

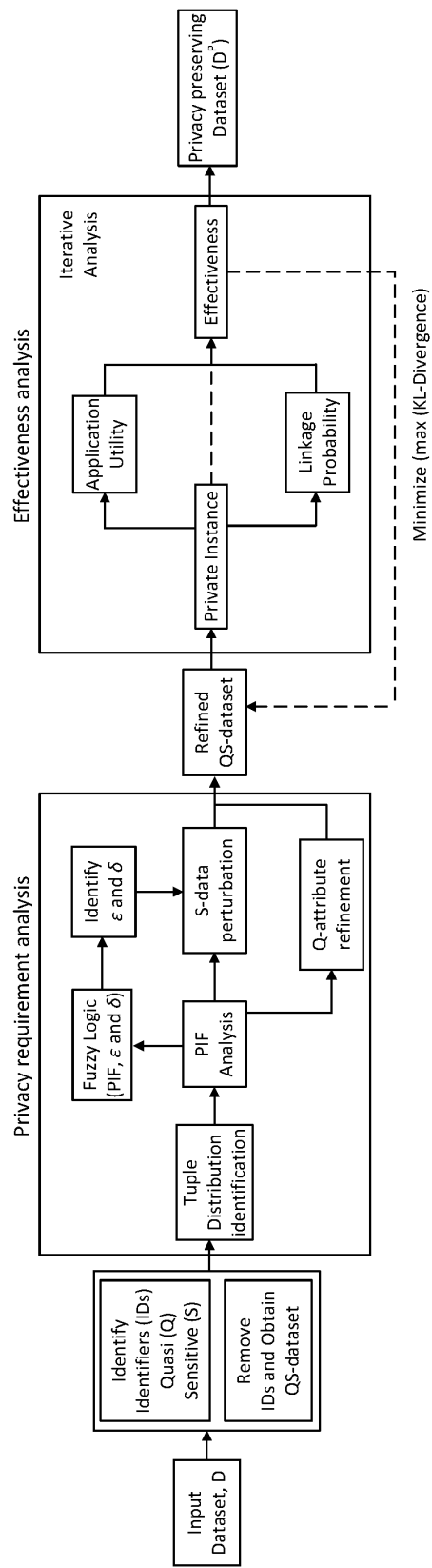


Fig. 1

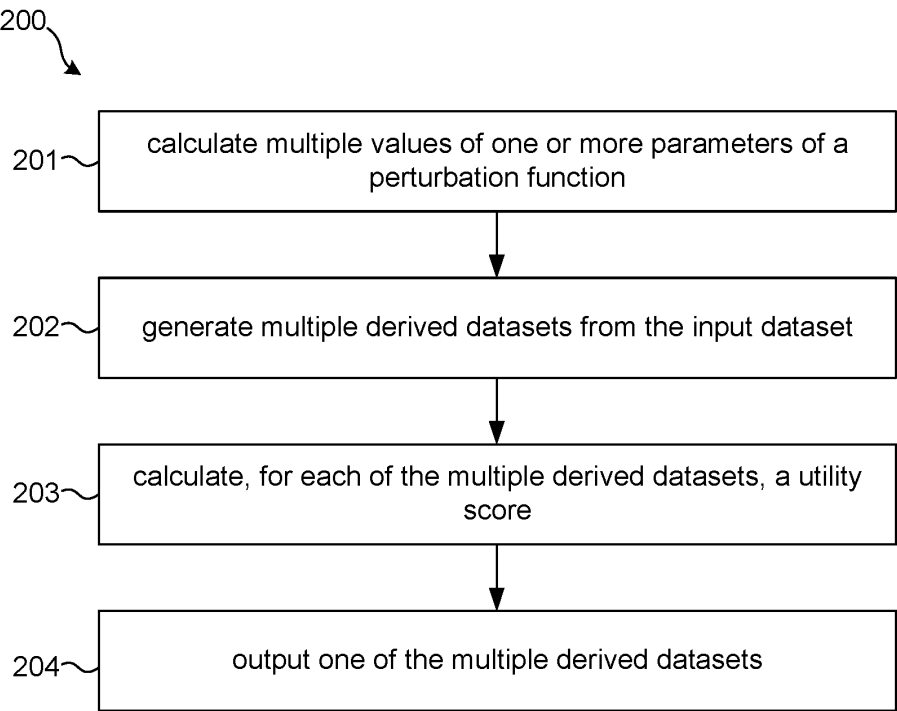


Fig. 2a

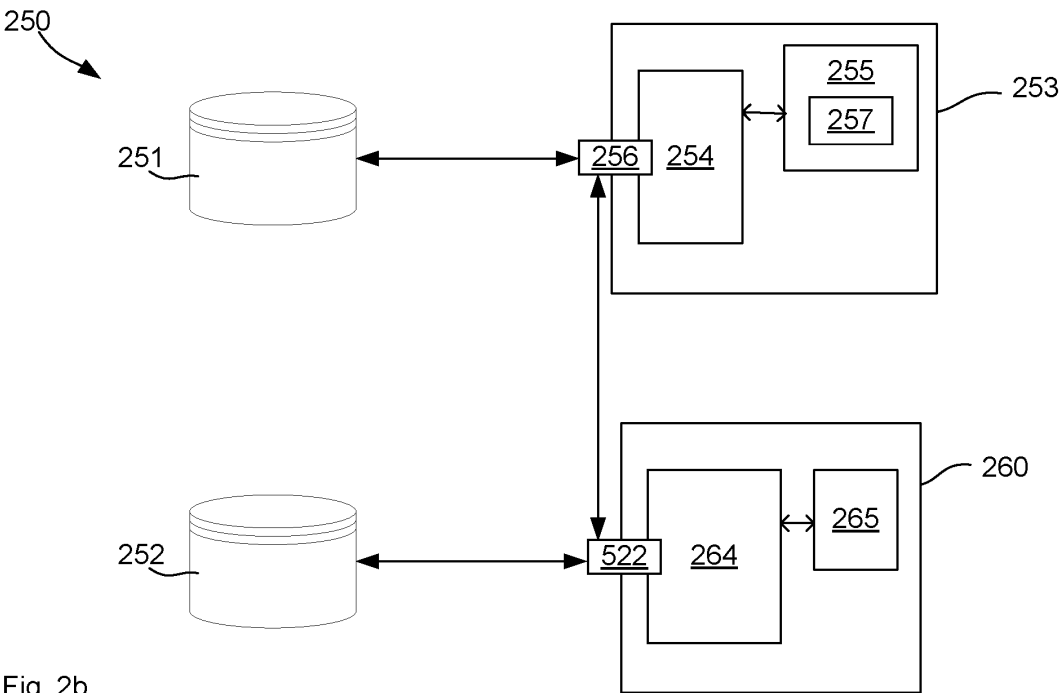


Fig. 2b

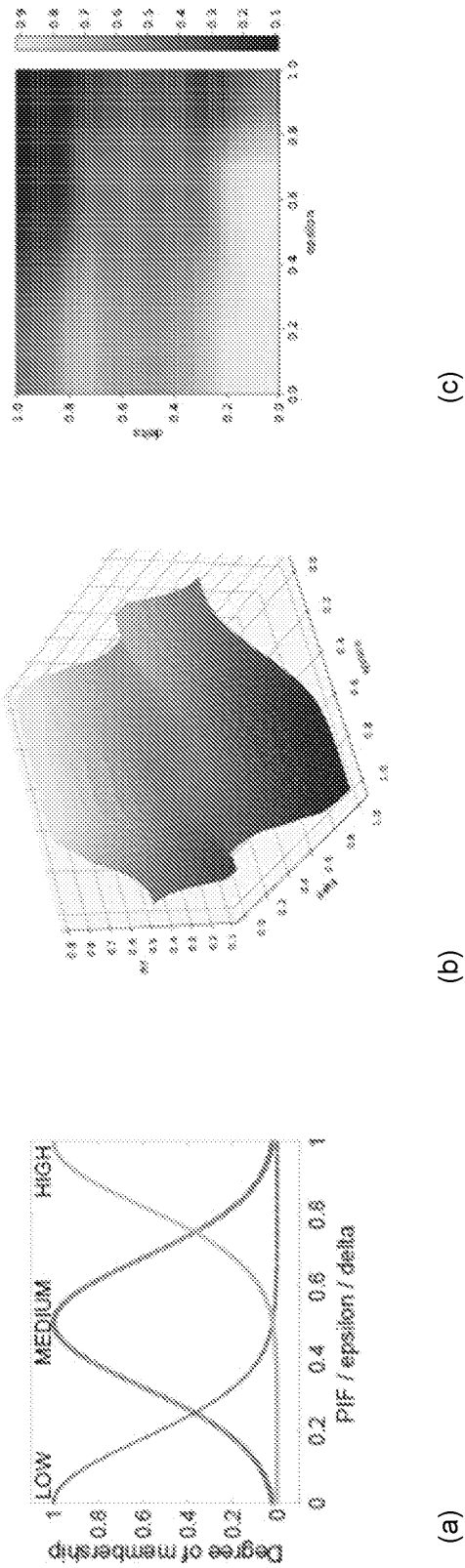


Fig. 3

---

|                                                                                     |                                                                                                                                                                                                                                                                                    |
|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Algorithm</b> : OptimShare algorithm for generating a privacy-preserving dataset |                                                                                                                                                                                                                                                                                    |
| <hr/>                                                                               |                                                                                                                                                                                                                                                                                    |
| <b>Input:</b>                                                                       |                                                                                                                                                                                                                                                                                    |
| $D$                                                                                 | $\leftarrow$ input dataset                                                                                                                                                                                                                                                         |
| $T_{id}$                                                                            | $\leftarrow$ probability threshold of $ID$ attribute selection                                                                                                                                                                                                                     |
| $G_q$                                                                               | $\leftarrow$ global $Q$ attribute set                                                                                                                                                                                                                                              |
| $cn_{range}$                                                                        | $\leftarrow$ cluster number of tuple distribution analysis                                                                                                                                                                                                                         |
| $U_\varepsilon$                                                                     | the upperbound for $\varepsilon$                                                                                                                                                                                                                                                   |
| $P_l$                                                                               | $\leftarrow$ percentage leak for $\delta$                                                                                                                                                                                                                                          |
| $TN_{\varepsilon,\delta}$                                                           | $\leftarrow$ total $(\varepsilon, \delta)$ combinations to be generated                                                                                                                                                                                                            |
| $TS$                                                                                | $\leftarrow$ total number of searches                                                                                                                                                                                                                                              |
| $t$                                                                                 | $\leftarrow$ perturbed instances per $DPA$                                                                                                                                                                                                                                         |
| $A$                                                                                 | $\leftarrow$ type of application                                                                                                                                                                                                                                                   |
| $C_\varepsilon$                                                                     | $\leftarrow$ effectiveness coefficient                                                                                                                                                                                                                                             |
| $E^T$                                                                               | $\leftarrow$ effectiveness threshold                                                                                                                                                                                                                                               |
| <b>Output:</b>                                                                      | $D^p \leftarrow$ perturbed dataset of $D$                                                                                                                                                                                                                                          |
| 1                                                                                   | identify identifiers ( $ID$ ) and quasi-identifiers ( $Q$ ) from $S'$ of $D$ , where $S' = \{x : x \in U \setminus S\}$ and $U$ represents the set of all attributes of $D$ ;                                                                                                      |
| 2                                                                                   | remove $ID$ from the dataset using $T_{id}$ to produce $QS$ -dataset;                                                                                                                                                                                                              |
| 3                                                                                   | identify tuple distribution of the $QS$ -dataset ;                                                                                                                                                                                                                                 |
| 4                                                                                   | identify initial $QS$ dataset and determine $PIF$ of $Q$ -dataset ;                                                                                                                                                                                                                |
| 5                                                                                   | determine the anonymity level necessary for $Q$ ;                                                                                                                                                                                                                                  |
| 6                                                                                   | determine $PIF$ of $Q$ -dataset conditioned to $S$ ;                                                                                                                                                                                                                               |
| 7                                                                                   | refine the $Q$ attributes and $S$ attributes;                                                                                                                                                                                                                                      |
| 8                                                                                   | recalculate $PIF$ of $Q$ ;                                                                                                                                                                                                                                                         |
| 9                                                                                   | generate potential $\varepsilon$ and $\delta$ combinations $\{(\varepsilon_1, \delta_1), (\varepsilon_2, \delta_2), \dots, (\varepsilon_n, \delta_n)\}$ (where $\varepsilon_i > \varepsilon_{i+1}$ and $\delta_i > \delta_{i+1}$ ) using the fuzzy model based on the $PIF$ value; |
| 10                                                                                  | <b>for each</b> $(\varepsilon_i, \delta_i) \in \{(\varepsilon_1, \delta_1), (\varepsilon_2, \delta_2), \dots, (\varepsilon_{TS}, \delta_{TS})\}$ <b>do</b>                                                                                                                         |
| 11                                                                                  | apply $k$ number of suitable differentially private algorithms from the pool of algorithms $(DPA_1, DPA_2, \dots, DPA_n)$ , and generate $t \times k$ perturbed instances;                                                                                                         |
| 12                                                                                  | merge $Q$ to all $t \times k$ perturbed instances $(DP_1, DP_2, \dots, DP_{t \times k})$ ;                                                                                                                                                                                         |
| 13                                                                                  | generate $t \times k$ normalized utility values;                                                                                                                                                                                                                                   |
| 14                                                                                  | generate $t \times k$ privacy leak normalized values;                                                                                                                                                                                                                              |
| 15                                                                                  | find effectiveness value $e_i$ of each $t \times k$ perturbed instances;                                                                                                                                                                                                           |
| 16                                                                                  | choose all $DP_i$ that satisfy $e_i \geq E^T$ ;                                                                                                                                                                                                                                    |
| 17                                                                                  | <b>return</b> $D^p$ to with largest $e_i$ ;                                                                                                                                                                                                                                        |

---

Fig. 4a

---

**Algorithm** : OptimShare algorithm for generating a privacy-preserving dataset

---

$D \leftarrow$  input dataset  
 $T_\epsilon \leftarrow$  threshold  $\epsilon$   
 $TN_{\epsilon,\delta} \leftarrow$  total  $(\epsilon, \delta)$  combinations  
 $TS \leftarrow$  total number of searches  
**Input:**  $t \leftarrow$  perturbed instances per combination  
 $A \leftarrow$  application  
 $C \leftarrow$  effectiveness coefficient  
 $E^T \leftarrow$  effectiveness threshold  
**Output:**  $D^p \leftarrow$  perturbed dataset of  $D$

- 1 Identify identifiers ( $ID$ ) and quasi-identifiers ( $Q$ );
- 2 Remove  $ID$  from the dataset to produce  $QS$ -dataset;
- 3 Identify tuple distribution of the  $QS$ -dataset;
- 4 Determine  $PIF$  of  $Q$ -dataset;
- 5 Determine  $PIF$  of  $Q$ -dataset conditioned to  $S$ ;
- 6 Refine the  $Q$  attributes and  $S$  attributes;
- 7 Generate  $TN_{\epsilon,\delta}$  combinations as  $\{(\epsilon_1, \delta_1), \dots, (\epsilon_{TN}, \delta_{TN})\}$  (where  $\epsilon_i < \epsilon_{i+1}$  and  $\delta_i < \delta_{i+1}$ );
- 8 **for each**  $(\epsilon_i, \delta_i)$  **at**  $TN_{\epsilon,\delta}$  **intervals in**  $TS$  **do**
  - 9 Apply DP generators ( $DPA_1, \dots, DPA_n$ ) to generate  $t$  perturbed instances;
  - 10 Merge  $Q$  to all  $t$  perturbed instances ( $DP_1, \dots, DP_t$ );
  - 11 Generate normalized utility values of all  $DP_i$ ;
  - 12 Generate  $t$  residual leak normalized values;
  - 13 Find effectiveness loss value  $el_i$  of each  $t$  perturbed instances for  $A$  using  $C$  and  $T_\epsilon$ ;
  - 14 Choose all  $DP_i$  that satisfy  $e_i \geq E^T$  (where  $e_i = 1 - el_i$ );
- 15 Return  $D^p$  with the highest  $e_i$ ;

---

Fig. 4b

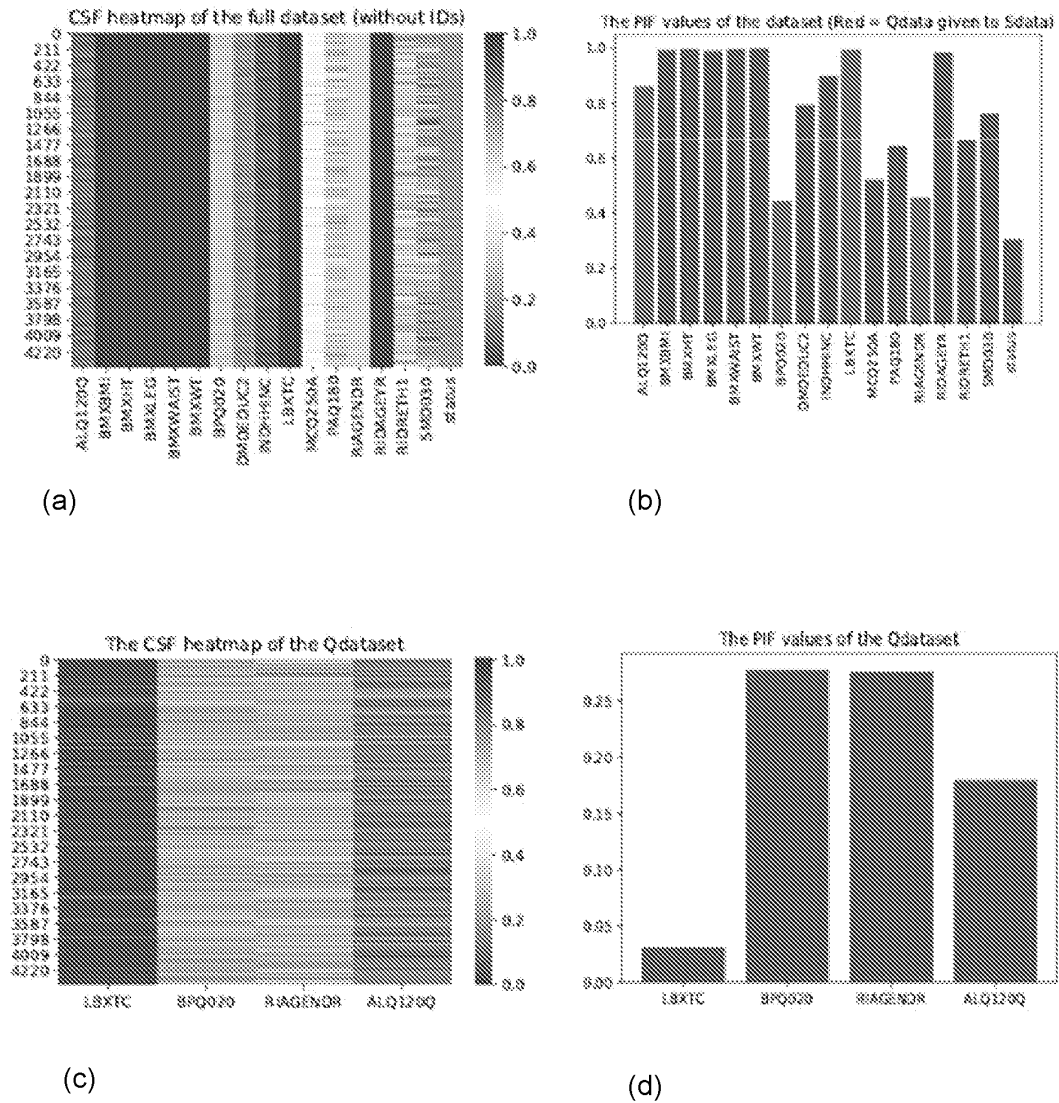
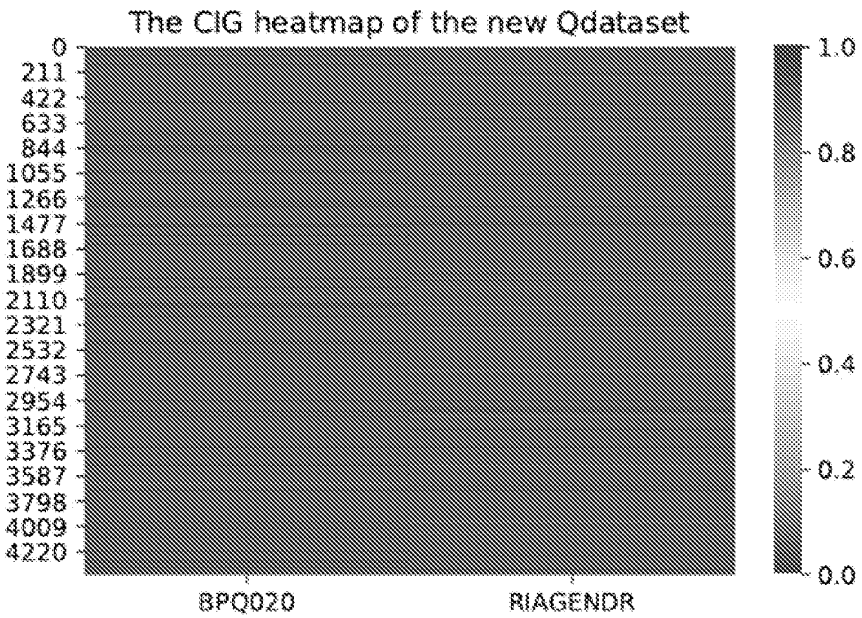
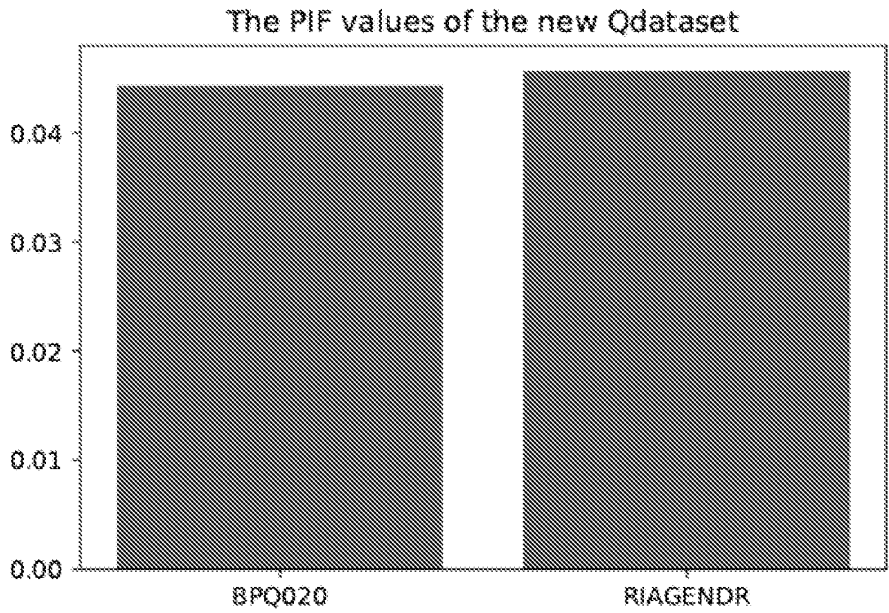


Fig. 5



(a)



(b)

Fig. 6

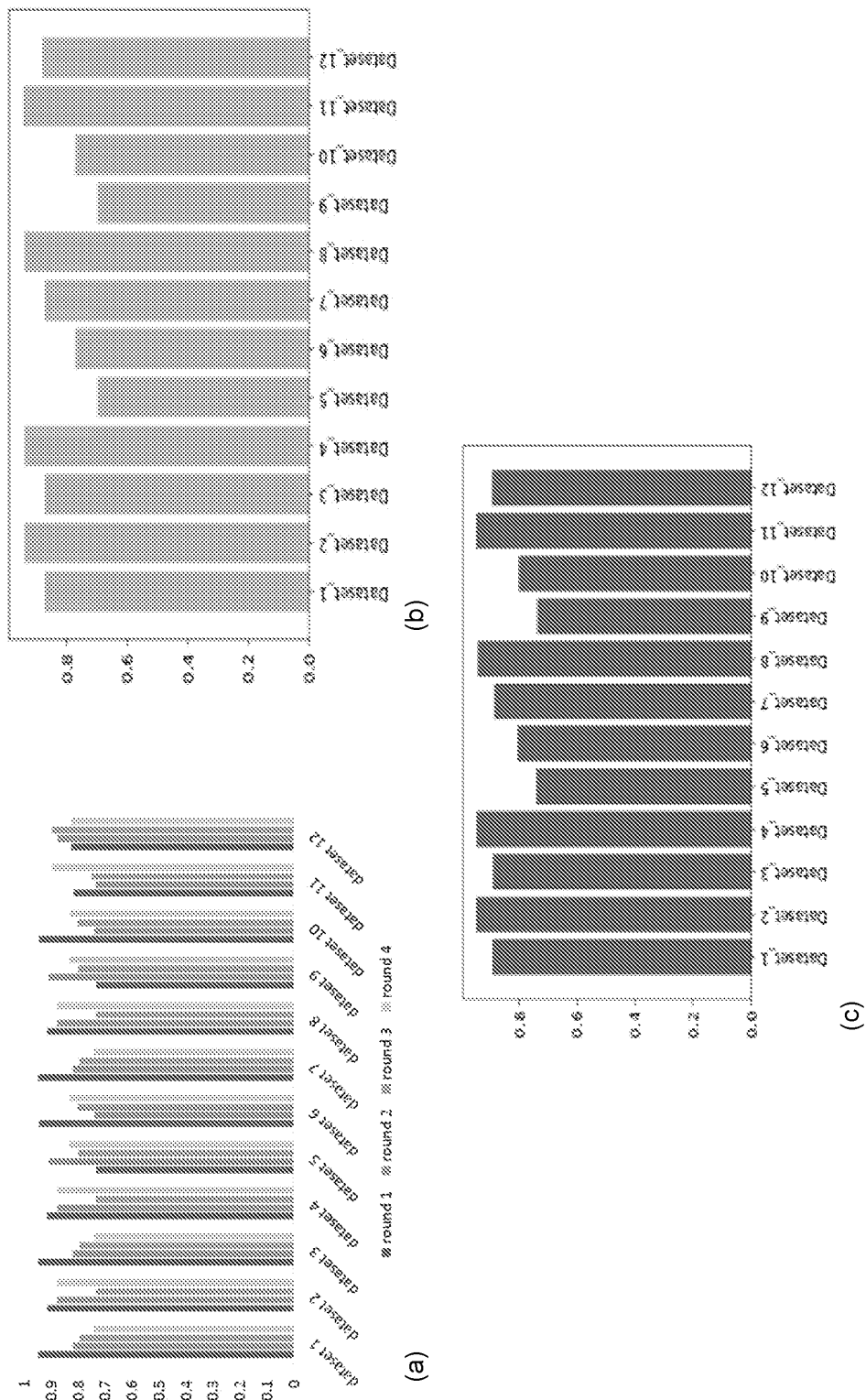
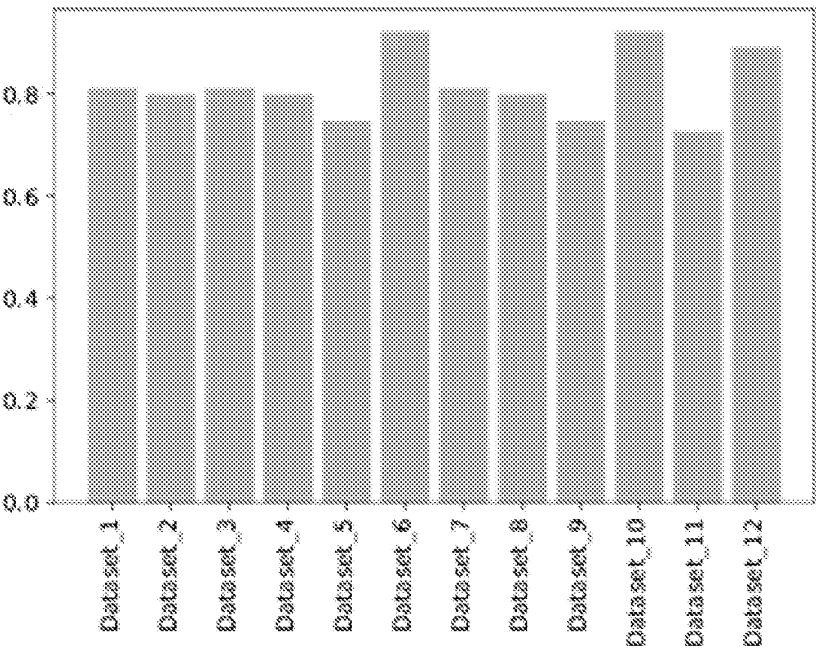
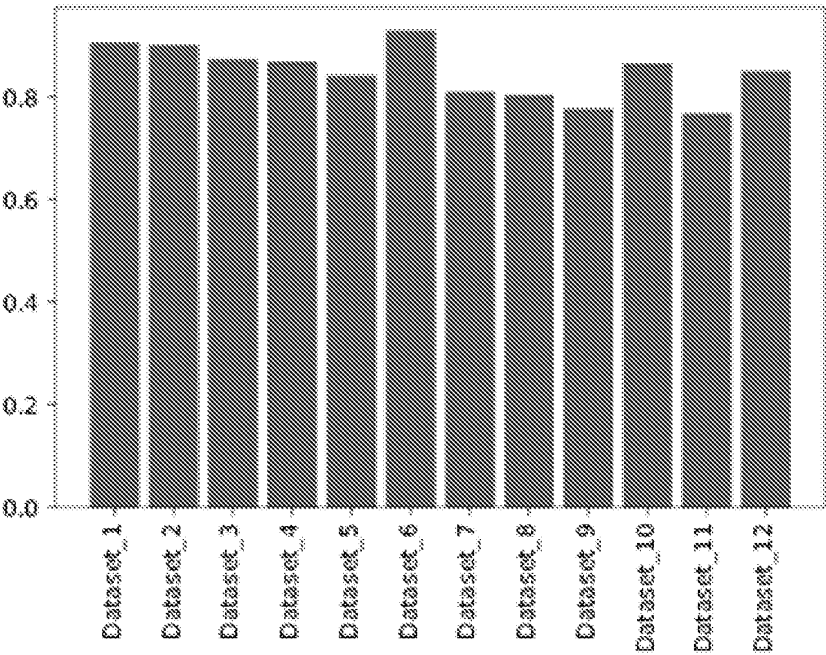


Fig. 7



(a)



(b)

Fig. 8

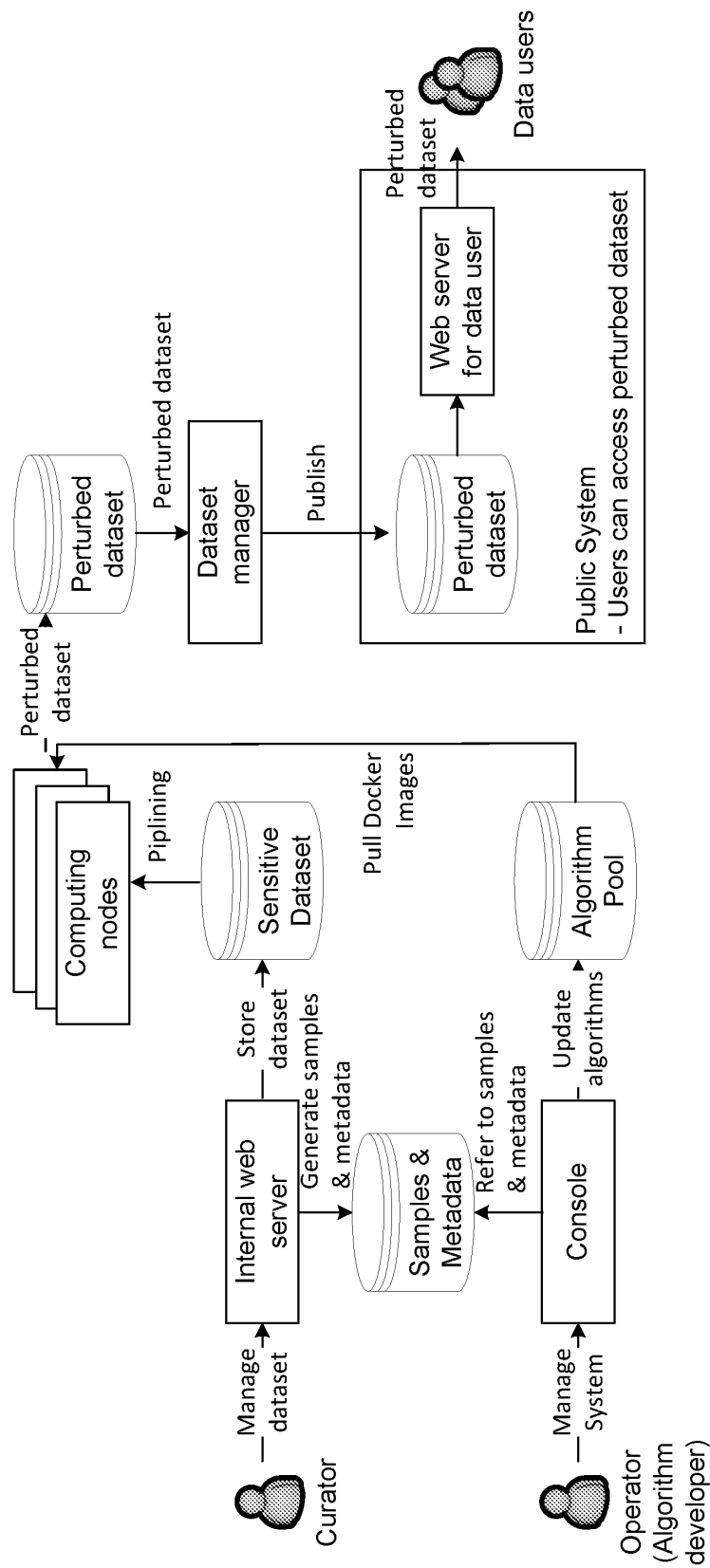


Fig. 9

| Dataset                                 | Abbr. | Records | Attributes | Classes | Global Q                                                             | Attributes |
|-----------------------------------------|-------|---------|------------|---------|----------------------------------------------------------------------|------------|
| NHANES diabetes Kaggle <sup>1</sup>     | NHDS  | 4,412   | 17         | 2       | 'BPQ020', 'RIAGENDR', 'ALQ120Q', 'LBXTC'                             |            |
| Wine Quality <sup>2</sup>               | WQDS  | 4,898   | 12         | 7       | 'free sulfur dioxide', 'total sulfur dioxide'                        |            |
| Page Blocks Classification <sup>3</sup> | PBDS  | 5,473   | 11         | 5       | 'at1', 'at2', 'at10'                                                 |            |
| Letter Recognition <sup>4</sup>         | LRDS  | 20,000  | 17         | 26      | 'letter', 'x-box', 'y-box', 'width', 'high', 'xy2br'                 |            |
| Statlog (Shuttle) <sup>5</sup>          | SSDS  | 58,000  | 9          | 7       | 'b', 'd', 't'                                                        |            |
| Credit Score Kaggle <sup>6</sup>        | CSDS  | 150,000 | 11         | 2       | 'ID', '#ofOCLL', '#ofT90DL', '#RELL', '#ofT60DPDNW', '#ofDependents' |            |

Fig. 10