A Survey on Differential Privacy with Machine Learning and Future Outlook

Samah Saeed Baraheem
Department of Computer Science
University of Dayton
Dayton, OH 45469
baraheems1@udayton.edu

Zhongmei Yao
Department of Computer Science
University of Dayton
Dayton, OH 45469
zyao01@udayton.edu

November 22, 2022

Abstract

Nowadays, machine learning models and applications have become increasingly pervasive. With this rapid increase in the development and employment of machine learning models, a concern regarding privacy has risen. Thus, there is a legitimate need to protect the data from leaking and from any attacks. One of the strongest and most prevalent privacy models that can be used to protect machine learning models from any attacks and vulnerabilities is differential privacy (DP). DP is strict and rigid definition of privacy, where it can guarantee that an adversary is not capable to reliably predict if a specific participant is included in the dataset or not. It works by injecting a noise to the data whether to the inputs, the outputs, the ground truth labels, the objective functions, or even to the gradients to alleviate the privacy issue and protect the data. To this end, this survey paper presents different differentially private machine learning algorithms categorized into two main categories (traditional machine learning models vs. deep learning models). Moreover, future research directions for differential privacy with machine learning algorithms are outlined.

Keywords Differential privacy · DP · Differentially private machine learning algorithms · Differential privacy with machine learning models

1 Introduction

Machine learning has proven its capability in effectively solving real-world problems, even with previously unsolvable problems. The objective of machine learning is to simulate and imitate the human behaviors so that machines and computers are able to learn and acquire new skills and/or knowledge from the given data. Therefore, machine learning has been, and still is, an active and hot topic among researchers. Recently, more and more machine learning models and applications have been developed and deployed. However, vulnerabilities and privacy leaks might be a serious threat to the participants' data. Particularly, when the dataset contains sensitive personal information. For instance, for health care applications, the dataset might include some very sensitive information, such as patient names, contact phone numbers, addresses, email addresses, dates of birth, insurance details, photo ID, and tax ID numbers. Indeed, machine learning task has the ability to extract useful and meaningful information from data to learn new knowledge/skills; and thus, it enhances the progress in companies, commerce, industry, academia, and science. Nonetheless, the ability of machine learning to capture fine-grained details may compromise data providers' privacy.

Several research [1, 2, 3] states that inferring data about specific records in the training dataset is possible even with black-box settings. Different types of attacks might threaten data privacy. These attacks can be categorized into two categories: passive adversaries, i.e., model inversion [1] and membership inference [2], and active adversaries, such as [3], just to name a few. Fredrikson et al. [1] introduced a model inversion attack. This attack can reveal the faces of participants by providing the face recognition system API along with the participant's name. Shokri et al. [2] proposed a membership inference attack. The attack is capable of predicting if the training dataset includes a particular record though black-box access to a machine learning model. Hitaj et al. [3] proposed a vigorous attack against collaborative/distributed deep learning using GANs that results in inferring sensitive information from the user's device.

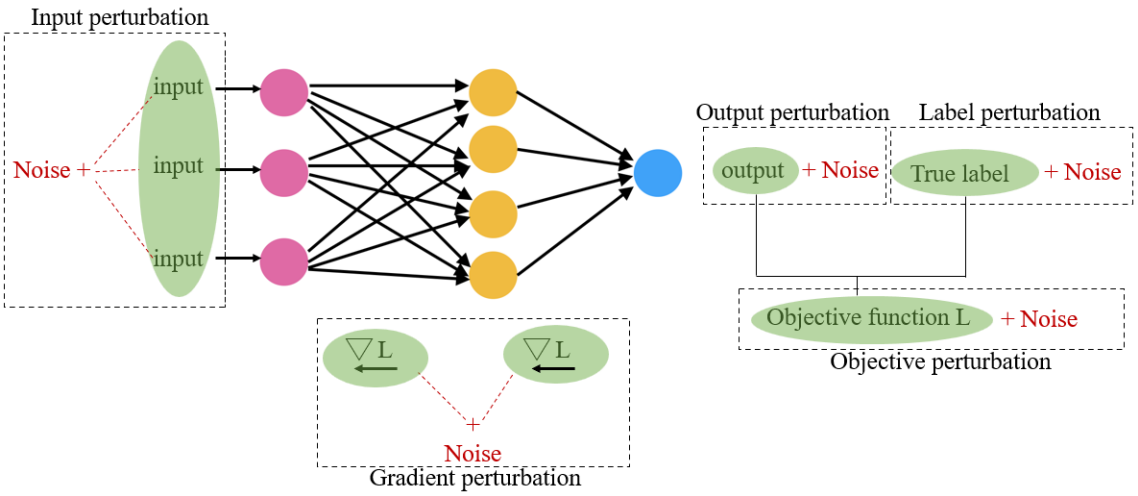

Figure 1: An overview of the differential privacy positions in the machine learning pipeline.

Hence, there is an urgent demand for privacy protection and preservation. Various solutions and mechanisms have been proposed to tackle this serious problem. One of the strongest and most popular solutions is differential privacy (DP) [4]. DP is a powerful technique for privacy guarantees. It was initially used in simple statistics [5, 6], but then the research community has leveraged DP in a machine learning environment [7, 8, 9, 10] to protect the dataset from being revealed based on the outcomes. DP works by injecting an additional statistical noise to the dataset, whether to the inputs, the outputs, the ground truth labels, the objective functions, or the gradients, as shown in Figure 1.
Furthermore, adding noise could be accomplished locally (client side) or globally (server side). Based on this, DP is categorized into two categories, global differential privacy (GDP) and local differential privacy (LDP), as illustrated in Figure 2. GDP requires a trusted data curator. The perturbation occurs at output time in the server side and by the data curator. This leads to more accurate results because the amount of added noise is not significant and happens at the end of the process. However, it is more vulnerable to attacks because the data curator needs the original data to add statistical noise. Thus, this might make the system vulnerable to attacks, i.e., membership inference [2] and model memorizing attacks [11]. As a result, this makes GDP mechanisms unsuitable for machine learning applications [12]. On the other hand, in LDP, there is no need to have data curator since the perturbation step is occurred at input time in client side and by the data owner. Thus, it is a more preservative and secure model. Therefore, it is more suitable for machine learning algorithms
The remaining sections of this paper are as follows. We first review various differentially private machine learning algorithms. Then, we conclude by suggesting possible future research directions for differential privacy with machine learning algorithms.

2 Differentially Private Machine Learning Algorithms

Machine learning is a powerful technique that extracts useful information about the underlying distribution from the given dataset. Recently, machine learning has attracted the researchers' attention due to its capability to solve even the previous unsolvable problems. However, machine learning models might leak the training data. This could be a serious threat if the information contained in the training dataset is sensitive and private. Therefore, many research has been recently conducted to protect the training data. One of the popular solutions to preserve the dataset from leaking while maintain the dataset quality is differential privacy (DP). Therefore, in this section, various differential privacy with machine learning algorithms are introduced. In general, DP guarantees privacy in machine learning models through adding noise to the inputs, outputs, ground truth labels, or even to the models. Thus, adversaries are unable to predict and infer any information for any individual record after publicly releasing the machine learning models or only the results. Figure 3 summarizes the taxonomy of differentially private machine learning algorithms. In this survey paper, differentially private machine learning algorithms are categorized into two groups based on the machine learning types (traditional or deep learning).

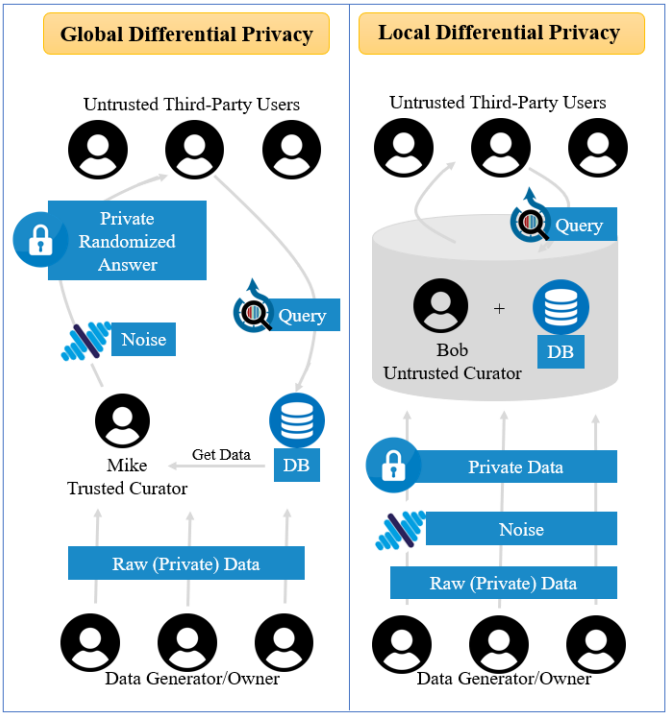

Figure 2: The difference between global differential privacy (GDP) and local differential privacy (LDP).

2.1 Differential Private Traditional Machine Learning

2.1.1 Differential Private Supervised Learning

Classification and regression models infer a variable Y based on a given combination of variables X. While the former infers categorical variable, the latter estimates a numerical variable.

Naive Bayes model [13]. It is one of the traditional machine learning algorithms which used for classification purpose. Specifically, it is a family or a collection of simple probabilistic algorithms based upon Bayes' theorem [14]. Bayes' theorem is based on computing conditional probabilities, where it attempts to find the probability of an event that is happening provided the probability of another event that has previously happened. Hence, it computes the conditional probability for all possible labels. The label with the highest probability is the predicted label.
To maintain the privacy, ϵ-differential privacy is leveraged with Naive Bayes classifier **[15]**. The idea behind it is based on the bound assumption, meaning that all features values in the training set are bounded by specific value. It could be based on the Gaussian assumption as well if the bound contains the Gaussian distribution. Next, the information sensitivity is computed, and Laplace noise is added to this information to protect the model from being compromised. Moreover, to preserve the individuals' training data, **Yilmaz et al.** [16] propose a local differential privacy (LDP) with Naive Bayes algorithm using locally differentially private frequency and statistics estimation mechanisms. This approach helps in collecting the training data to be used in Naive Bayes classifier while preserving privacy of individuals who provide the training data. In this proposed method, perturbed data from participants are first collected, where the relationships between the feature values and the corresponding labels should be maintained during collecting data. In order to maintain this relationship, each feature value and label from the user are transformed first into a new value. Then, LDP perturbation is performed. Five LDP techniques are used for perturbation which are Direct Encoding (DE), Symmetric and Optimal Unary Encoding (SUE and OUE), Summation with Histogram Encoding (SHE), and Thresholding with Histogram Encoding (THE). It works well with discrete and continuous data. However, for continuous data, the authors further apply dimensionality reduction techniques to enhance the accuracy.
As opposed to [15, 16], where the data are privately collected from a single data provider, **Li et al.** [17] propose a DP with Naive Bayes in multi-data provider setting. It works by initializing the cryptographic tools along with the public parameters. After that, each data provider encrypts their data and sends the encrypted data to the data collector. The data collector aggregates all received data from multiple owners and adds the Laplace noise based on the auxiliary information. This leads to achieve the privacy-preserving solution. Following this, the trainer receives the joint dataset

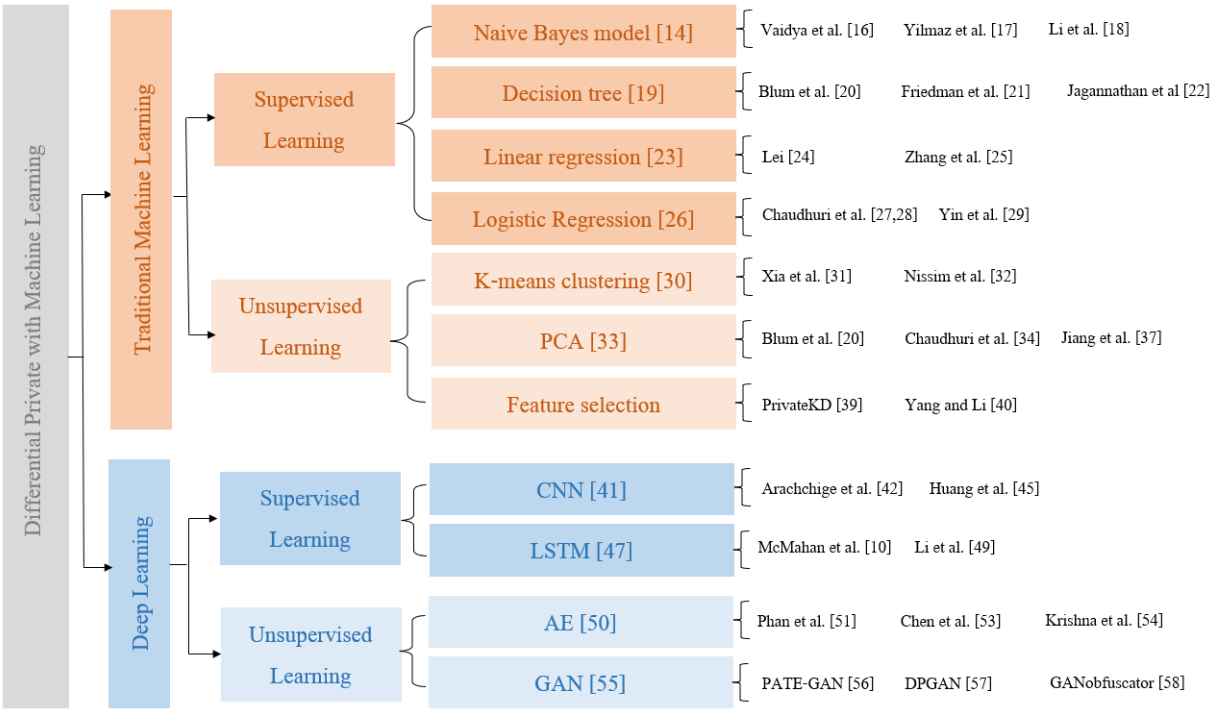

Figure 3: A summary of differentially private machine learning algorithms' taxonomy.

from the data collector and trains the model over different aggregated dataset while preserving the ownership privacy. Thus, adversaries are unable to infer whether a specific data owner holds a record that contains a particular feature value.

Decision tree [18]. It is a special form of probability tree which used to make a decision or a prediction based on prior knowledge. In particular, the data are split recursively based on the chosen features; and thus, the learning is an iterative and repetitive process. It first starts from the root which contains the whole training dataset. Then, based on the selected feature, the training set is segmented into sub-sets, where each and every sub-set satisfies the best classification. This process of choosing a feature and splitting the set into sub-sets is iterative. After that, a pruning process begins to eliminate unnecessary sub-sets from the tree and merge them to their parent nodes.
The first decision tree model without compromising the privacy is proposed in **[19]**. This model is based on the Sub-Linear Queries (SuLQ) framework that protect the privacy through adding noise to the features' information gain. To split a node into sub-nodes, the model selects a feature with less value than a particular threshold. Since the noise is added iteratively to the features in each splitting step, it may fail due to the large volume of added noise. To tackle this problem, **Friedman et al. [20]** leverage an exponential technique in the feature selection step. This leads to less amount of noise than [19]. However, [20] still results in a large amount of noise. Thus, to reduce the amount of noise in the feature selection step, **[21]** is proposed. In random decision tree with DP, the pruning phase is removed by eliminating the empty nodes. Then, reconstructing the tree in a way that all leaf nodes are put at the very same level. Following this, a Laplace noise is added to each leaf nodes which satisfies ϵ-DP.

Linear regression [22]. It is a simple traditional machine learning algorithm that is mainly used for predicting a numerical feature value (dependent variable) based on values of other features (independent variable(s)). If the number of independent variable is one, then it is called a simple linear regression. Meanwhile, when the number of independent variable is more than one, it is called multiple linear regression. As a result, both types of linear regression work by computing the conditional probability distribution of dependent variable given the values of the independent variable(s). In **[23]**, the privacy of linear regression models is preserved by generating a synthetic training dataset. Instead of using the original dataset, the data is augmented with noise. In particular, the noise is added to the inputs' histograms. This perturbed histogram guarantees privacy, but it only works well with low-dimensional training set. Additionally, it sometimes fails and obtains inaccurate regression outputs as if there is no privacy-preserving mechanism. For the above aforementioned reason, **[24]** was proposed based on functional mechanism. In this work, the objective/cost function is perturbed by inserting noise to its coefficients after approximating the objective function through Taylor expansion to

approximate it to a low order. Then, it attempts to find the weight which minimizes the approximated objective function. This leads to ϵ-differentially private linear regression model.

Logistic Regression [25]. Likewise linear regression, logistic regression works by predicting the value of the dependent variable based on its relationship to one or more independent variables. The difference is that it only works to infer a categorical variable rather than a continuous variable. Mostly, in logistic regression, a regularization term is incorporated into the adjective function to overcome the problem of overfitting. Hence, applying output perturbation or objective perturbation which add noise to the output or adjective/cost function, respectively as **in [26, 27]** can ensure privacy preserving over regularized logistic regression. However, this mechanism will not work on standard logistic regression. To tackle this problem, **Yin et al. [28]** propose to use local differential privacy (LDP) to protect the privacy of regularized and standard logistic regression. Three steps are employed, starting with adding noise, followed by selecting features, and ending with training the logistic regression model. In the first phase, Laplace mechanism is used to add noise to the training dataset. Then, feature selection starts to guarantee that the noisy data is not removed and to enhance the classification performance. At the end, a logistic regression classifier is trained on the perturbed data.

2.1.2 Differential Private Unsupervised Learning

Unsupervised traditional machine learning algorithms can be divided into three main categories, clustering, dimension reduction, and feature selection. Clustering models cluster and group unlabeled data into different clusters/groups depending on the similarity. Whereas dimension reduction models attempt to project the data from a high-dimensional space to a low-dimensional space by preserving the most useful and meaningful data, feature selection mechanisms attempt to select the most informative variables/features from the given dataset. Thus, both dimension reduction and feature selection are used as a pre-processing step for further studies.

K-means clustering [29]. It is an unsupervised learning method that works on unlabeled datasets. Thus, it works without intervention of human/annotator. Specifically, it iteratively groups the datapoints into different clusters based on the similarity, where the number of clusters (k) is predefined. Each datapoint belongs to one cluster based on the nearest centroid. Then, at each iteration, each centroid is updated based on the mean of datapoints in that cluster.
Xia et al.[30] propose the first LDP K-means clustering. They leverage LDP for privacy preservation of K-means clustering through direct and local perturbation mechanism over the training dataset. Furthermore, a budget allocation technique is used to improve the accuracy by decreasing the noise scale. An extended version is introduced in the same work to enhance both the utility and privacy. In the improvement version, not only the training data is perturbed, but also the intermediate outcomes of clusters in each iteration are perturbed. **Nissim et al. [31]** introduce differential privacy with K-means clustering through smooth sensitivity and sample-aggregate frameworks. Particularly, sample-aggregate framework randomly divides the training set into several subsets, and then apply K-means clustering on each subset individually. Thus, it produces several outputs, one output for each subset. Following this, a smooth sensitivity framework is used to add instance-specific noise to each output. This leads to add a control on the amount of noise; and thus, it improves the accuracy. Additionally, it aids in publishing the output from a differential private dense set which leads to preserve the privacy while maintaining the K-means clustering.

Principal component analysis (PCA) [32] is a popular dimension reduction model that is used to transform from a high-dimension to a low-dimension while maintaining most of the information in the original dataset. Therefore, it helps in analyzing and easily visualizing the dataset.
To preserve privacy while maintaining the performance of PCA, **Blum et al. [19]** propose to add noise to the second moment matrix. Following this, PCA is run on the perturbed matrix. However, because of the noise amount added, this approach might significantly impact the approximation quality. To tackle this issue, **Chaudhuri et al. [33]** present PPCA. This method is based on the exponential mechanism in [34] which maintain the data privacy. In PPCA, it randomly samples a k-dimensional subspace from the matrix Bingham distribution [35]. This distribution not only ensures privacy, but also ensures the quality of approximation. Hence, PPCA is a differential privacy PCA method. To enhance the privacy guarantee, **[36]** is introduced. In this approach, a novel input perturbation technique is proposed for obtaining a covariance matrix that achieves (ϵ,0)-differential private. Specifically, the Wishart distribution [37] is utilized to produce noise. Then, the Wishart noise matrix is added to the original covariance matrix, providing a noisy covariance matrix prior finding the eigenspace.

Feature selection. Feature selection is the process of reducing the number of inputs that are fed into the model later. In particular, it is the method for the automatic selection of the most informative and relevant features.
Vinterbo proposes **PrivateKD [38]**, a differential privacy feature selection for classification. Two assumptions are held in PrivateKD. The first assumption is that all variables are categorical. The second assumption is that each variable is limited to certain potential values. The idea behind private projected histogram (PPH) is that it first specifies the

number of features to be selected. Then, it incrementally selects the features. Specifically, the selected features set is first initialized to empty set, and then it adds new features one at a time through a greedy method along with the exponential mechanism to increase the distinguishability of the selected features. Thus, the new perturbed dataset is generated before feeding it to a classifier. However, this method might not work well with high-dimensional features. **Yang and Li [39]** propose a differential privacy feature selection algorithm. It relies on local learning, where each datapoint is grouped to the nearest neighbor using the Manhattan distance. This results in scaling each feature; and thus, it produces a weighted feature space. Moreover, to fit the model and overcome the overfitting problem, the logistic regression loss with L2-regularizer is incorporated. To guarantee privacy, output perturbation mechanism is used to add noise to the output based on the sensitivity analysis of the model.

2.2 Differential Private Deep Learning

Recently, due to its effectiveness and powerfulness when trained on a large dataset, deep learning has attracted many researchers in image classification, object detection, and natural language processing; just to name a few. Nevertheless, the training sensitive data are at risk of adversaries. Therefore, to alleviate privacy concerns, differential privacy mechanisms are leveraged with deep learning. In this section, different differential private deep learning models are briefly summarized into two categories (supervised and unsupervised models).

2.2.1 Differential Private Supervised Learning

Convolutional Neural Network (CNN) [40]. It is a special type of artificial neural network (ANN). It is commonly used to analyze the visual data in many applications, such as classification, recognition, and so on. The reason of its popularity is that it has the ability to automatically extract the important features during the training process.
To ensure privacy-preserving deep learning, i.e., CNN, **Arachchige et al. [41]** present a local differential privacy (LDP) deep learning algorithm named LATENT. This method allows the data providers to insert a randomization layer (i.e., LDP layer) prior data leave their devices and before the data are sent to a machine learning service. Thus, the data providers perturb the data prior releasing them. This way the data are preserved from leaking in the server side. It efficiently works with a convolutional neural network (CNN), where it first divides the CNN architecture into three layers. The first layer is the convolutional module. The second layer is the randomization module (i.e., LDP layer) which adds a privacy preserving service. The third and last layer in CNN architecture is the fully connected module. The second layer leverages the features of randomized response [42] which guarantees LDP and uses a new LDP protocol called utility enhancing randomization (UER). UER depends on a modified version of optimized unary encoding protocol (OUE) [43] to enhance the flexibility in selecting randomization probabilities by adding a coefficient epsilon (privacy budget). **Huang et al. [44]** propose a new optimization algorithm for CNN, named DPAGD-CNN. DPAGD-CNN is short for Differential Privacy Adaptive Gradient Descent for Convolutional Neural Network. Instead of using a fixed privacy budget in each iteration during training process, DPAGD-CNN adaptively allocates different privacy budget per iteration. The privacy budget is split into two parts in each iteration. While one part is used to calculate the noisy gradient, the rest is used to choose the optimal step size. In particular, in each step, different noise is injected into the gradient via adaptive approach, but the overall privacy budget should be the same. The amount of noise is bigger than gradient norm in the first iteration of the optimization since it will not impact the gradient descent direction at the beginning. However, the noise is reduced in latter iterations to ensure precise gradient descent direction. Zero-concentrated differential privacy (ZCDP) [45], a relaxed variant of differential privacy, is used.

Long Short-Term Memory Networks (LSTM) [46]. It is a type of recurrent neural network (RNN) that is able to learn long-term dependencies. Thus, LSTM is able to memorize the past and then find out patterns throughout time to obtain next guesses that make sense. It is widely used in machine translation, speech recognition, language modeling, and sentiment analysis, just to name a few.
McMahan et al. propose differentially private LSTM language models **[10]**. This model preserves privacy while it maintains the accuracy using LSTM to predict the next word in mobile keyboards. User-level privacy guarantees are achieved based on the federated averaging algorithm [47]. The federated averaging algorithm gathers many stochastic gradient descent (SGD) updates which in turns are averaged to calculate the final update. The final update is perturbed using Gaussian noise. Moreover, each-user update is clipped to make the total update bounded in L2 norm. The federated average algorithm leads to large-step updates; and hence, fewer training steps which results in better accuracy and privacy. As a result, this model achieves differential privacy without reducing the accuracy. However, the computation is increased, leading to increase the training time. **Li et al. [48]** present differential privacy deep learning called (DP-LSTM) for predicting stock price based on financial news articles and sentiment analysis. In this paper, the first step is to formulate a sentiment-ARMA based on the autoregressive moving average model (ARMA) that considers the financial news articles information to extract the information and analyze the sentiment. Following this, an LSTM is implemented. This LSTM model includes three components which are VADER, LSTM, and DP mechanism. The

model uses valence aware dictionary and sentiment reasoner (VADER) to compute the sentiment scores. VADER is a rule-based sentiment analysis and lexicon. To improve the robustness of LSTM predictions and guarantee privacy, DP mechanism is adopted by injecting noise from Laplace distribution. The model works well in predicting the stock price with less errors and high robustness.

2.2.2 Differential Private Unsupervised Learning

Autoencoders (AE) [49]. It is a type of unsupervised learning, where the training dataset only contains inputs without outputs/labels. The target values (outputs) are set to be equal to the corresponding inputs, leading to forming the task as a supervised learning task. Then, the model attempts to minimize the reconstruction error which is the difference between the original inputs and the corresponding reconstructed outputs. Therefore, it attempts to learn compressed representations of the original inputs. In general, it has two components (an encoder and a decoder). The encoder attempts to transform a high-dimensional sample to a low-dimensional representation, while the decoder attempts to reconstruct the high-dimensional sample from the low-dimensional representation.
Phan et al. [50] propose a DP-based deep autoencoders named deep private autoencoder (dPA) to predict the human behavior in a health social network while preserve privacy. Deep autoencoders [51] is a model that consists of many autoencoders. The objective is to extract useful and meaningful latent representations through an unsupervised learning. To add a privacy level to deep autoencoders in order to protect the data, objective perturbation is leveraged. Thus, it first approximates the cross-entropy objective function of the reconstructed inputs to a polynomial approximation via Taylor series. After that, noise is added into the coefficients of the polynomial approximation objective function. Finally, a normalization layer is added into the private autoencoder (PA) to guarantee privacy in a deep autoencoders since multiple private autoencoders can be stacked on top of each other. **Chen et al. [52]** propose a differentially private autoencoder-based generative model (DP-AuGM). In this model, a private data using a differentially private algorithm [7] is fed into an autoencoder during the training process. [7] uses differentially private stochastic gradient descent (DPSGD), where clipping operation is applied and Gaussian noise is added to the computed gradients. Then, the encoder is released and published, while the decoder is dropped. Following this, to encode and generate new differentially private data, a small amount of original data is fed into the encoder by the trainer. This new encoded/generated data is leveraged to train any machine learning models in the future which guarantees privacy. Thus, since it uses not only private data, but also public data, it provides high utility while preserving the data. Moreover, **Krishna et al. [53]** present an autoencoder-based differentially private text transformation (ADePT). ADePT is based on text-based autoencoder, such as LSTM sequence-to-sequence models, to transform the given text. It begins by transforming the given input text into a latent representation via the encoder (transformation phase). In this phase, a randomized algorithm is used. Specifically, the latent representation produced by the encoder is clipped, and then a Laplacian noise is added. Following this, it generates a new data based on the transformed data in previous phase through the decoder (generation phase). Thus, this model preserves the data privacy while maintains the dataset quality.

Generative Adversarial Network (GAN) [54]. GAN is unsupervised learning that consists of two networks, namely, the generator (G) and the discriminator (D). The generator tries to fool the discriminator by synthesizing and generating samples that look like the original real inputs. In the meantime, the discriminator tries to distinguish between real and synthetic samples. Thus, these two networks contest with each other via a minimax two-player game. To protect the privacy of the training data from leaking and revealing sensitive information, differential privacy (DP) is adopted in the training process of GANs.
PATE-GAN [55] is proposed to produce synthetic tabular data while maintaining the privacy. In this model, PATE method [8] is adopted to guarantee differential privacy. PATE method [8] divides the training dataset into k disjoint subsets, and then, k classifiers (in PATE-GAN K teacher discriminators) are trained individually on each subset. To produce a differentially private output during testing/classifying a new sample in PATE, a noisy aggregation of classifier outputs is performed. Furthermore, Jordon et al. propose to train a student discriminator with the synthesized data. The outputs of these synthesized data are given by the teachers through PATE method [8]. Hence, the student network is trained privately. **DPGAN [56]** is introduced to generate synthetic images without compromising privacy. It works by adding designated noise on the discriminator's gradients during the training process. In addition, gradient clipping is applied to enforce iterative gradient descent. **GANobfuscator [57]** is proposed to alleviate information leakage when GAN is used. In particular, a designated noise is added to the gradients during the learning process, and gradient clipping is leveraged to enhance the stability of the training process and improve the privacy as in DPGAN [56]. However, in DPGAN [56], the model weights are clipped to a bounded box $[-c_p, c_p]$ which in turn automatically bounds the gradients by constant c_g. Meanwhile, GANobfuscator [57] incorporates wasserstein GAN (WGAN) [58] and adaptive clipping. In adaptive clipping, public data is used to adjust the gradients' clipping bounds throughout the training process.

3 DP Future Research Directions

Although the aforementioned surveyed studies have shown great success in preventing machine learning models from leaking sensitive information and satisfying differential privacy requirement, there are some existing challenges that need to be addressed in future work. To this end, in this section, future research directions for differential privacy with machine learning algorithms are discussed.

- One major element that plays a key role in differential privacy is ϵ value since the amount of added noise depends on this value. The amount of added noise in turn affects the utility and accuracy of the model. The larger the ϵ value, the less noise is added; and hence, the model is more accurate but might be less privacy-preserving and vice versa. Thus, choosing the proper ϵ value is crucial; and consequently, more research should be conducted to determine the right ϵ value automatically.
- Further, differential privacy is a hot topic among researchers and in academic settings. However, it is less used in industry and practical applications because its privacy guarantee is strong which might affect the utility and accuracy of the models. Thus, one possible direction for future work can be through leveraging a relaxation to DP, i.e., $(\epsilon; \delta)$-LDP, on industrial practical applications. Since some dimensions/variables are not sensitive and cannot reveal any sensitive information, a relaxed DP should be used. For instance, country, state, or even city is not sensitive. Hence, participants probably do not care to reveal this information, but they really care about the precise and exact location which should have a strong and high level of privacy-preserving.
- Moreover, a hybrid method, that allows some participants to report their real values while perturbing other participants' values, may help in reducing the amount of noise added to the dataset. Thus, this might enhance the overall model accuracy while still protecting the data because privacy preferences are different among participants. Having an access even to a small subset of public dataset without perturbation may aid in estimating the hyper-parameters; and thus, improving the model accuracy and utility.
- In addition, an adjustable and adaptive privacy budget is a good approach instead of static and equal assignment of the privacy budget to every layer of the model.
- Furthermore, using DP with complex machine learning models has many challenges. First, differentially private complex machine learning models increase the information leaking risk. Hence, they require excessive and more sanitization. This might lead to distort the data which in turn might affect the utility and accuracy of the models. Second, differentially private complex machine learning models are probably time-consuming in terms of computation. Therefore, it could be hard for researchers to incorporate DP with complex machine learning models. Consequently, new designated distributed protocols might be one possible solution for DP with complex models.
- Another path for future work could be by incorporating cryptography techniques with differential private machine learning algorithms, particularly, through Distributed Differential Privacy (DDP).
- Since differential privacy has its limitations especially when the dataset size is large due to time complexity and data utility, leveraging other privacy preserving approaches with machine learning should be explored. Examples of these other privacy preserving techniques are outline as follows.
  - **Homomorphic encryption (HE) [59]:** The data owner encrypts the data before sending the data to the model especially if the data provider uses the cloud to train the model. Then, the cloud homomorphically trains the model on the encrypted data and returns the encrypted result. Following this, the data owner can decrypt and share the result. Hence, the participants' data are protected; however, it is time-consuming where the complexity of computation is increased.
  - **Secure multiparty computation (MPC) [60]:** This technique allows training the model on data from different parties without having the parties to share their data with any party. Thus, the data privacy is preserved; however, it is time-consuming even though it is less computation than HE.
  - **K-anonymity [61]:** Anonymization can be done by altering the data before feeding it to the model. In particular, k-anonymity modifies the data with k value, where k denotes the number of indistinguishable records. Generalization and suppression [62] are some techniques that can be used to modify data and achieve anonymization. Generalization works by replacing the data with a broader/interval one, i.e., age of 50 can be replaced with [50,59]. Suppression works by removing sensitive data either removing the entire attribute values or some values of the attribute in some records. The downside of this mechanism is that it is vulnerable to background knowledge attack and homogeneity attack.
  - **L-diversity [63]:** It is an extension of k-anonymity [61], where it requires at least L well-represented values to be existed for sensitive attributes for every equivalence class. One main issue with L-diversity is that when the whole rows in a table are distributed into few equivalence classes, semantic closeness could lead to information revealing. Thus, it might be vulnerable to similarity attack.

- **T-closeness [64]:** It is an extension of L-diversity [63], where equivalence class is a T-closeness if the distance between the distributions of sensitive feature in the class and the distribution of the feature in the entire table is less than or equal a specific threshold based on the Earth Mover's Distance[65]. This approach guarantees data privacy, but the data distribution may not be appropriate each time when implementing T-closeness.
- **Condensation approach [66]:** it starts by constructing restricted clusters in the dataset. Then, it produces pseudo-data based on the statistics of the constructed clusters. The restrictions on the clusters depend on the cluster size that should be selected in a way to maintain k-anonymity [61]. This approach efficiently preserves the data privacy, but it might disclose some sensitive data because of the similarity in terms of values between some records in the generated/constructed data and original data.
- **Data distribution approach [67, 68]:** It works by distributing the data across multiple sites in two ways (horizontal and vertical distributions). Horizontal distribution [67] works by partitioning the records of the dataset across several entities, each partition has same attributes. Meanwhile, vertical distribution [68] works by partitioning the attributes across many entities, each attribute holds all records.

4 Conclusion

Differential privacy (DP) is a rigid definition of privacy based on injecting noise into the data. Different approaches can be used to inject a specific amount of noise into the data. The noise can be added to the inputs, the outputs, the ground truth labels, the objective functions, or even to the gradients to mitigate the privacy concern and preserve the data. Different mechanisms can be leveraged, such as Laplacian, Gaussian, exponential, and randomized response mechanisms, just to name a few.
Recently, after the advancement in machine learning and deep learning, more research has been conducted in this field due to its capability in solving real-world problems. However, vulnerabilities and revealing sensitive information may be a big concern with machine learning. Thus, DP can be incorporated with machine learning models to preserve the data from leaking and revealing sensitive information.
Therefore, this paper reviews various differentially private machine learning algorithms grouped into two categories: traditional machine learning models and deep learning models. Additionally, it concludes by discussing the future outlook and directions for differentially private with machine learning algorithms.

Acknowledgments

The first author would like to thank Umm Al-Qura University, in Saudi Arabia, for the continuous support.

References

[1] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15*, New York, New York, USA, 2015. ACM Press.

[2] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016.

[3] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: Information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2017. ACM.

[4] Cynthia Dwork. *Differential Privacy*, page 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09*, New York, New York, USA, 2009. ACM Press.

[6] Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2), 2010.

[7] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, 2016. ACM.

[8] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv [stat.ML]*, 2016.

[9] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv [stat.ML]*, 2018.

[10] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv [cs.LG]*, 2017.

[11] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX Security Symposium (USENIX Security 20)*, page 1605–1622, 2020.

[12] Mahawaga Arachchige Pathum Chamikara, Peter Bertók, Ibrahim Khalil, Dongxi Liu, and Seyit Camtepe. Local differential privacy for deep learning. *CoRR*, abs/1908.02997, 2019.

[13] Daniel Lowd and Pedro Domingos. Naive bayes models for probability estimation. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, New York, New York, USA, 2005. ACM Press.

[14] G. Coletti and R. Scozzafava. A coherent qualitative bayes' theorem and its application in artificial intelligence. In *1993 (2nd) International Symposium on Uncertainty Modeling and Analysis*. IEEE Computers Society Press, 2002.

[15] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, page 571–576. IEEE, 2013.

[16] Emre Yilmaz, Mohammad Al-Rubaie, and J. Morris Chang. Locally differentially private naive bayes classification. *arXiv [cs.LG]*, 2019.

[17] Tong Li, Jin Li, Zheli Liu, Ping Li, and Chunfu Jia. Differentially private naive bayes learning over multiple data sources. *Information sciences*, 444:89–104, 2018.

[18] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An introduction to decision tree modeling. *Journal of chemometrics*, 18(6):275–285, 2004.

[19] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '05*, New York, New York, USA, 2005. ACM Press.

[20] Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, New York, New York, USA, 2010. ACM Press.

[21] Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N. Wright. A practical differentially private random decision tree classifier. *Transactions on data privacy*, 5(1):273–295, 2012.

[22] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression: Linear regression. *Wiley interdisciplinary reviews. Computational statistics*, 4(3):275–294, 2012.

[23] Jing Lei. Differentially private m-estimators. *NIPS*, 2011.

[24] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 5(11):1364–1375, 2012.

[25] Michael P. LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008.

[26] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[27] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *arXiv [cs.LG]*, 2009.

[28] Chunyong Yin, Biao Zhou, Zhichao Yin, and Jin Wang. Local privacy protection classification based on human-centric computing. *Human-centric computing and information sciences*, 9(1), 2019.

[29] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. An efficient k-means clustering algorithm. 1997.

[30] Chang Xia, Jingyu Hua, Wei Tong, and Sheng Zhong. Distributed k-means clustering guaranteeing local differential privacy. *Computers and security*, 90(101699):101699, 2020.

[31] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing - STOC '07*, New York, New York, USA, 2007. ACM Press.

[32] Hervé Abdi and Lynne J. Williams. Principal component analysis: Principal component analysis. *Wiley interdisciplinary reviews. Computational statistics*, 2(4):433–459, 2010.

[33] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. Near-optimal algorithms for differentially-private principal components. *arXiv [stat.ML]*, 2012.

[34] Per Austrin. Towards sharp inapproximability for any 2-csp. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, page 307–317. IEEE, 2007.

[35] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer New York, New York, NY, 2003.

[36] Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 30(1):1730–1736, 2016.

[37] Ingram Olkin and Herman Rubin. A characterization of the wishart distribution. *The annals of mathematical statistics*, 33(4):1272–1280, 1962.

[38] Staal A. Vinterbo. *Differentially private projected histograms: Construction and use for prediction*, page 19–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[39] Jun Yang and Yun Li. Differentially private feature selection. In *2014 International Joint Conference on Neural Networks (IJCNN)*, page 4182–4189. IEEE, 2014.

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, 86(11):2278–2324, 1998.

[41] Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE internet of things journal*, 7(7):5827–5842, 2020.

[42] James Alan Fox. *Randomized response and related methods: Surveying sensitive data*. SAGE Publications, Thousand Oaks, CA, 2 edition, 2015.

[43] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, page 729–745, 2017.

[44] Xixi Huang, Jian Guan, Bin Zhang, Shuhan Qi, Xuan Wang, and Qing Liao. Differentially private convolutional neural networks with adaptive gradient descent. In *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, page 642–648. IEEE, 2019.

[45] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *arXiv [cs.CR]*, 2016.

[46] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[47] H. B. McMahan, Eider Moore, D. Ramage, and B. A. Y. Arcas. Federated learning of deep networks using model averaging. *ArXiv*, 2016.

[48] Xinyi Li, Yinchuan Li, Hongyang Yang, Liuqing Yang, and Xiao-Yang Liu. Dp-lstm: Differential privacy-inspired lstm for stock prediction using financial news. *arXiv [q-fin.ST]*, 2019.

[49] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv [cs.LG]*, 2020.

[50] Nhathai Phan, Yue Wang, Xintao Wu, and Dejing Dou. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 30(1):1309–1316, 2016.

[51] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2(1):1–127, 2009.

[52] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. Differentially private data generative models. *arXiv [cs.CR]*, 2018.

[53] Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. Adept: Auto-encoder based differentially private text transformation. *arXiv [cs.CR]*, 2021.

[54] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv [stat.ML]*, 2014.

[55] James Jordon, Jinsung Yoon, and M. Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. *ICLR*, 2018.

[56] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv [cs.LG]*, 2018.

[57] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE transactions on information forensics and security*, 14(9):2358–2371, 2019.

[58] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc.

[59] Abbas Acar, Hidayet Aksu, A. Selcuk Uluagac, and Mauro Conti. A survey on homomorphic encryption schemes: Theory and implementation. *arXiv [cs.CR]*, 2017.

[60] Yehuda Lindell. Secure multiparty computation (mpc). *IACR Cryptol. ePrint Arch.*, 2020.

[61] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, page 217–228. IEEE, 2005.

[62] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[63] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, page 24–24. IEEE, 2006.

[64] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. T-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, page 106–115. IEEE, 2007.

[65] Yossi Rubner. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[66] Charu C. Aggarwal and Philip S. Yu. *A condensation approach to privacy preserving data mining*, page 183–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[67] Vincent Yan Fu Tan and See-Kiong Ng. *Privacy-preserving sharing of horizontally-distributed private data for constructing accurate classifiers*, page 116–137. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[68] Yi Deng, Xiaoqian Jiang, and Qi Long. Privacy-preserving methods for vertically partitioned incomplete data. *AMIA Annual Symposium proceedings*, 2020:348–357, 2020.