label		Class = 1	Class = 0
	Class = 1	TP	FN
	Class = 0	FP	TN

Figure 4: A Confusion Matrix

guesses. For example, TP represents the "True Positive", where the class label for the test data was 1, and the classification model also stated that the label was 1, whereas FN represents the "False Negative", where the class label for the data was 1, but the classification model stated that the label was 0. These values are used to draw specific data about the model's ability to classify data.

The sensitivity of a model can be given as:

$$\frac{TP}{TP + FN}$$

Sensitivity represents the ratio of true positives to the total actual positives in the data. It is used in cases where false negatives in datasets are unacceptable. This is one of the metrics used in the evaluation of the results.

The specificity of a model can be given as:

$$\frac{TN}{TN + FP}$$

Specificity represents the ratio of true negatives to total negatives in the data. It is generally used where false positives are unacceptable in the final data output.

Finally, the accuracy of a model can be given as:

$$\frac{TP + TN}{TP + FP + FN + TN}$$

It represents the proportion of correctly classified data points compared to the total number of data points.

5.1 Ungeneralised Results

For comparison and results to be drawn from the accuracy, sensitivity and specificity values collected from the generalised data in the models, first it is important that a baseline from ungeneralised data is recorded. Table 1 shows the accuracy collected from ungeneralised data being passed into each of the classification models.

	Random Forest	Decision Tree	K-Nearest-Neighbour
Ungeneralised Data	0.7282	0.7186	0.68

Table 1: Ungeneralised Data Testing Accuracy

Table 2 shows the sensitivity, and specificity of the ungeneralised data in each of the models. These will be used as a baseline comparison point, to determine the loss in utility and privacy caused by the generalisation techniques - as the data fed into the model for these results is meant to be the least "private" data, while holding the highest "utility".

	Random Forest	Decision Tree	K-Nearest-Neighbour
Sensitivity	0.83	0.82	0.74
Specificity	0.61	0.6	0.6

Table 2: Ungeneralised Data Sensitivity and Specificity

5.2 Measured Accuracy Results

Tables 3, 4 and 5 represent the recorded testing accuracy of each of the models where the p and ε values (denoted by E in the tables) are changed - that is the proportion of correctly classified values when compared with their true labels. The tables have been coloured according to the proportion of values recorded compared to the ungeneralised testing accuracy as outlined in table 1 - where a darker shade of green represents closer to (or exceeding) the ungeneralised data accuracy.

Random Forest	E				
P	1	2	3	4	5
0.6	0.72	0.718	0.711	0.7226	0.718
0.7	0.71013	0.705	0.71253	0.72	0.7116
0.8	0.7228	0.72	0.72106	0.7192	0.71546
0.9	0.728	0.711066	0.7244	0.7156	0.726

Table 3: Generalised Data, Random Forest Test Accuracy

Decision Tree	E				
P	1	2	3	4	5
0.6	0.7144	0.711	0.7029	0.71453	0.71213
0.7	0.7092	0.698	0.7128	0.713	0.70853
0.8	0.7144	0.7148	0.71626	0.7113	0.7126
0.9	0.7264	0.70506	0.7152	0.711466	0.7186

Table 4: Generalised Data, Decision Tree Test Accuracy

KNN	E				
P	1	2	3	4	5
0.6	0.65	0.68	0.6828	0.6956	0.692
0.7	0.6672	0.688	0.6858	0.69	0.686
0.8	0.6785	0.6817	0.6888	0.68086	0.6796
0.9	0.70706	0.68333	0.68226	0.671866	0.69346

Table 5: Generalised Data, K-Nearest-Neighbour Test Accuracy

5.3 Measured Sensitivity and Specificity Results

The results shown in tables 6, 7 and 8 are the measured recordings of the sensitivity and specificity of the model's performance under different individual privacy requirements "p" and ε values represented by "E". The colours of the cells of the tables have been changed to represent their value as a proportion of the values calculated on the performance of models with ungeneralised data. A darker shade of yellow represents calculated sensitivity scores closer to or exceeding the sensitivity value of the corresponding model's performance under ungeneralised conditions as per table 2. Similarly, a darker shade of green represents calculated specificity scores closer to or exceeding the specificity value of the corresponding model's performance under ungeneralised conditions as per table 2.

Random Forest	Sensitivity					Specificity				
	E									
P	1	2	3	4	5	1	2	3	4	5
0.6	0.84	0.77	0.83	0.81	0.79	0.59	0.66	0.58	0.62	0.63
0.7	0.8	0.81	0.82	0.76	0.81	0.61	0.58	0.59	0.67	0.59
0.8	0.76	0.81	0.81	0.82	0.8	0.68	0.62	0.62	0.6	0.61
0.9	0.76	0.8	0.79	0.77	0.81	0.69	0.61	0.65	0.65	0.62

Table 6: Generalised Data, Random Forest Sensitivity and specificity

Decision Tree	Sensitivity					Specificity				
	E									
P	1	2	3	4	5	1	2	3	4	5
0.6	0.82	0.79	0.82	0.83	0.81	0.59	0.62	0.57	0.59	0.59
0.7	0.82	0.81	0.83	0.82	0.83	0.58	0.57	0.58	0.59	0.57
0.8	0.82	0.82	0.82	0.82	0.81	0.6	0.59	0.6	0.58	0.59
0.9	0.82	0.79	0.79	0.8	0.82	0.61	0.61	0.63	0.61	0.6

Table 7: Generalised Data, Decision Tree Sensitivity and specificity

KNN	Sensitivity					Specificity				
	E									
P	1	2	3	4	5	1	2	3	4	5
0.6	0.73	0.71	0.74	0.74	0.72	0.57	0.66	0.61	0.65	0.66
0.7	0.7	0.73	0.73	0.74	0.75	0.63	0.65	0.64	0.64	0.61
0.8	0.66	0.71	0.77	0.7	0.7	0.7	0.65	0.6	0.66	0.65
0.9	0.78	0.73	0.67	0.74	0.72	0.62	0.63	0.7	0.59	0.67

Table 8: Generalised Data, K-Nearest-Neighbour Sensitivity and specificity

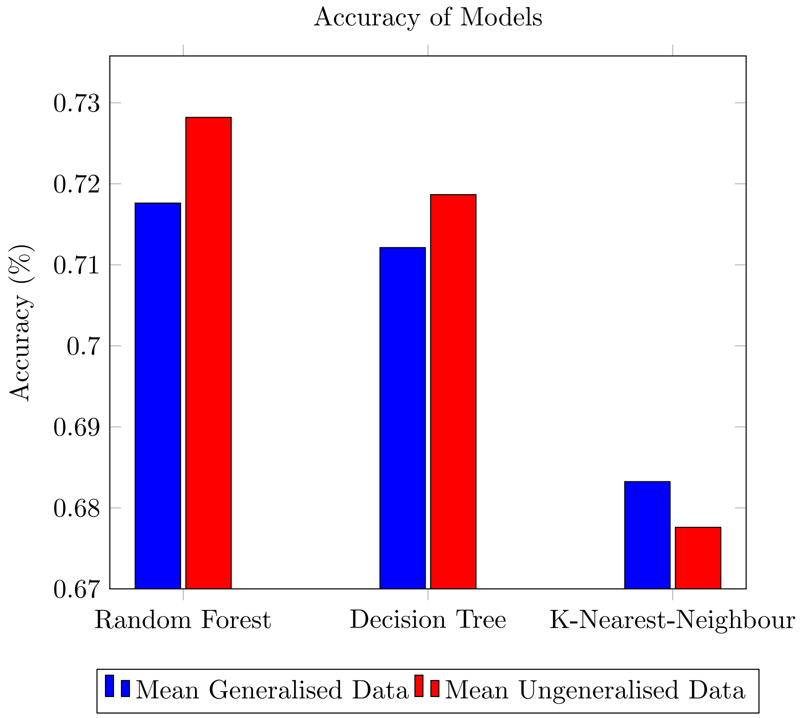
5.4 Discussion

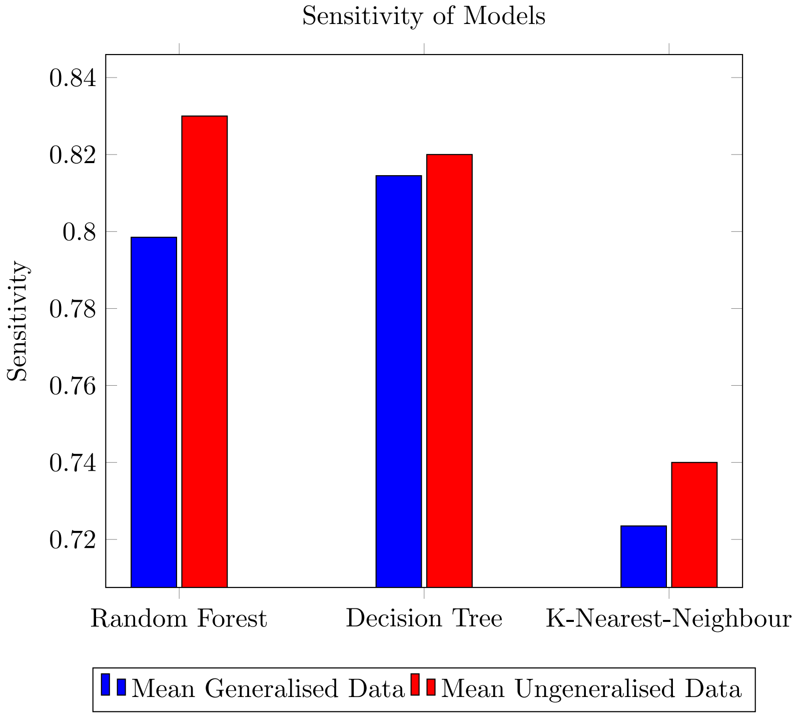
The results calculated allow one clear conclusion to be drawn, which is that the process of LDP Feature Selection coupled with Data Generalisation does not notably take away from the utility of the final data in any meaningful or predictable way. That is to say, the performance of the models under different generalised conditions is high relative to the same models with ungeneralised data. This could be due to a number of reasons.

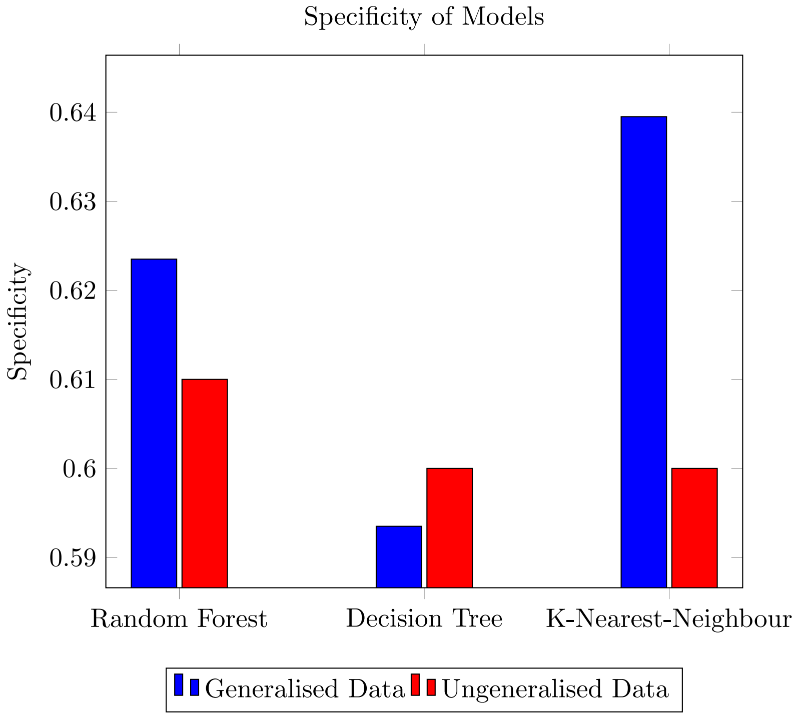
Firstly, the trade-off process could be removing "noise" in the final dataset by generalising data, that is to say that the classification models are more easily able to see the data which will bring them the most utility - which means that the higher privacy levels which prioritises removal of poor-utility data first would reduce the overall noise of "useless" data in the final output.

Secondly, it is possible that since the original dataset is not extremely high-performance in and of itself, the models do lose utility but that the loss is unnoticeable because of the already lower statistics produced by the dataset. This uncertainty can be resolved through further work being conducted with different datasets - especially those with more than simply binary values, more privacy values and more epsilon values.

Finally, it is possible that the *"p"* and *"E"* scores chosen for taking measurements in this project aren't the most restrictive for this specific use-case - however the chosen values do reflect what real-life scenarios would require from the privacy-preserving mechanisms.







6 Related Work

This project's overall structure is based on Alishahi et. al (2022) [1], where a novel framework is proposed which allows for Feature Selection based on the the utility of the data - combined with LDP Feature Selection. This work is combined with the Data Generalisation structure as outlined in Martinelli and SheikhAlishahi (2019) [12] - whereby encryption is used as opposed to LDP as a privacy-preserving mechanism. As such, this work serves as a combination of the LDP Feature Selection and Data Generalisation frameworks presented in these papers. In this case, the Data Generalisation framework - relying on calculations of privacy and utility scores, is combined with LDP as a privacy-preserving mechanism.

Besides the work conducted in this framework - and the underlying mechanisms as proposed above, the field of privacy-preserving data collection enjoys active research and innovation. The notion of LDP was formalised in [9], and has been expanded on by major organisations who wish to deploy its uses in private data collection such as Apple [18] Microsoft [6], and Google [4].

Existing LDP implementations in academia explore different applications of LDP, for example Arachchige et al. (2020) [11] employ frequency estimation to deep learning of LDP. Further to this, Truex et al. (2020) [20] explore LDP's applications in Federated Learning environments.

Much like the framework proposed in this report, much of LDP focuses on distributed applications. Thông et. al (2016) [13] explores LDP's usages in distributed IoT use-cases, while Zhao et. al (2021) [24] explore the combined approaches of LDP in IoT and Federated Learning.

Another area of active research in privacy-preserving methods of Feature Selection, one school of thought revolves around investigating Feature Selection using encryption as a privacy-preserving mechanism, as Ono et. al (2022) [14] and Li et. al (2021) [10] explore for full Homomorphic encryption and Secure Multiparty communication respectively. Alishahi et al. (2022) [1] explore Feature Selection where Local Differential Privacy is used as a Feature Selection mechanism as opposed to encryption.

The proposed framework uses LDP as a secure Feature Selection mechanism, and incorporates Data Generalisation to explore these privacy-preserving mechanisms in the context of Classification models - which, to the best of the author's knowledge is a novel approach to LDP.

7 Conclusion

This report has proposed a novel framework for Feature Selection and privacy-preserving Data Generalisation, where distributed nodes have no trusted third party, individual privacy requirements and a shared interest in the data being published.

The proposed framework has been created and tested - while in a limited form due to the scope of the project. The framework has shown that the overall utility of the data when processed by the LDP-Feature Selection and Data Generalisation process is reduced, however this reduction in utility is very limited. This means that the generalised data produced by the framework still retains a large degree of the utility which the original data held.

As per **RQ1**, the proposed framework does ensure the accuracy of classifiers trained on the consolidated dataset as per the results as outlined in Section 5 to a great degree. The performance of the models does not drop by a large margin and in some measures improves in performance as a result of the process completed by the proposed framework.

As per **RQ2**, the framework contributes to privacy improvements for each participating entity by allowing them to set an individual privacy requirement for the framework to require the overall data to fit to, this means that the data's final overall privacy score will always fulfill the greatest privacy requirement given by the data-holding individuals. It does this by generalising the features of the distributed dataset until the framework is certain that the privacy requirements have been met, at which point the data is published and processed by the classification models. It ensures the privacy of the data during processing through LDP - which is a powerful privacy-preserving mechanism for each individual participating entity. The proposed framework allows granular control over the privacy achieved by both the Data Generalisation process (the "p" score), and the LDP process (ε).

7.1 Ethical Considerations

This project has created a novel framework designed to increase the privacy of inputted data. The framework has not been presented as anything other than a novel framework, and will not be represented as a catch-all privacy mechanism. The dataset used in this report is a published dataset from the US Government CDC. It contains information freely available to the public only.

8 Further Work

There are a number of ways the framework presented in this paper can be expanded on. Firstly, there are limitations in the dataset chosen for testing and training - in that only binary data which could be generalised to "*any*" was used in order to limit the overall scope of the project. A different dataset could be used which incorporates binary, discrete and continuous data into one data set - allowing for a much more granular data set and likely more high-quality classification results. This could allow for additional insight into the impact which LDP-Feature Selection and Data Generalisation has on the overall utility of a dataset - as well as its privacy under more complicated conditions.

Secondly, additional classification models could be explored, namely Neural Networks could be created and configured to optimally handle the generalised data. This, again, could produce higher quality final data which could provide a greater insight into the amount of utility lost in the framework as presented in this paper.

Thirdly, additional LDP models could be explored to determine the performance - both in terms of raw performance in a distributed system, and privacy - of differently selected models which could be designed to fit the specific use-case in mind for this framework. This could allow for higher quality results to be collected from the models as higher-quality data is inserted into the system.

Finally. more "p" values and "ε" values could be selected in order to get more details on the overall performance of the models - and test the limits of the performance of the framework as established in this paper.

Finally, additional metrics could be used to evaluate the utility of the framework's output data. A common metric used for this purpose is the NCP metric defined in [23]. By implementing further metrics, additional insight can be gained from the performance of framework - and the framework can be further optimised.

References

[1] Mina Alishahi, Vahideh Moghtadaiee, and Hojjat Navidan. "Add noise to remove noise: Local differential privacy for feature selection". In: *Computers and Security* 123 (2022), p. 102934. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2022.102934. URL: https://www.sciencedirect.com/science/article/pii/S0167404822003261.

[2] CDC. *UCI Machine Learning Repository, CDC Diabetes Health Indicators.* 2015. DOI: https://doi.org/10.24432/C53919. URL: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators.

[3] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers and Electrical Engineering* 40.1 (2014). 40th-year commemorative issue, pp. 16–28. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024. URL: https://www.sciencedirect.com/science/article/pii/S0045790613003066.

[4] Girish Chandrashekar and Ferat Sahin. "A survey on feature selection methods". In: *Computers and Electrical Engineering* 40.1 (2014). 40th-year commemorative issue, pp. 16–28. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024. URL: https://www.sciencedirect.com/science/article/pii/S0045790613003066.

[5] Mariana Cunha, Ricardo Mendes, and João P. Vilela. "A survey of privacy-preserving mechanisms for heterogeneous data types". In: *Computer Science Review* 41 (2021), p. 100403. ISSN: 1574-0137. DOI: https://doi.org/10.1016/j.cosrev.2021.100403. URL: https://www.sciencedirect.com/science/article/pii/S1574013721000435.

[6] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. "Collecting Telemetry Data Privately". In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/253614bbac999b38b5b60cae531c49 Paper.pdf.

[7] Haroon Elahi, Guojun Wang, and Dongqing Xie. "Assessing privacy behaviors of smartphone users in the context of data over-collection problem: An exploratory study". In: *2017 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).* 2017, pp. 1–8. DOI: 10.1109/UIC-ATC.2017.8397613.

[8] Robert M. Grant. "Toward a knowledge-based theory of the firm". In: *Strategic Management Journal* 17.S2 (1996), pp. 109–122. DOI: https://doi.org/10.1002/smj.4250171110. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.4250171110. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.4250171110.

[9] Shiva Prasad Kasiviswanathan et al. "What Can We Learn Privately?" In: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826. DOI: 10.1137/090756090. eprint: https://doi.org/10.1137/090756090. URL: https://doi.org/10.1137/090756090.

[10] Xiling Li, Rafael Dowsley, and Martine De Cock. "Privacy-Preserving Feature Selection with Secure Multiparty Computation". In: *Proceedings of the 38th International Conference on Machine Learning.* Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 6326–6336. URL: https://proceedings.mlr.press/v139/li21e.html.

[11] Pathum Chamikara Mahawaga Arachchige et al. "Local Differential Privacy for Deep Learning". In: *IEEE Internet of Things Journal* 7.7 (2020), pp. 5827–5842. DOI: 10.1109/JIOT.2019.2952146.

[12] Fabio Martinelli and Mina SheikhAlishahi. "Distributed Data Anonymization". In: *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. 2019, pp. 580–586. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00113.

[13] Thông T. Nguyên et al. "Collecting and Analyzing Data from Smart Device Users with Local Differential Privacy". In: *CoRR* abs/1606.05053 (2016). arXiv: 1606.05053. URL: http://arxiv.org/abs/1606.05053.

[14] Shinji Ono et al. "Privacy-Preserving Feature Selection with Fully Homomorphic Encryption". In: *Algorithms* 15.7 (2022). ISSN: 1999-4893. DOI: 10.3390/a15070229. URL: https://www.mdpi.com/1999-4893/15/7/229.

[15] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[16] *PyCharm: the Python IDE for Professional Developers by JetBrains*. June 2021. URL: https://www.jetbrains.com/pycharm/.

[17] Yuji Roh, Geon Heo, and Steven Euijong Whang. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective". In: *IEEE Transactions on Knowledge and Data Engineering* 33.4 (2021), pp. 1328–1347. DOI: 10.1109/TKDE.2019.2946162.

[18] Differential Privacy Team. *Learning with Privacy at Scale*. 2017. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[19] Patrick Tobler et al. "AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size". In: *European Radiology* 31.9 (Sept. 2021), pp. 6816–6824. ISSN: 1432-1084. DOI: 10.1007/s00330-021-07811-2. URL: https://doi.org/10.1007/s00330-021-07811-2.

[20] Stacey Truex et al. "LDP-Fed: federated learning with local differential privacy". In: *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*. EdgeSys '20. Heraklion, Greece: Association for Computing Machinery, 2020, pp. 61–66. ISBN: 9781450371322. DOI: 10.1145/3378679.3394533. URL: https://doi.org/10.1145/3378679.3394533.

[21] Ke Wang, P.S. Yu, and S. Chakraborty. "Bottom-up generalization: a data mining solution to privacy protection". In: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. 2004, pp. 249–256. DOI: 10.1109/ICDM.2004.10110.

[22] Stanley L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69. ISSN: 01621459. URL: http://www.jstor.org/stable/2283137 (visited on 01/09/2024).

[23] Jian Xu et al. "Utility-based anonymization using local recoding". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 785–790. ISBN: 1595933395. DOI: 10.1145/1150402.1150504. URL: https://doi.org/10.1145/1150402.1150504.

[24] Yang Zhao et al. "Local Differential Privacy-Based Federated Learning for Internet of Things". In: *IEEE Internet of Things Journal* 8.11 (2021), pp. 8836–8853. DOI: 10.1109/JIOT.2020.3037194.