

# モデルの情報効用を最大化する タスク誘導型差分プライバシ保護合成データ生成

税所 修<sup>1,a)</sup> 三浦 堯之<sup>1</sup> 岩花 一輝<sup>1</sup> 紀伊 真昇<sup>1</sup>

**概要：**本論文は人間を系に含む AI ライフサイクル全体を通じたプライバシ保護に着目し、人間によるラベルアノテーション工程を差分プライバシ下で安全かつ効率的に実現する手法を提案する。従来の AI セキュリティ研究では学習や推論に限定して着目し、アノテータへのデータ提示に伴うプライバシリスクを扱ってこなかったが、本論文では、有益なサンプルを優先的に差分プライバシ合成データとして生成、提示する新しい枠組みを確立することで、その未解決課題に取り組む。具体的には、差分プライバシ保護合成データ生成手法 AIM の適応的な逐次更新過程に、能動学習の獲得関数 BALD を統合し、プライバシ予算とアノテーション予算を適応的に消費する設計により実現する。Adult データセットを用いた実験的評価により、提案手法はランダムサンプリングや従来手法の単純な組み合わせを大きく上回り、少データ環境においても実データ利用時に近い精度を達成することを示した。これにより、本研究は効率的なデータ利用と厳密なプライバシ保証を両立する新たなプライバシ・バイ・デザイン型 AI 開発の包括的なアプローチを提示する。

**キーワード：**差分プライバシ保護合成データ、能動学習、プライバシ・バイ・デザイン、人間を系に含む AI ライフサイクル

## Task-Guided Differentially Private Synthetic Data Generation via Active Model Utility Maximization

OSAMU SAISHO<sup>1,a)</sup> TAKAYUKI MIURA<sup>1</sup> KAZUKI IWANAHANA<sup>1</sup> MASANOBU KII<sup>1</sup>

**Abstract:** This paper focuses on privacy protection across the entire AI lifecycle involving human participants, and proposes a method to realize human-in-the-loop label annotation safely and efficiently under differential privacy. While prior research on AI security has mainly focused on training and inference, the privacy risks of presenting data to annotators have been largely overlooked. In this work, we address this unresolved issue by establishing a novel framework that generates and presents differentially private synthetic data, prioritizing informative samples. Specifically, we integrate AIM for differentially private synthetic data generation with BALD acquisition function, and design adaptive consumption of both privacy and annotation budgets. Through empirical evaluation on the Adult dataset, we demonstrate that the proposed method significantly outperforms random sampling and naive combinations of existing methods, achieving accuracy close to models trained on actual data even in low-data regimes. These results indicate that our work provides a comprehensive privacy-by-design approach to AI development that simultaneously ensures efficient data utilization and strict privacy guarantees.

**Keywords:** differentially private synthetic data, active learning, privacy-by-design, AI lifecycle involving human participants

## 1. 諸論

医療ヘルスケアなど、ユーザ自身の機密性の高い実世界

<sup>1</sup> NTT 社会情報研究所  
NTT Social Informatics Laboratories  
a) osamu.saisho@ntt.com

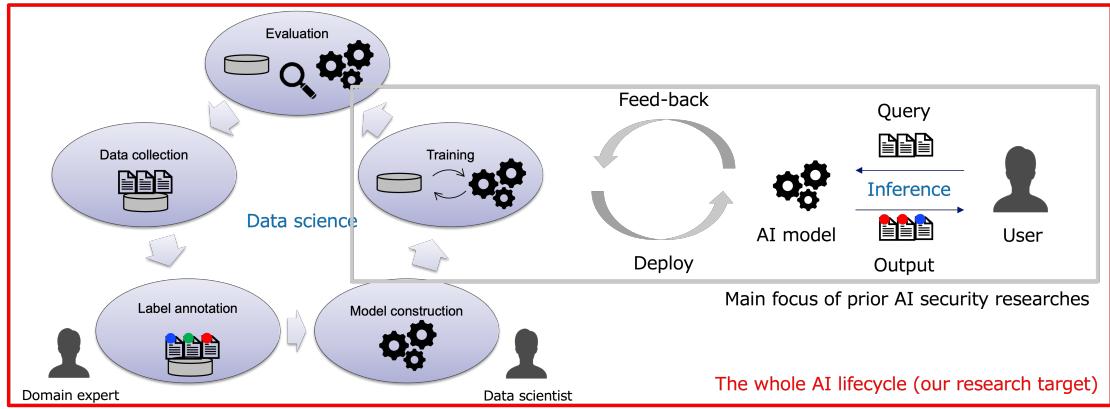


図 1 本論文が対象とする領域。AI モデルの学習や推論に限らず、人間を系に含む AI ライフサイクル全体を通じたプライバシ保護を実現する。

データを用いた AI 活用が身近となった今日では、デジタル領域においてもプライバシ・バイ・デザインの原則が不可欠である [5], [34]。プライバシ・バイ・デザインの原則は、プライバシへの配慮をシステム開発そのものに組み込むことで、保護対策を後から追加するのではなく最初から統合することを保証するものである。AI をユーザが使用するシステムとして捉えると、AI 開発のフロー全体を対象とすることが求められる。すなわち、AI 適用先のタスクに関する専門家がデータ 1 つ 1 つに対して正解情報を付与するアノテーションを行う学習データセット構築やデータサイエンティストによる識別モデル構築といった、第三者の人間を系に含む AI ライフサイクル全体でプライバシ保護などを保証することが必要となる。ことに医療ヘルスケアやユビキタスコンピューティングといった領域における AI 研究や AI 開発では、AI ライフサイクルの第一ステップとして実世界データの収集と人間によるアノテーションが今日でも一般的に行われており、AI システム開発における最重要要素の 1 つを占めている。

プライバシ保護技術の著しい進歩にもかかわらず、AI セキュリティにおける既存アプローチは AI 開発のフロー全体を網羅できていない。AI セキュリティ研究では AI モデルや学習データがすでに存在するものとし、それらに含まれるプライバシや情報に関して攻撃や防御、保護を語るのが一般的なためである。例えば連合学習や秘密計算は、モデルの学習と推論におけるプライバシ保証を提供するが [22]、これらは人間を系に含む AI ライフサイクルの小さなサブセットでしかない [20]。問題設定をサブセットの範囲で閉じたうえでいくら効用とプライバシ保護のトレードオフを解消したとしても、他の工程でプライバシ漏洩をするのであれば、データセットや AI の構築の構築から AI システムのデプロイまでのフロー全体を安心安全に流すことはできない。

人がデータそのものに触れるということから、AI の学習に用いるデータそのもののプライバシ保護を実現する

手法として、差分プライバシ保護合成データ生成が注目されている。合成データは、実世界データの全ての属性の分布、統計量、相関関係に基づいて生成され、実世界データと似て非なる架空のデータである。差分プライバシ (DP) 保証 [9] をデータ生成プロセスに組み込むことで、合成データそのものや合成データのみで学習した AI モデルが公開されても、元の実世界データのプライバシが理論的に保証される [15], [23], [32]。プライバシ保護合成データは、k-匿名化 [31] や Pk-匿名化 [17] のような伝統的なプライバシ保護手法と比較して、高次元のデータセットでもより良いパフォーマンスを発揮し、プライバシ保護とデータ有用性の両立が期待されている [7]。

本論文の目的は、人間を系に含む AI ライフサイクル全体におけるプライバシ保護実現に向けて、人がプライバシ保護合成データにラベルアノテーションを行う実用的な工程を構築、検証することである。人間を系に含む場合にプライバシ保護と同様に課題となるのが、人間の有限な稼働である。通常の実データを用いた AI 開発においては、この課題に対して人間のラベルアノテーション稼働の最小化と AI モデルの有用性の最大化を同時に能動学習 [28] により対応するのが一般的である。そこで本論文ではプライバシ保護合成データ生成と能動学習を組み合わせることで、AI ライフサイクル全体におけるプライバシ保護、学習により得られる AI モデルの有用性の最大化、人間のアノテーション作業コストの最小化の 3 つを同時に満たす手法を確立する。具体的には、ラベルなし実データプールの存在を前提として、適応的かつ逐次的に高品質なプライバシ保護合成データを生成できる AIM [23] における生成パラメータにベイジアン能動学習手法の BALD [14] の獲得関数を組み込んで、逐次的にラベルなしプライバシ保護合成データを生成し、人がその合成データにアノテーションを行う枠組みを確立する。これにより、実データにおける差分プライバシを保証する条件下で、実データ分布とその獲得関数のスコアを参照し、AI モデルの有用性を最大化

する合成データのみを適応的かつ逐次的に生成、人間に提示する。なお差分プライバシ保証については、獲得関数の組み込みを AIM を保証理論を崩さずに実現することで担保する。また本枠組みでは、ラベルアノテーションにおいて保証を行い、その後実データを参照しないため、差分プライバシの後処理定理により、AI ライフサイクル全体の差分プライバシ保証を実現できる。

本論文の主な貢献は以下の通りである。第一に、AIM により生成したプライバシ保護合成データに対してアノテーションを行うことで厳密なプライバシ保護と高い AI モデル性能のトレードオフを克服できる一方、能動学習をそのまま適用しても効果が出ないことを実証する。第二に、プライバシ予算とアノテーション予算を同時に消費し、BALD スコアを考慮した AIM により差分プライバシ保護合成データを適応的かつ逐次的に生成、提示することで上記課題を解決できることを実証する。これらの貢献は、人間を系に含む AI ライフサイクル全体にプライバシ・バイ・デザインの原則を持ち込み、機密性の高いアプリケーションのための安全な AI エコシステムの開発において一貫したプライバシ保護、AI の高性能化、人間稼働の省力化の 3 課題を同時に満たす包括的なソリューションにつながる。

## 2. 準備

### 2.1 差分プライバシ

差分プライバシ [9] はプライバシ保護の程度を定量化する代表的な概念である。これは 1 つのデータ点の追加や削除によってアルゴリズムの出力が大きく変化しないことを保証することで、プライバシ保護性能を示す。具体的には、ランダム化アルゴリズム  $\mathcal{M}$  が  $(\epsilon, \delta)$ -差分プライバシを満たすのは、レコードが 1 つ異なる全ての隣接するデータセット  $D$  と  $D'$ 、および出力空間の全部分集合  $S$  に対して、次が成り立つときである：

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta.$$

なお、 $\epsilon$  や  $\delta$  が小さいほどプライバシがより厳しく保護されていることを示す。本論文では、後述する AIM における理論保証を活用して、ラベルアノテーション過程に含まれる合成データ生成プロセスにおいて差分プライバシを保証する。

### 2.2 プライバシ保護合成データ生成

人間が直接触れるデータそのもののプライバシ保護を目的とした手法として、合成データが活用されている [29]。合成データは、実世界のデータの分布、統計量、相関関係に基づいて生成されるため、実世界データと似て非なる架空データであり、主にデータ拡張や実世界データの代替として使用されるが、有用性の高さに反して実世界データのプライバシは保証されていないため、合成データを通じて

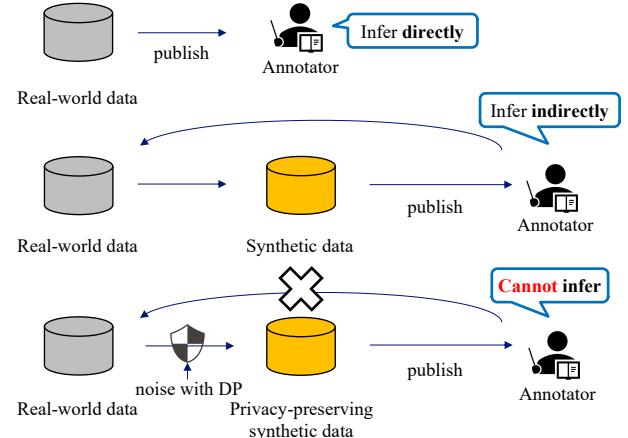


図 2 データ種によるプライバシ漏洩リスクの比較。差分プライバシ保護合成データのみが保証できる。

機密情報が漏洩する危険性がある [4], [6], [30]。差分プライバシ保護合成データは、その生成プロセス中に差分プライバシを適用することで実世界データのプライバシを保証する。図 2 のように、実世界データのプライバシを保証できるのは差分プライバシ保護合成データのみである。様々な生成アプローチが提案されており、統計に基づく手法 [1], [21]、グラフィカルモデルに基づく手法 [37]、深層生成モデルに基づく手法 [35] などが代表的である。

#### 2.2.1 AIM

AIM は差分プライバシ下で実データのマージナル統計を逐次的に推定し、それに整合するように合成データを適応的に更新していく方式による差分プライバシ保護合成データ生成手法である [23]。ここでマージナルとは、データセットの属性集合の部分集合に着目して周辺化した確率分布を指す。例えば年齢単独の分布（1 次元マージナル）や、性別と収入の同時分布（2 次元マージナル）が該当する。AIM はこれらのマージナル候補のうち精度改善に寄与するものを選んで推定し、合成データ分布をその結果に基づいて適応的に更新することで、限られたプライバシ予算で高品質な合成データを生成できる。逐次的な生成過程の 1 ラウンドはマージナル選択、差分プライバシ測定、合成データ更新の 3 ステップで構成される。

マージナル選択では、候補となるマージナル集合  $\mathcal{M}$  の各マージナル  $m \in \mathcal{M}$  に対して、実世界データのデータ分布  $P_{\text{true}}$  と暫定の合成データ分布  $P_{\text{synth}}^{(t)}$  の誤差が最大となるものを下記の通り選択する：

$$m_t = \operatorname{argmax}_{m \in \mathcal{M}} |m(P_{\text{true}}) - m(P_{\text{synth}}^{(t)})|. \quad (1)$$

差分プライバシ測定では、選択されたマージナル  $m_t$  の集計表を実世界データセット  $X$  に対して計算し、差分プライバシ機構  $M_{\epsilon_t}$  を適用することで、ノイズ付きマージナル推定値  $\tilde{m}_t(X)$  を得る：

$$\tilde{m}_t(X) \sim M_{\epsilon_t}(m_t(X)). \quad (2)$$

---

**Algorithm 1:** 一般的な能動学習アルゴリズム

---

**Input:** Unlabeled dataset  $\mathcal{D}_U$ , initial labeled dataset  $\mathcal{D}_L$ , acquisition function  $A(\cdot)$ , query budget  $B$

**Output:** Trained model  $f$

Train model  $f$  on labeled dataset  $\mathcal{D}_L$ ;

**for**  $t = 1$  **to**  $B$  **do**

Compute acquisition scores  $A(x)$  for all  $x \in \mathcal{D}_U$ ;

Select  $x^* = \operatorname{argmax}_{x \in \mathcal{D}_U} A(x)$ ;

Query label  $y^*$  for  $x^*$  from annotators;

$\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \{(x^*, y^*)\}$ ;

$\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \{x^*\}$ ;

Retrain model  $f$  on updated  $\mathcal{D}_L$ ;

**return**  $f$

---

このときプライバシ予算は逐次的に消費され、全ラウンドを通じて

$$\sum_{t=1}^T \epsilon_t \leq \epsilon \quad (3)$$

を満たすように管理され、効率的に予算を使い切るためにアニーリング戦略（初期ラウンドで大きなノイズ、後半で小さなノイズ）も用いられる。合成データ更新では、これまでに得られたノイズ付きマージナル推定値  $\{\tilde{m}_1(X), \dots, \tilde{m}_t(X)\}$  に整合するように、合成データ分布を更新する。これは次の最適化問題として定式化される：

$$P_{\text{syn}}^{(t+1)} = \arg \min_P \sum_{i=1}^t \left( P(m_i) - \tilde{m}_i(X) \right)^2. \quad (4)$$

ここで  $P(m)$  は分布  $P$  におけるマージナル  $m$  の期待値である。実装においては、Private PGM [24] が用いられ、ノイズ付きマージナルに整合する確率グラフィカルモデルを推定し、そこから暫定合成データを生成する。

### 2.3 能動学習

能動学習 [28] は人間のラベルアノテーション稼働を最小化しつつ、AI モデルの性能を最大化する一般的な手法である。最も一般的な戦略であるプールベースサンプリングは、説明変数のみを含む大規模なラベルなしデータプールから、アノテーションにより得られるモデル効用が大きいと推定されるデータを順次選択する。 $\mathcal{D}_L$  をラベル付きデータセット、 $\mathcal{D}_U$  をラベルなしデータセット、 $f$  を識別モデルとし、各データのアノテーションから得られる AI モデルの性能の期待される向上を獲得関数  $A(x)$  として定義すると、

$$x^* = \operatorname{argmax}_{x \in \mathcal{D}_U} A(x)$$

によりクエリ  $x^*$  を選択し、人がラベル  $y^*$  を付与した後  $(x^*, y^*)$  を  $\mathcal{D}_L$  に追加する（アルゴリズム 1）。獲得関数には様々な定義があるが、不確実性を評価する情報量 [36] や実データの分布に着目する代表性 [16] を定量化することが一般的である。

#### 2.3.1 BALD

BALD はベイジアン能動学習における代表的な獲得関数である。サンプル  $x$  に対するラベル  $y$  の取得がモデルパラメータ  $\theta$  の不確実性をどれだけ減らすかを相互情報量として定義する。具体的には、現在のラベル付きデータセット  $\mathcal{D}_L$  に基づく予測分布を用いて次のように表される：

$$\begin{aligned} A(x) &= I(\theta; y | x, \mathcal{D}_L) \\ &= H[p(y | x, \mathcal{D}_L)] - E_{\theta \sim p(\theta | \mathcal{D}_L)} [H[p(y | x, \theta)]]. \end{aligned} \quad (5)$$

ここで  $I[\cdot]$  は相互情報量を、 $H[\cdot]$  はエントロピーを、それぞれ表す。第一項はモデル全体としての予測エントロピー（ラベル予測の不確実性）を、第二項はモデルパラメータ  $\theta$  ごとの予測エントロピーの期待値をそれぞれ表す。実装上は、ペイズニューラルネットワークの事後分布を近似するために MC Dropout [11] が広く用いられる。すなわち、ニューラルネットワークにドロップアウトを施したまま複数回フォワードパスを行い、得られた確率分布をサンプル平均して近似することで、上式の期待値項を推定する。

### 3. 関連研究

能動学習に関する研究のほとんどが実世界データに適用した場合のものであるが、差分プライバシとの組み合わせも検討されている。しかしそれらの研究は、能動学習の枠組みで学習されたモデルにおけるデータプライバシ保護に焦点を当て、ラベルアノテーションを行う人間は単なる第三者ではなく、同意取得済みかつ信頼できることを暗黙の前提としている [2], [3], [12], [27]。例えば K. Schwethelm et al. [27] は、能動学習ではサンプル選択自体がデータ依存であるため従来の DP 保証がそのままでは成り立たないという課題を指摘し、その上でサンプル選択によりデータ利用を効率化しつつ学習更新に差分プライバシ機構を組み込むことで効率とプライバシを両立する手法を提案している。しかし彼らの枠組みでも、アノテータは信頼された存在とみなされている。アノテータに提示される時点でのデータプライバシを扱った研究は数少ないものの存在する。その場合でも、 $k$ -匿名化による性能劣化を許容するか [10]、厳格な制限のあるオープンデータに依存するか [26] であり、どちらも枠組みとして成立はするものの、性能や設定に実用性に欠けている。

また、AI ライフサイクル全体ではなくラベルアノテーションのみに着目したプライバシ保護としては、LLM による自動アノテーションを活用して人間によるラベルアノテーションを不要とする提案がされている [13], [33]。しかし、自動アノテーションの適用範囲はデータが画像やテキストで構成され、タスクが明確に定義されている場合に限られる。実世界データに AI を適用して新たな価値を創造することを目的とするような新しい AI アプリケーションの開発では、プライバシ保護の適用範囲の不足のみならず、

---

**Algorithm 2:** 提案手法のアルゴリズム

---

**Input:** Unlabeled dataset  $\mathcal{D}_U$ , initial labeled dataset  $\mathcal{D}_L$ , acquisition function  $A(\cdot)$ , query budget  $B$ , total privacy budget  $\epsilon$

**Output:** Trained model  $f$

Initialize synthetic distribution  $P_{\text{synth}}^{(0)}$ ;

**for**  $k = 1$  **to**  $K$  **do**

// AIM 外側ループ (privacy-consuming)  
Calculate normalized  $A(x)$  via sigmoid to weight  $P_{\text{true}}$  and  $P_{\text{synth}}^{(k)}$ ;  
Select marginal  $m_k$  by  
$$m_k = \arg \max_{m \in \mathcal{M}} |m_A(P_{\text{true}}) - m_A(P_{\text{synth}}^{(k)})|.$$
  
Apply DP mechanism  $M_{\epsilon_k}$  as in Eq. (2), ensuring the global constraint in Eq. (3);  
Update synthetic distribution by  
$$P_{\text{synth}}^{(k+1)} = \arg \min_P \sum_{i=1}^k \left( P(m_i) - \tilde{m}_i(X) \right)^2.$$

**for**  $t = 1$  **to**  $T_k$  **do**  
// BALD 内側ループ (annotation-consuming)  
Retrain model  $f$  and compute  $A(x)$  for synthetic samples  $x \sim P_{\text{synth}}^{(k+1)}$ ;  
Select top sample  $x^*$  and query its label  $y^*$  from annotator;  
 $D_L \leftarrow D_L \cup \{(x^*, y^*)\};$   
 $B \leftarrow B - 1;$

**return**  $f$ ;

---

このような前提は全く当てはまらない。

#### 4. タスク誘導型差分プライバシ保護合成データ生成

本論文では AIM によるプライバシ保護合成データ生成と BALD による能動学習を組み合わせることで、人間によるラベルアノテーションも含めた AI ライフサイクル全体におけるデータプライバシを保証しつつ、効率的なアノテーションにより高性能な AI モデルを構築できる手法を確立する。単純な組み合わせとしては、ラベルなし実世界データセットに対して、すべてのプライバシ予算を消費してラベルなし合成データセットを生成し、以後はその合成データセットのみを使用して能動学習を適用することで枠組みとしては成立する。しかし、能動学習では通常よりも少数の教師データで学習して次のクエリ選択を逐次的に行うため、差分プライバシノイズの影響がより大きくなり、十分な性能を期待できない。一方、能動学習の逐次処理ごとに実世界データセットを参照し、合成データを生成し直すとすると、逐次処理ごとに消費できるプライバシ予算が微小となり、合成データの品質の劣化が著しくなり、現実的でない。これはタスクによる能動学習の逐次処理回数が 1000 などのオーダーになるためである。

よって両者を単純に組み合わせるだけでなく、AIM の適

応的かつ逐次的処理の中に BALD による能動学習の逐次処理を組み込むことが必要となる。このとき、大きく 2 つの課題が生じる。まず、実世界データおよび合成データにおける BALD スコアの AIM の合成データ生成におけるパラメータへの反映方法である。次に、プライバシ予算とアノテーション予算を同時に消費することになる設定における両者の消費方法である。

マージナル選択におけるマージナル  $m_t$  の集計表算出時に、実世界データおよび合成データの各クエリを BALD スコアにより重み付けすることで実現する。この際、後続の差分プライバシ測定において付与するノイズを通常の AIM と同等とし、理論保証を担保するために、感度に影響しないことが必要となる。そこで BALD スコアをそのまま重みとせずに、sigmoid 関数を適用して  $[0, 1]$  の範囲に正規化を行う。

2 種類の予算消費に関しては、一般に AIM のラウンド数よりも能動学習のラウンド数の方が圧倒的に大きくなることから、AIM の 1 ラウンドの中で能動学習を複数ラウンド実施する。すなわち、AIM 由来の外側ループにおいて、プライバシ予算を消費して、実世界データセットを参照しつつ合成データセットの更新を行う。そして能動学習由来の内側ループにおいて、アノテーション予算を消費して、合成データセットの中からクエリを選択、アノテーションを実施する。また、初期段階ではプライバシ予算消費とアノテーション予算消費の往復を増やすことで組み合わせ効果を最大化すべく、内側ループのループ回数は外側ループのラウンドに比例して大きくなる設定を導入する。

#### 5. 実験

AIM で生成したプライバシ保護合成データにアノテーションを行うことの有効性を示すことと、能動学習を適用する際の課題および提案手法の有効性を示すことを目的として、2 つの実験を行った。1 つ目の実験では、ラベルがアノテータによって事後的に付与された場合 (Ann) と、ラベルを含めて AIM により生成された場合 (No Ann) について、プライバシ保護性能とモデル性能のトレードオフを検証する。2 つめの実験では、プライバシ保護合成データ生成と能動学習を組み合わせる際の提案手法 (Proposed) と一括で合成データを生成した後に能動学習を行う場合 (Post Ann) と合成データにおけるランダムサンプリング (Random) について、能動学習ラウンドごとのモデル性能を検証する。

##### 5.1 実験設定

UCI Machine Learning Repository [8] に公開されていて、二値分類におけるプライバシ保護合成データの評価に広く利用されている Adult データセット [19] により実験を行った。Adult データセットのタスクは、年齢や職業な

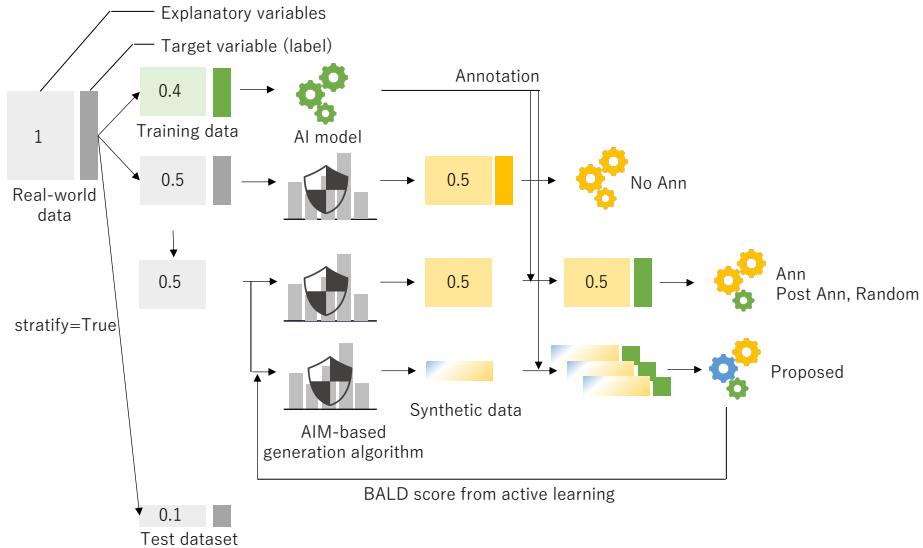


図 3 データセットの分割方法. 全量を 1 として, 0.5 を合成データ生成用, 0.4 を模擬アノテーション用, 0.1 を評価用とすることで, 検証中の実世界データの情報漏洩を防ぐ.

ど 13 の説明変数を用いて, 年収が \$50,000 を超えるか否かを予測する二値分類問題である. 欠損データを除去する前処理後の有効サンプル数は 30,162 個で, クラス別では 6,508 個と 22,654 個である. 図 3 にデータセットをどのように分割して実験に用いたかを示す. アノテーションは本来人間が行うものであるが, フレームワークの検証のために, 明に分割した実世界データで学習した教師あり学習モデルで代用した. 検証にかかる非現実的な稼働への対策のみならず, よく知られた比較的単純なタスクであれば, 機械学習モデルによるラベルアノテーションは人間によるアノテーションと類似しており [25], 恣意的な評価も防げるためである. トータルのプライバシ予算は,  $(\epsilon, \delta)$ -差分プライバシで明示的に制御した.  $\epsilon$  が小さいほどプライバシ保証が厳しいことを示し,  $\epsilon = 0.4, 0.2, 0.1$  の 3 条件を行った.  $\delta$  は  $\delta = 1.0 * 10^{-5}$  で固定とした. AIM におけるラウンド数は 20 で固定とし, 能動学習のラウンド数は 1000 で固定とした. 提案手法における  $k_{AIM}$  回目の外側ループにおける内側ループの回数は線形に増加させるため,  $k_{BALD} = 5 * k_{AIM}$  とした. 識別モデルは中間層のノード数が [512, 256, 128] のニューラルネットワークとし, 正解率を評価指標として, 各 5 回の試行における平均と標準偏差を算出した.

## 5.2 実験結果

図 4 に実験 1 の結果として, 各手法で生成した識別モデルのテストデータセットに対する正解率を示す. エラーバーは標準偏差を示す.  $\epsilon$  ごとの Ann と No Ann の結果に合わせて, 参考として実世界データで学習した結果 (Actual) を並べている. これらの結果から, No Ann では差分プライバシノイズの影響が合成データセットのラベルにも及び,

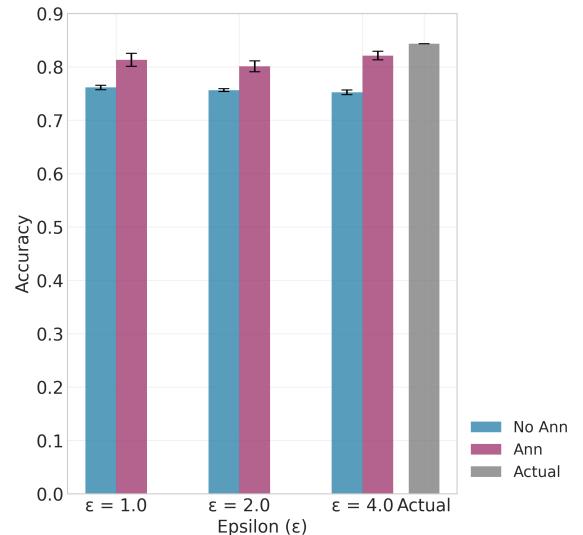


図 4 実験 1 の結果. AIM で生成したプライバシ保護合成データにラベルアノテーションを事後的に行うことでも実世界データによる学習に近い性能が得られる.

モデル性能を低下させる傾向があるが, Ann ではプライバシ保護しつつも性能の大幅な低下を効果的に防いでいることが確認できる.

図 5 に実験 2 の結果として, 各手法の各能動学習ラウンドで得られる暫定モデルのテストデータセットに対する正解率を示す. 各線の囲む網掛け部分は標準偏差を表す.  $\epsilon$  ごとの Proposed, Post Ann, Random の結果に合わせて, 参照として実世界データで学習した結果 (Actual) を並べている. 実験 1 において十分なデータ量があれば Ann と Actual に大きな差はないことが示されたが, 実験 2 では Random に加えて Post Ann も Actual から大きく下回っている. このことから少数データ環境では性能が大きく劣り, また能動学習もそのままでは期待通りに機能しないこ

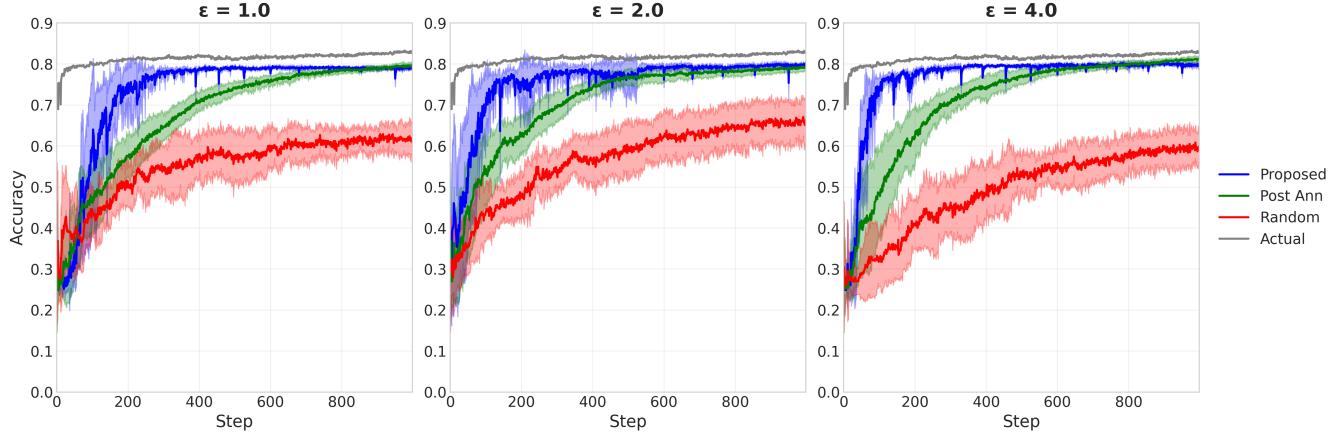


図 5  $\epsilon$  ごとの実験 2 の結果. プライバシ保護合成データにそのまま能動学習を適用しても性能が得られないが、提案手法による適応的かつ逐次的な組み込みでは実世界データ活用時に近い性能が得られる.

とが確認できる。一方 Proposed は Post Ann と Random を大きく上回り、また Actual にたいしても初期フェーズでは下回るもの、プライバシ保護を保証したうえで近しい性能を示せていることが分かる。この結果から提案手法では、差分プライバシ保護、AI モデルの有用性の最大化、人間のアノテーション作業コストの最小化の 3 つを同時に満たしていることが分かる。またラベルアノテーションの AI ライフサイクルにおける位置づけと差分プライバシの後処理定理から、本提案手法により、人間を系に含む AI ライフサイクル全体における差分プライバシ保証を実現できていると言える。

### 5.3 制約と展望

本論文では、人間を系に含む AI ライフサイクル全体における差分プライバシ保証を実現する手法を提案し、その性能を実証した。本節では、本論文における制約を明確にしたうえで、さらなる手法の高度化、実用化に向けた将来課題を示す。まず、本論文ではラベルアノテーションを実際の人間ではなく、擬似アノテータで実施している。本論文ではプライバシ保護合成データを対象としているため、実際にアノテーションを行った際の性能に差が出るのか、行いやすくするにはどのようなインターフェースがいいのかの検証が必要である。本手法は 2 クラス識別に特化しておらず、そのまま多クラス識別にも適用できるが、多クラス識別問題などより複雑なタスクにおける検証も必要である。

手法の高度化と実用化の観点では、本論文ではプライバシ予算とアノテーション予算の消費を固定している点と能動学習 1 ラウンドで抽出するクエリを 1 つに固定している点が挙げられる。AIMにおいてもプライバシ予算に対してアニーリングを行っているように、本手法においても、どちらの予算をどれだけ消費するのが適切かを同時に最適化する手法が求められる。また、BALD に対して batch

BALD [18] が提案されているように、1 ラウンドで最適な複数クエリの組み合わせを抽出できる手法が求められる。

## 6. 結論

本論文では、人間を系に含む AI ライフサイクルにプライバシ・バイ・デザインの原則を適用すべく、AI ライフサイクルの第一ステップである人間によるラベルアノテーションを差分プライバシ化することに取り組んだ。具体的には、AIM による差分プライバシ保護合成データ生成の適応的かつ逐次的なプロセスの内部に、能動学習を統合し、BALD で得られるクエリごとのスコアを考慮することで、厳密なプライバシ保証、人間のアノテーション稼働の効率化、高い AI モデル性能の獲得を同時に実現する手法を確立した。実験結果より、AIM による差分プライバシ保護合成データに対する事後的なラベルアノテーションが差分プライバシノイズの悪影響を除去する一方、そのまま能動学習を適用しても効果が得られない課題を確認したうえで、提案手法ではその課題を解決できることを実証した。今後は、提案手法の高度化、実用化を進めつつ、実際の人間が合成データにアノテーションを行う実用的な検証を行う。

## 参考文献

- [1] Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S. and Kaafar, D.: Differentially private release of datasets using Gaussian copula, *Journal of Privacy and Confidentiality*, Vol. 10, No. 2 (2020).
- [2] Balcan, M. F. and Feldman, V.: Statistical Active Learning Algorithms for Noise Tolerance and Differential Privacy, *Algorithmica*, Vol. 72, No. 1, p. 282–315 (2015).
- [3] Bittner, D. M., Brito, A. E., Ghassemi, M., Rane, S., Sarwate, A. D. and Wright, R. N.: Understanding Privacy-Utility Tradeoffs in Differentially Private Online Active Learning, *Journal of Privacy and Confidentiality*, Vol. 10, No. 2 (2021).
- [4] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D. and Wallace, E.:

- Extracting Training Data from Diffusion Models, *arXiv preprint arXiv:2301.13188* (2023).
- [5] Cavoukian, A.: Privacy by Design: Origins, Meaning, and Prospects for Assuring Privacy and Trust in the Information Era, *Privacy Protection Measures and Technologies in Business Organizations: Aspects and Standards*, pp. 170–208 (2011).
- [6] Chen, D., Yu, N., Zhang, Y. and Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models, *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362 (2020).
- [7] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. and Mahmood, F.: Synthetic data in machine learning for medicine and healthcare, *Nature Biomedical Engineering*, Vol. 5, No. 6, pp. 493–497 (2021).
- [8] Dua, D. and Graff, C.: UCI Machine Learning Repository (2017).
- [9] Dwork, C.: Differential Privacy, *Automata, Languages and Programming*, pp. 1–12 (2006).
- [10] Feyisetan, O., Drake, T., Balle, B. and Dieth, T.: Privacy-preserving active Learning on sensitive data for user intent classification (2019).
- [11] Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, *Proc. of ICML*, Vol. 48, pp. 1050–1059 (2016).
- [12] Ghassemi, M., Sarwate, A. D. and Wright, R. N.: Differentially Private Online Active Learning with Applications to Anomaly Detection, *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, p. 117–128 (2016).
- [13] Hathurusinghe, R., Nejadgholi, I. and Bolic, M.: A Privacy-Preserving Approach to Extraction of Personal Information through Automatic Annotation and Federated Learning, *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pp. 36–45 (2021).
- [14] Housby, N., Huszar, F., Ghahramani, Z. and Lengyel, M.: Bayesian Active Learning for Classification and Preference Learning, *arXiv:1112.5745* (2011).
- [15] Hu, J.: Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data, *Transactions on Data Privacy*, Vol. 12, No. 1 (2019).
- [16] Kim, Y. and Shin, B.: In Defense of Core-Set: A Density-Aware Core-Set Selection for Active Learning, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 804–812 (2022).
- [17] Kimura, E., Chida, K., Ikarashi, D., Hamada, K. and Ishihara, K.: Statistical disclosure limitation of health data based on Pk-anonymity, *Studies in health technology and informatics*, Vol. 180, pp. 1117–9 (2012).
- [18] Kirsch, A., van Amersfoort, J. and Gal, Y.: Batch-BALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).
- [19] Kohavi, R.: Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207 (1996).
- [20] Kreuzberger, D., Kuhl, N. and Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture, *IEEE Access*, Vol. 11, pp. 31866–31879 (2023).
- [21] Li, H., Xiong, L., Zhang, L. and Jiang, X.: DPSynthesizer: Differentially private data synthesizer for privacy preserving data sharing, *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, Vol. 7, No. 13, p. 1677 (2014).
- [22] Lindell, Y.: Secure Multiparty Computation, *Commun. ACM*, Vol. 64, No. 1, pp. 86–96 (2020).
- [23] McKenna, R., Miklau, G. and Sheldon, D.: Winning the NIST Contest: A scalable and general approach to differentially private synthetic data, *Journal of Privacy and Confidentiality*, Vol. 11, No. 3 (2021).
- [24] McKenna, R., Sheldon, D. and Miklau, G.: Graphical model based estimation and inference for differential privacy, *Proc. of ICML*, pp. 4435–4444 (2019).
- [25] Nanda, V., Majumdar, A., Kolling, C., Dickerson, J. P., Gummadi, K. P., Love, B. C. and Weller, A.: Do Invariances in Deep Neural Networks Align with Human Perception?, *Proc. of AAAI* (2023).
- [26] Saisho, O., Miura, T., Iwahana, K., Kii, M. and Okada, R.: Active Learning for Human Annotation of Privacy-Preserved Synthetic Data, *Proc. of PSD2024*, pp. 1–15 (2024).
- [27] Schwethelm, K., Kaiser, J., Kuntzer, J., Yigitsoy, M., Rueckert, D. and Kaassis, G.: Differentially Private Active Learning: Balancing Effective Data Selection and Privacy (2025).
- [28] Settles, B.: Active Learning Literature Survey, Technical report (2009).
- [29] Shirai, S. and Whitehill, J.: Privacy-Preserving Annotation of Face Images Through Attribute-Preserving Face Synthesis, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 21–29 (2019).
- [30] Stadler, T., Oprisanu, B. and Troncoso, C.: Synthetic Data – Anonymisation Groundhog Day, *Proc. of USENIX Security 2022*, pp. 1451–1468 (2022).
- [31] Sweeney, L.: K-Anonymity: A Model for Protecting Privacy, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, No. 5, pp. 557–570 (2002).
- [32] Takagi, S., Takahashi, T., Cao, Y. and Yoshikawa, M.: P3GM: Private High-Dimensional Data Release via Privacy Preserving Phased Generative Model, *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 169–180 (2021).
- [33] Tornberg, P.: ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning (2023).
- [34] Valadares, D. C. G., Will, N. C., Spohn, M. A., de Souza Santos, D. F., Perkusich, A. and Gorgônio, K. C.: Confidential computing in cloud/fog-based Internet of Things scenarios, *Internet of Things*, Vol. 19, p. 100543 (2022).
- [35] Xu, L., Skouliaridou, M., Cuesta-Infante, A. and Veeramachaneni, K.: Modeling Tabular data using Conditional GAN, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 7335–7345 (2019).
- [36] Yuan, J., Hou, X., Xiao, Y., Cao, D., Guan, W. and Nie, L.: Multi-criteria active deep learning for image classification, *Knowledge-Based Systems*, Vol. 172, pp. 86–94 (2019).
- [37] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. and Xiao, X.: PrivBayes: Private Data Release via Bayesian Networks, *ACM Trans. Database Syst.*, Vol. 42, No. 4 (2017).