IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 24, 2019, accepted December 12, 2019, date of publication December 25, 2019,
date of current version January 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962368

Comprehensive Survey on Big Data Privacy Protection

MOHAMMED BINJUBEIR[1], ABDULGHANI ALI AHMED[1], (Senior Member, IEEE), MOHD ARFIAN BIN ISMAIL[1], ALI SAFAA SADIQ[2,3], (Senior Member, IEEE), AND MUHAMMAD KHURRAM KHAN[4], (Senior Member, IEEE)

[1]Faculty of Computer Systems and Software Engineering, University Malaysia Pahang, Kuantan 26300, Malaysia
[2]Wolverhampton Cyber Research Institute, School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton WV1 LY, U.K.
[3]Center of Artificial Intelligence Research and Optimization, Torrens University, Brisbane, QLD 4006, Australia
[4]Center of Excellence in Information Assurance, King Saud University, Riyadh 12372, Saudi Arabia

Corresponding author: Muhammad Khurram Khan (mkhurram@ksu.edu.sa)

This work was supported by the Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang under the internal grant No. RDU190311 and Fundamental Research Grant Scheme (FRGS) with Vot No. RDU190113 and in part by the Researchers Supporting Project under RSP-2019/12, King Saud University, Riyadh, Saudi Arabia.

ABSTRACT In recent years, the ever-mounting problem of Internet phishing has been threatening the secure propagation of sensitive data over the web, thereby resulting in either outright decline of data distribution or inaccurate data distribution from several data providers. Therefore, user privacy has evolved into a critical issue in various data mining operations. User privacy has turned out to be a foremost criterion for allowing the transfer of confidential information. The intense surge in storing the personal data of customers (i.e., big data) has resulted in a new research area, which is referred to as privacy-preserving data mining (PPDM). A key issue of PPDM is how to manipulate data using a specific approach to enable the development of a good data mining model on modified data, thereby meeting a specified privacy need with minimum loss of information for the intended data analysis task. The current review study aims to utilize the tasks of data mining operations without risking the security of individuals' sensitive information, particularly at the record level. To this end, PPDM techniques are reviewed and classified using various approaches for data modification. Furthermore, a critical comparative analysis is performed for the advantages and drawbacks of PPDM techniques. This review study also elaborates on the existing challenges and unresolved issues in PPDM.

INDEX TERMS Security, big data, privacy protection, privacy-preserving data mining.

I. INTRODUCTION

Recently, various organizations in different sectors (e.g., government, banking, medical, and insurance sectors, as well as public and private institutions) have been striving to make their data electronically available. That is, these organizations have been collecting the data of their clients or users for exploration, analysis, research, or any other purposes. In several instances, the output data size comprises terabytes of huge and complex data, which is defined as big data [1].

Recently, some researchers considered big data as the revolution of the digital era compared to ''the new oil'' in terms of significance to the society [2], [3]. Most of these data are often unstructured or complex; a significant portion of the data is generated from several sources, such as business sales records, sensors used in the internet of things, social media, medical patient records in healthcare organizations, video and image archives [4].

The practice of extracting patterns (i.e., knowledge) from big data sets is conducted to generate new or useful information, which can be used to represent, interpret, or discover interesting patterns. This practice is referred to as data mining, which is an interdisciplinary subfield of computer science [5]–[9]. The term ''data mining'' has been considered a substitute explanatory term for ''knowledge discovery from data'' (KDD), which is another term that denotes the goal of data mining. Data mining methods involve patterns of discovery and extraction. These methods also encompass patterns of recognition techniques and infer algorithms that are recurrently applied in data mining. However, data

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-hoon Kim.

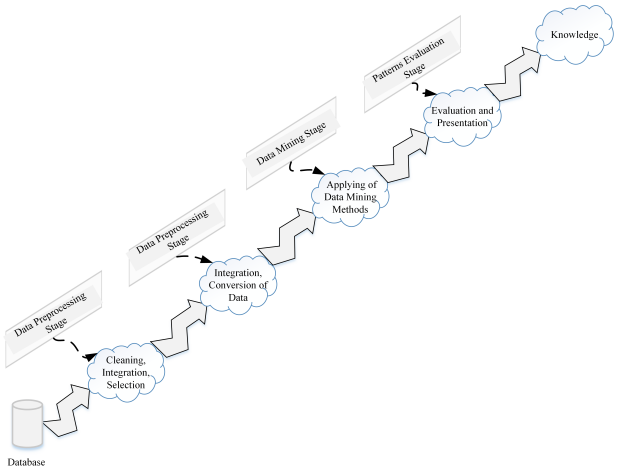

FIGURE 1. Outline of KDD.

mining in certain cases is simply a basic stage in the course of knowledge discovery because it contains varied stages. Figure 1 shows an iterative pattern and interactive of the four stages of knowledge discovery [5], [6].

Stage 1: Data preprocessing: involves data selection, data cleaning (i.e., to eliminate noise and redundant data), and data integration (i.e., to integrate data collected from various sources);

Stage 2: Data transformation: entails the integration and conversion of data into formats that are suitable for mining;

Stage 3: Data mining: involves the implementation of intelligent methods to mine data sequences (e.g., classification rules, clusters); and

Stage 4: Basic operations (i.e., pattern evaluation and knowledge presentation): involves the identification of fascinating patterns that represent knowledge. These basic operations also include showing the mined knowledge in an easy-to-comprehend manner.

As shown in Figure 1, KDD consists of various stages of operations. The mining stage and other stages of the KDD operations have resulted in the emergence of many privacy-related issues, such as data phishing, which have evolved into one of the foremost drawbacks affecting the advancement of big data [1]. Data phishing can arise in one stage of KDD, such as data preprocessing, or possibly in the delivery of the mining results, with each stage viewing the security issue from its own standpoint [5], [10]. Apart from the importance of the mining stage, which is significant in many applications, an increasing concern has been focused on the privacy threats that emerge from data mining. Consequently, numerous establishments regularly need to distribute partial data, which can be useful in enhancing the efficiency of organizations and aid their future plans [5], [11]–[13]. However, the human processing level is known for collecting large volumes of data, which increase exponentially [14]. Thus, the privacy of individuals may be violated as a result of such reasons as the unlawful entry into personal information, unwanted unearthing of individuals' disturbing private data, and usage of personal information for purposes unrelated to the original reasons for the data collection [5]. Evidently, this gap is an opportunity to improve the KDD field and resolve issues on privacy; therefore, filling in this gap becomes increasingly important and necessary with the advancements in learning technology [1], [5], [14].

To deal with the privacy issues during data mining, a sub field of data mining, referred to as privacy preserving data mining (PPDM) has gained a great development in recent years. PPDM is a subfield of data mining and has been extensively studied recently. A key issue of PPDM is how to preserve the utility of the data and safeguard sensitive information from unsolicited or unsanctioned disclosure [5].

The remainder of this paper is organized as follows. Section II describes the privacy-preserving data mining (PPDM) methods and their application in the data preprocessing stage. Section III discusses the privacy preservation in the data preprocessing stage. Section IV concisely reviews the data modification approaches. Section V discusses the primary tasks of data mining. Section VI concludes this study.

II. PPDM

Pervasive computing, which is also referred to as ubiquitous computing, involves generating large volumes of data, which is the concept known as big data. The analysis of big data has been confirmed to be a driver of development and advantageous to numerous services, including health care, banking, cyber security, commerce, and transport [11]. Organizations distribute data among themselves and share their data with the public owing to the interest in sharing reciprocal benefits and the requirements of publishing some data. The mining community eventually realized privacy issues when people publish their specific data in their original form, thereby possibly leading to violations of people privacy. They also realized the phishing of data over the Internet, which arises because data can contain some confidential information. In addition, recent advancements in the field of learning technology have significantly threatened individuals' privacy [9], [15]. Large investments have also been committed to issues on privacy protection, such as privacy preserving data publishing (PPDP) [16] and privacy aware learning [17].

Evidently, the concept of privacy should be defined. Although privacy has various definitions, providing an accepted standard definition of this concept is difficult [11], [12], [18], [19]. Privacy was established as a right in the Universal Declaration of Human Rights [18] in 1948. Nonetheless, this right is considered in an extremely limited scope because privacy can be found in specific contexts, such as correspondence, at home, and with family. According to [12], [18], the scope of information privacy is described in forms of bodily privacy, communication privacy, and territorial privacy as illustrated in Figure 2.

Information privacy concerns gathering and managing personal data. Bodily privacy is related to the protection of the bodies of individuals from invasive measures, such as drug testing and others. The privacy of communications entails any

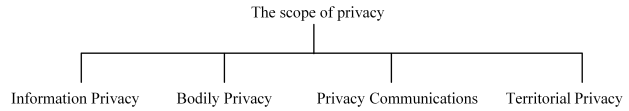

FIGURE 2. Scope of privacy.

form of communication. Lastly, territorial privacy focuses on placing boundaries on incursion into a local environment.

The scope of this study is centered on the form of information privacy. In general, the study of information privacy focuses on content privacy and interaction privacy. Content privacy refers to the prevention of disclosure of individuals' identities from an anonymized or encrypted database, such as extracting information from their credit card records from a state or national level database. By contrast, interaction privacy refers to the prevention of disclosure of a given content of an individual, such as checking victims' encrypted web traffic or using voice fingerprint to access services [1].

Thus, the current study adopts the definition of privacy in the contexts of content and interaction [12], [18], which is related to research path in terms of the collection and analysis of individual data. This can be valuable in boosting the effectiveness of organizations or support prospective plans. Furthermore, they contain some sensitive data on individuals, whose privacy is also threatened. However, transforming data or anonymizing individuals may minimize the utility of the transferred data and lead to inaccurate knowledge [12], [20]. Hence, numerous endeavors have been dedicated to privacy, which involve the preservation of individuals' information using data mining algorithms, to avert the disclosure of individuals' identities or sensitive data in the course of knowledge discovery [21]. This paradigm is referred to as PPDM. Hence, PPDM [12], [22]–[24] is an innovative research path concerned with providing guarantee to a certain level of privacy and security for big data in the application of mining research and statistical records.

Conversely, secrecy is the protection of people's information from unauthorized disclosure, alteration, or loss when transferred over a network. As the data reach the data collection point, no additional restrictions are levied on data security to disclose the personal data of persons. Therefore, data security should be correlated with data privacy because the former is a requirement of the latter. Privacy is specific, and can be achieved by hiding people's identity or screening personal information that may result in the people's recognition [25].

PPDM has recently garnered considerable interest among academics and designers. Consequently, several methods have been developed to protect privacy or far-reaching policies have been imposed for sensitive data protection [12], [21], [25]. The form of privacy varies depending on the data used and the way they are used; hence, many methods are used to provide privacy [25]. At present, no existing generic solutions can handle all privacy issues regarding the protection of sensitive information from unwanted

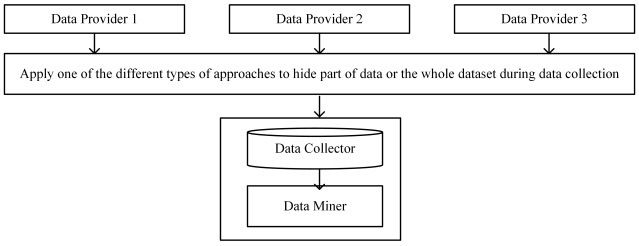

FIGURE 3. Describe user roles scenario of data that aims to safeguard raw data from being divulged.

disclosure while simultaneously preserving the utility of the data. The related studies have solely focused on searching for effective protocols for specific problems. Nonetheless, data utility and information loss are trade-offs when effective data mining is conducted for privacy measures [25]–[29]. In this survey, the privacy preservation in the big data life cycle is considered at the data preprocessing and data mining task stages.

III. PRIVACY PRESERVATION IN THE DATA PREPROCESSING STAGE

By observing the four different stages of KDD (Figure 3), privacy disclosure can occur when private data are transmitted from one stage to another. Thus, preventing private information disclosure reduces data utility, which can produce erroneous or even infeasible extraction of knowledge through data mining. An important issue of KDD is how to transmit the minimum necessary private data for data mining among the various KDD stages [27]. A commonly used privacy protection measure is to enforce privacy preservation in the data preprocessing stage by different user roles of data, which aims to protect raw data from disclosure. In general, user roles have two different types that prevent disclosure of private information in the data preprocessing stage [5].

A. PRIVACY PRESERVATION IN DATA PROVIDERS

Data providers are data owners (i.e., individuals or organizations) who are expected to provide their original raw data to data collectors (which hold data warehouse servers) that could contain some sensitive information (e.g., academic records of students, financial transcripts of customers). The main issue of data providers is their ability or inability to control the sensitivity of data they provide to data collectors. The theory is that data collectors are unreliable. Therefore, the data provider protocols (which protect privacy during data generation and data transmission to data warehouse servers) considerably aim to hide their sensitive information or prevent unauthorized access to prevent privacy disclosure and obtain adequate returns for the possible loss in privacy. However, the following question should be answered: What type of and how much information that counterpart individuals can acquire from their data? [1]. The data provider approach to disclosure behaves in line with one of the following policies.

1. Data providers cannot disclose any information because they regard their data as extremely sensitive. They decline the command to provide such information and endeavor to take effective measures to safeguard sensitive data.
2. Data providers opt to never to release person-identifiable information (private) because they are aware of the value of their data to data collectors. Accordingly, data providers distort their data that will be transmitted to data collectors to prevent true information from being easily revealed.
3. Data providers may be willing to disclose some of their sensitive information for specific rewards, such as improved services or financial benefits. They are requiring understanding how to negotiate with the data collector to obtain sufficient reimbursement for any potential loss in privacy.
4. If data providers cannot either block access to their sensitive personal information or make a profitable transaction with the data collector, then data collectors can misrepresent data collected by the data providers in such a manner that actual information cannot be easily revealed.

B. PRIVACY PRESERVATION IN DATA COLLECTION

In data warehouse servers, data collectors collect large amounts of data from data providers to maintain the ensuing data mining operations and stored in well-disciplined physical structures (e.g., multi-dimensional data cube). The data collected possibly holds the sensitive personal data of individuals. Thus, the goal of preserving data privacy is to safeguard privacy during data collection and transmission to different data mining servers by finding the minimum portion of private information required to construct accurate data mining models [30], [31]. The direct disclosure of data to data miners will infringe on the privacy of data providers [1], particularly in cases where data miners execute mining algorithms using the data provided by data collectors and extract valuable information from data. In accordance with the adopted techniques for ensuring privacy during data collection, three types of approaches have been generally developed to conceal the raw data from their original value [32]. As illustrated in Figure 3, these approaches are data exchange [33], data cryptographic [34], and data modification [35], [36].

With the data exchange technique, private information can be disseminated from (at least) one data provider to another. Hence, this technique is only applicable in systems with trusted data providers. That is, any of the data providers have no intention to compromise the disseminated private information. In the majority of practical systems, data providers are not trusted because they may want to compromise the disseminated private data. Hence, private information cannot be protected from compromise with this data exchange technique [37], [38].

Cryptography mainly provides the security concepts required for information. Cryptography has several definitions but the simplest is the art of writing in secret characters [39]. The majority of the cryptographic algorithms are based on the difficulty involved in solving difficult mathematical problems. In addition, the initial three main concepts of information security are confidentiality, integrity, and availability, which are referred to as the CIA triad [40]. Since the CIA triad was established, additional concepts, such as authenticity, accountability, privacy, and non-repudiation, have been developed [41]. Initially, an original comprehensible message, which is referred to as the plainest, is inputted into an algorithm. The algorithm conducts different tasks and transforms the plaintext to a scrambled unintelligible message called ciphertext. The conversion process is referred to as encryption or enciphering. The counter procedure of generating the plaintext from a ciphertext is known as decryption or deciphering

In cryptography, multiple parties (i.e., data providers) typically cooperate for the computation of results or jointly participate in analyzing non-sensitive information, where pairs of public and private keys are available to each data provider. Moreover, the public keys of all data providers should be distributed to everyone, including the data warehouse servers (data collectors). Initially, all data providers are provided with the sum of the public keys as their reference to encrypt their data on the basis of the provided reference for onward transmission to the data warehouse servers. Hence, no involved people know anything beyond their own input. Through mathematical manipulations, accurate models can be built by the data warehouse servers on the basis of the received encrypted data; these models can be used to solve PPDM problems among mutual untrusted parties or competitors [42], [43]. However, the complexity of this method may lead to large computational costs with enormous data for data providers and data warehouse servers, thereby making this method practically useless [44], [45]. Given the data modification is the main focus of this review paper, it is further reviewed and deeply analyzed in the following section.

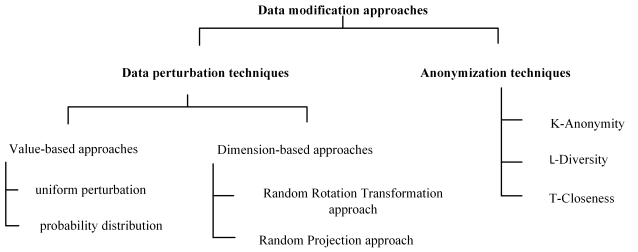

FIGURE 4. Classification framework of the data modification approaches.

IV. DATA MODIFICATION APPROACHES

Data modification approaches can be classified into two categories in accordance with the type of privacy protection: data perturbation and anonymization-based techniques. Figure 4 shows the recommended classification framework [46].

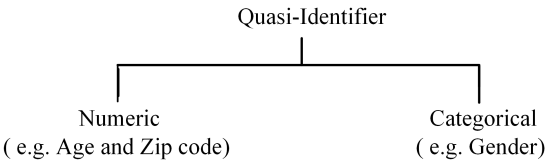

FIGURE 5. Types of QI attributes.

TABLE 1. Medical patient database.

Identifier (IAs)	Quasi-Identifier (QI)			Sensitive (SA)
Name	Age	Gender	Zip code	Disease
Bob	29	Male	462350	Heart Disease
Mike	22	Male	462351	Cancer
Michel	27	Male	462352	Flu
David	43	Male	462350	Heart Disease
Alice	52	Female	462350	Heart Disease
Sofia	38	Female	462350	Heart Disease
Carl's	33	Female	462355	Heart Disease
Abraham	49	Male	462356	Heart Disease
William	39	Male	462355	Cancer
Linda	41	Female	462351	Heart Disease
Camila	28	Female	462356	Heart Disease

Anonymization techniques attempt to prevent attempts to recognize the records' owner identity. When data perturbation techniques are used, data providers need to (independently) modify their original data before sending them to the data warehouse servers [35], [36], [47]. In this manner, garbled values are used instead of original values when applying data mining algorithms, thereby ensuring the privacy of individuals. The following subsections will describe in detail each of these approaches. Readers can refer to [12], [48], [49], and [30] for a comprehensive study on these subjects.

A. ANONYMIZATION TECHNIQUES

The collected data should be treated as a private table that encompasses multiple records (see Table 1) [50]. Each record (row) represents a single client and comprises several attributes that are specific to a particular individual [49]. These attributes can be categorized into three [27], [51]: Identity attributes (IA) explicitly identify the records of an owner (e.g., name, cellular phone number, social security number, and driver's license number). Quasi-identifier (QI) attributes denote a sequence of individuals' non-explicit attributes (e.g., race, age, date of birth, ZIP code, and gender), which can potentially identify the records of owners; sensitive attributes (SA). QI attributes consist of two types: numeric and categorical (Figure 5). SAs contain confidential data of individuals, such as salary and disease [52].

Entities may have the intention to publish partial data derived from big data sets that can be valuable in raising the effectiveness of the entities and aid their prospective plans without divulging the proprietorship of the sensitive data. Solely eliminating attributes (IAs), which explicitly identify users from the table prior to disclosing them, has been demonstrated to be inefficient [1], [27]. In this setting, the effective preservation of privacy can be attained by controlling for the disclosure of information. That is, a person identification (QI) attribute, which represents a set of individuals' non-explicit attributes by using anonymization techniques prior to release, is a previous technique used for the preservation of privacy and functions as a development platform for advanced convolution techniques. Examples of such a technique are explained in the following subsections [12].

TABLE 2. Three anonymous versions of the medical patient database relating to Table 1.

ID	equivalence class	Age	Gender	Zip code	Disease
1	1	2*	Person	462***	Heart Disease
2		2*	Person	462***	Heart Disease
3		2*	Person	462***	Flu
4		3*	Person	462***	Heart Disease
4	2	3*	Person	462***	Heart Disease
5		3*	Person	462***	Cancer
6		3*	Person	462***	Cancer
7	3	≥ 40	Person	462***	Heart Disease
8		≥ 40	Person	462***	Heart Disease
8		≥ 40	Person	462***	Heart Disease

1) K-ANONYMITY APPROACH

The k-anonymity approach is an extensively applied and recognized privacy technique [10]. The concept of k-anonymity for limiting disclosure of information was proposed by [53] and [54] as an attempt to protect the privacy of persons. The idea of k-anonymity is based on modifying the values of the QI attributes to make it difficult for an attacker to unravel the identity of persons in a particular data set while the released data remain as useful as possible (see Table 2) [53], [54]. The K value is used as a measure of privacy. The lower the K value, the lower the probability of de-anonymizing. Conversely, if the K value is higher, then an attacker will have more difficulty unraveling the identity of individuals (i.e., the higher the probability of de-anonymizing). However, increasing the K value will simultaneously lower the usefulness of the data [12].

Although k-anonymization-based technique provides a certain level of privacy preservation, it also has some limitations. First, the k-anonymization-based technique will have difficulty identifying the QI attributes selected in the external tables and determining the extent by which information can be disclosed to others [46]. Recent studies [49], [55] have shown that approximately 87% of the populace can be distinctly recognized using the seemingly innocuous QI attributes. In previous studies [56] [1], mobility data set has been collected for 1.5 million people and a basic anonymization operation has been applied (eliminating apparent ID attributes). Nonetheless, these studies were able to identify

a person with 95% precision using only four spatiotemporal points. The drawback of simple anonymization was additionally confirmed by a recent study [1], [57], which analyzed a data set of 90-day financial dealings of over 1 million persons. The aforementioned study has demonstrated that four spatiotemporal points effectively re-identified approximately 90% of the persons.

Second, Table 2 gives an example of three anonymous versions of sick individuals' database relating to Table 1. The k-anonymity approach attempts to work on the attributes of QI, which involves identifying the age, gender, and ZIP code of a person, with no investment on the sensitive attributes [1]. Hence, the k-anonymity-based method is subjected to indirect attacks that enable the possibility of precisely deducing the features of an individual, thereby leading to the disclosure of identity. Examples of such an attack are homogeneity attack (i.e., absence of variety in sensitive attributes within anonymized group; see the equivalence class 3 in Table 2) and background knowledge attack, which is based on the following aspects: an opponent has sufficient background knowledge from the relationship between sensitive and QI attributes to conduct probabilistic attacks [1], [58] or when the QI attributes are connected with other public database, thereby possibly aiding an adversary to disclose the identities and other sensitive attributes of individuals [1], [27], [46], [59]. In addition, information loss with the use of anonymization-based techniques is inevitable when attempting to attain a high level of privacy [60]. However, anonymization technique possibly affects the use of data, thereby resulting in the production of imprecise or even impractical extraction of knowledge by data mining. Thus, initiating a balance between privacy and utility is essential in big data applications.

2) L-DIVERSITY APPROACH

L-diversity was designed by Machanavajjhala *et al.* [52] (2007) to protect the identities of individuals from disclosure [10]. This approach is considered an extension of the k-anonymity approach. The primary aim of L-diversity is to preserve privacy by increasing the diversity of sensitive values. This technique involves treating the values of a specific attribute in a similar manner, regardless of its distribution in the data, thereby resulting in the sufficient representation of sensitive attributes within each equivalence class in an anonymized group of data set, which prevents probabilistic inference attacks [61].

However, the major drawback of the L-diversity approach lies in the distribution of values of such sensitive attributes because different values have varying degrees of sensitivity. In an equivalence class, one value may emerge considerably more often than other values in an anonymized group (see the equivalence class 1 in Table 2). This recurrence of a value poses a serious privacy risk, thereby enabling an opponent to deduce the possibility of another entity in the equivalence class having the same value. This attack is referred to as skewness attack [50]. The production of viable l-diverse representations is difficult because the attribute values may be skewed. In addition, this approach is inadequate to prevent the disclosure of attribute to similarity attack (in an equivalence class, the values of the sensitive attribute are different while they are semantically similar). An opponent can easily have access to the sensitive attribute because the global distribution information of this attribute is markedly available to opponents, thereby resulting in divulging the identities of individuals. L-diversity guarantees the diversity of sensitive values in every group but does not consider their semantical nearness. This drawback motivated the development of the T-closeness approach [50], [61].

TABLE 3. A summary of data anonymization approaches.

Anonymization Approach	Advantage	Disadvantages
K-Anonymity Approach	Data remains truthful	Prone to background knowledge attack and Homogeneity attack
L-diversity Approach	Data remains truthful	Prone to similarity and skewness attack
T-Closeness Approach	Data remains truthful	Excessive information loss

3) T-CLOSENESS APPROACH

T-closeness was presented by Li *et al.* [50] as an extension of the l-diversity group-based anonymization, which is commonly used to protect privacy in data sets. In this approach, sensitive attribute distribution in any equivalence class should be similar to the distribution of the attribute in an overall table; for example, the distance between the two distributions should not exceed the threshold t [50], [58].

Overall, anonymization techniques are simple and attempt to protect the privacy of individuals. Nonetheless, they have an intrinsic drawback. Thus, they cannot continually and effectively protect the records' critical values against attacks. Furthermore, optimal anonymization has been demonstrated to be an NP-Hard problem [62]. Moreover, high dimensionality renders this technique ineffective because the identities of the primary record holders can be unmasked by merging the data with either the public or background information [62], [63].

Taking into the consideration that the form of privacy varies according to the data used and the way it is used, and there is no single technique that is entirely perfect. However, the limitations in one technique could be partially or adequately addressed by some other technique. Table 3 presents a summary of data anonymization techniques, by presenting the advantages and limitations of each approach [25].

a: VALUE-BASED APPROACHES

Value-based distortion approaches are a form of PPDM and they consist of two major types. The first type is the

fixed-data perturbation (i.e., uniform perturbation) category and the second type is the probability distribution category (see Figure 4) [25], [65].

b: UNIFORM PERTURBATION CATEGORY

To ensure that individual values are hidden during data collection, data providers can separately alter the value of each data item or attribute before sending to the collectors using one of the following two approaches: (1) by adding fixed data perturbation or substituting an attribute value with a new one (e.g., location from California to Washington and change age from 30 to 40) and (2) by generalizing data values or aggregating on the basis of the related domain hierarchy [48] (e.g., generalize age from 33 to range 31–35).

Both approaches are effective in protecting sensitive data from unauthorized use and performing the anonymization process. Accordingly, both approaches address different attributes independently and separately. That is, they adjust only the chosen values that minimalize the utility loss [48], while certain attributes that have no mining value are disclosed to the data warehouse servers [30]. In addition, perturbation-based methods are suitable for random data owing to the addition of fixed data. Meanwhile, aggregation-based methods can be applied only to data with domain hierarchy that has been disclosed to the data warehouse server. They can also ensure k-anonymity [53]. However, data collectors can retrieve the original data distribution from the perturbed data [30].

Probability Distribution Category: The randomization technique is one of the most commonly used methods for modifying the data in the probability distribution category [12], [25], [58]. This technique involves the addition of noise on the bases of some recognized probability distribution to mask the attribute values of records [66]. In general, the Gaussian distribution is used to generate noise values. This method endeavors to preserve data privacy for individuals by reconstructing the distributions. This method involves introducing a specific random perturbation for the original data values using a randomized process. Thus, an individual perturbed data value can be relatively dissimilar from its original data value. Accordingly, the real values are reserved in private and they cannot be deduced by the opponents by relating private attributes to a specific person [12]. The key point in this approach is that the owner of the data set publishes the resulting tuples from x_i+r instead of x_i, where $(x_{1,}x_{2,}\cdots,x_{n,})$ are the original data values of a column (one-dimensional distribution) are randomly drawn from a random variable x, and r is a random value drawn from a certain distribution.

Despite the simplicity and intuitive nature of the randomization technique, it also has certain drawbacks, the most common of which is privacy breach [12]. Several studies have [45], [64], [67], [68] experimentally demonstrated how unproductive the randomization technique may be at preserving privacy. In addition, a private data recovery algorithm can reasonably retrieve the original data from the perturbed data. When a relationship or strong correlation exists among the different attributes, this strong correlation is typically maintained after randomization. The introduced noise to each attribute is also independent. Thus, a private algorithm for data recovery can exploit the spectral structure of the perturbed data by using a filtering method. Consequently, the original data can be accurately recovered from the randomized data.

c: DIMENSION-BASED APPROACHES

In this approach, the data sets have several correlated attributes (multiple dimensions) rather than single column distribution to obtain exceptional results for data mining process in privacy preservation. Previous value-based approaches rearrange the data distributions to execute mining for privacy preservation, which involves analyzing each dimension separately, thereby overlooking the correlations among various attributes (dimensions) [64]. Previous value-based approaches require all data providers to assume the same level of privacy disclosure. That is, t value-based approaches typically require a significant amount of noise to hide the sensitive information, thereby overpowering the initial features enclosed in the actual data [69].

The clever approach to resolving this problem is to employ methods that deal with multiple-dimensions to obtain valuable results. For this reason, several dimension-based approaches applied in data collection have been newly proposed. Among these approaches, random rotation transformation and random projection are the most widely applied approaches [58]. They overcome the problem of a large noise included in the true data by transforming the original data to another space, thereby offering a considerable level of privacy guarantee, although certain features and relations in the original space are preserved [69].

Random Rotation Transformation Approach: Random rotation perturbation was developed by Chen and Liu [64] for privacy preserving data classification of data with multiple dimensions (attributes). This approach is a key module in geometric perturbation, although the quality of data mining remains unaffected, that will impact on the Quality of Service (QoS) of the processed data in the cloud [70]. The fundamental concept of this approach is changing (rotation) the data in a specific manner to protect private information in public data sets. The major drawback of the random rotation perturbation is that the domain-specific properties of data, such as the inner product or Euclidean distance, are not preserved. This result confirms that the majority of the available modeling techniques are perturbation invariant while bringing distance inference attacks [71], [72].

In this approach, data owners substitute the initial data X_{iXd} with $f(x) = X_{iXd} * R_{dXd}$, where X_{iXd} denotes the matrix that represents i objects and d attributes and R_{dXd} signifies a random rotation orthonormal matrix [73]. Privacy is guaranteed as long as the data values of the published matrix X_{iXd} relatively differ from the data values of the original matrix $f(x)$.

TABLE 4. Comparison of the dimension-based approaches.

Dimension-based Approach	Information Loss	Privacy Preservation Level
Random Projection Approach	Lower	High
Random Rotation Transformation Approach	Lower	High

Random Projection Approach: Random projection can generate the perturbation data f (x) by using random matrices. This approach is good for maintaining data utility rather than incorporating some random values into the definite data: $f(x) = X * P$, where X denotes the data sets matrix, which has dimensions $m * n$ (where n is rows and m is columns), and P is $r * m$ random matrix where $r \leq m$ [65].

The random projection technique is mainly based on the Johnson-Lindenstrauss Lemma [74], which requires the transformation of a set (N) of the original data points from its initial high dimensional space to a lower-dimensional subspace (randomly selected). This technique offers high-level privacy to the original data. Consequently, the extracted value and dimensionality of the original data set are unattainable, even if the random matrix is revealed. However, identifying the approximation of the original data is feasible [58].

A study for random projection matrices have been utilized as tools for the preservation of the privacy of mined data sets [75]. This study provides several attributes of the random projection matrices that are applicable to several data mining tasks, such as estimation of Euclidean distance, inner product, correlation, and linear classification.

Table 4 summarizes the comparison of the three types of data perturbation methods. It has been pointed out that the measurement of privacy preservation level and information loss are usually carried out through methods of data perturbation [71]. The two important concepts that should be mentioned here are the privacy preservation and information loss. The privacy preservation level refers to the degree of difficulty of estimating original data from perturbed data [64]. On the other hand, the information loss is a situation in which a significant portion of information of the original data set is lost after perturbation.

V. PRIMARY TASKS OF DATA MINING

Data mining tasks involve pattern detection and extraction from large data sets (big data) and the subsequent transformation into a readable format for future use. Thus, these tasks can be generally categorized into two common types of functions, namely, descriptive and predictive tasks. These tasks are based on the specific tasks to be achieved. The objective in a descriptive data-mining task is to modify the observed patterns in a given data set into a format that can be read by humans to generate new nontrivial knowledge on the basis of the considered data set. By contrast, the aim of predictive data mining is to depend on some fields or variables in a given data set to predict the unknown/prospective data [76], [77].

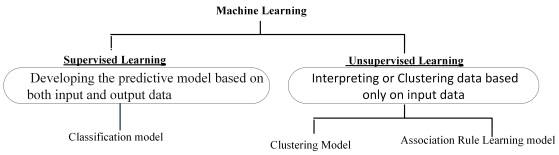

FIGURE 6. Machine-learning techniques include unsupervised and supervised learning.

Moreover, machine learning and pattern discovery extraction are perceived as ''two facets of the same field'' [12], [78]. The tasks are formed using machine learning techniques, which can be broadly categorized into supervised and unsupervised learning (see Figure 6) [79]. Supervised learning is a type of system where the input data and corresponding anticipated output data are provided. Accordingly, machine learning differentiates data, thereby developing the model. The classification model is a common type of learning task. Although the function of this model is to predict distinct classes, such as blood groups, the regression model predicts numerical values. In the unsupervised learning scheme, the learning system attempts to find relations in the data or associations among variables from unlabeled data. The association rule learning and clustering models are two common types of learning tasks under the unsupervised learning scheme. We will describe these tasks in considerable detail in the following subsections [12].

A. ASSOCIATION RULE MINING

Association rule mining algorithm is a widely applied data mining technique that is designed to determine the relationships among items or discern repeated patterns in the same transaction. This algorithm was initially presented as a market basket analysis tool [80] in the context of frequent item sets and association rule mining. Association rule mining has recently garnered significant interest in database communities [81] Furthermore, it has evolved into a useful tool for conducting unsupervised exploratory data analysis over a broad array of research and commercial areas [12], [79].

Thereafter, these associations are presented as if/then rules that facilitate the discovery of frequent patterns, where the pattern is a set of items that occurs frequently in a data set. The binary format of market basket data is presented in Table 5.

Let $A = A_1, A_2, \ldots A_n$, be a set of attributes called items and each row corresponds to a transaction (T), where T is a database of transactions $T = \{t_1, t_2, \ldots t_N$, where each T_i holds a subset of items referred to as item set (a set of zero or more items) chosen from A, such that $T \subseteq A$. The transaction width is the number of items present in T. An item can be analyzed as a binary variable, the value of which is 1 if the item is present in T and 0 otherwise. If an item set is K and contains items $X = \{x_1, x_2, \ldots x_k\}$, then it is called $K -$ itemset.

TABLE 5. Binary 0/1 representation of market basket data.

TID	A_1	A_2	A_3	A_4	A_5	A_n
t_1	1	1	0	0	0	0
t_2	1	1	1	1	1	0
t_3	1	1	1	1	0	1
t_4	1	1	1	1	0	0
t_N	1	1	0	0	0	1

To identify the valid rules from a given set of transactions T, the role of the mining association rules is to establish all the possible association rules, the confidence and support of which are more than the user-defined minimum support denoted as minconf and minsuprespectively.

The support of a rule is the probability (percentage), which denotes the quantity of transactions contained in a specific item set. Let $X \subset \mathcal{A}$ and $Y \subset \mathcal{A}$, where X and Y represent the disjoint item sets of $\mathcal{A}$. Thereafter, the rule support is $X \cup Y$. In the data set shown in Table 5, the support for $\{A_1, A_2, A_3$ is equal to three because only three transactions currently hold all three items. The confidence of a rule is denoted as percentage, which indicates the number of times X and Y are present in the entire transactions divided by the number of times X is found. The formulated expressions of these metrics are outlined as follows:

$$Support,\, s\,(X \rightarrow Y) = \begin{cases} \dfrac{|(X \cup Y)|}{|N|} & as \text{ percentage} \\ |(X \cup Y)| & as \text{ probability} \end{cases} or \quad (1)$$

$$\text{Confidence},\, c\,(X \rightarrow Y) = \left(\frac{|X \cup Y|}{|X|} \right) \quad (2)$$

Two steps are needed to mine the association rules on the bases of the support and confidence metrics [82], [83]:

- Finding the entire frequent itemsets in the database that exceed or equal to the minsup threshold and
- Generating the strong association rules from these frequent itemset.

Therefore, data-mining techniques are widely used in the field of data analysis and classification as its great benefit in discovering knowledge and hiding patterns out of big data sets. There are many related studies recently proposed and published using some meta-heuristics algorithms in solving different issues with big data analysis such as the ones in [84]–[86]. On the other hand, there are recently several of meta-heuristics algorithms that proposed that gave a very promising performance in optimizing the existing data-mining methods to achieve high accuracy. Hence, we would like to shed the light on the great possibility of exploring the usefulness of such algorithms in advancing the performance of PPDM. For example, but not limited to, one of the recently populated optimization algorithms as Grey Wolf Optimizer (GWO) [87].

B. CLUSTERING

Clustering is a type of unsupervised learning that entails the identification of valuable cluster of objects (observations) that are similar to one another in groups (clusters). The use of a few clusters to represent data unavoidably results in the loss of some minute details. However, simplification can be achieved. That is, the separation of an entire data set into groups of data has more similar characteristics than objects from different clusters [81]. The reason is that groups where an object belongs are not similar and may not be prespecified. The groups may also be revealed to have unknown relations in the data. Therefore, clustering is occasionally called the automatic classification of objects [81], [12].

The number of clustering depends on individuals' perception of unlabeled training data set, which is used to represent these groups. The number of clusters is the main constraint in clustering [88] because the notion of ''cluster'' is not precisely defined [89], [90]. Consequently, various algorithms have been recently proposed and every algorithm follows a set of rules to find cohesive groups in large data sets [89]. Users understand the problem and the equivalent data types will be the most effective measure in selecting the suitable method [88].

1) CATEGORIZATION OF ALGORITHMS

Han *et al.* [6] presented one of the numerous categorizations, whereas others (e.g., Agyapong *et al.* [77] and Jain et al.) have suggested similar categories for clustering methods. The most important categorization of clustering methods has been previously reported [6] and is based on the following properties.

- Partitioning methods: Data partitioning algorithms involve dividing data into several subsets (clusters). Partitioning methods are based on mathematical models (probabilistic and fuzzy membership models) such as Expectation–Maximization, k-mean, c-mean, and FUZZY [88]. Each cluster attempts to improve a certain clustering criterion locally (on a subset of objects), such as the computation of the values of the similarity or distance or globally (defined over all of the objects). Hence, the majority of these clusters could be considered greedy-like algorithms and computationally complex in applications involving large data sets [6], [91], [92].
- Hierarchical methods: These method are the most commonly used [93] and they work by grouping data objects that have many attributes into a cluster hierarchy (i.e., a tree of clusters) [92], [94]. The two types of hierarchical methods are (1) divisive and (2) agglomerative. The divisive method starts with a single cluster and subsequently separates into smaller clusters (called splitting or top down). The agglomerative method starts with each object in an individual cluster and attempts to combine with similar clusters thereafter to form larger clusters (referred to as merging or bottom up). Typically, each stage of hierarchical clustering entails the integration

or splitting of a pair of clusters in accordance with a specific criterion. Several criteria, such as single link (nearest neighbor) and explicit formulations of induction principle, have been proposed to optimize some criteria [88], [95]. Although this technique may appear simple and fast, locating objects that are similar among a large collection of objects requires comparing each object with every other object, thereby making the process cumbersome for large data sets [94]. In general, agglomerative clustering has time complexity ($\vartheta\left(n^2 \log(n)\right)$), whereas divisive clustering demands comprehensive search ($\vartheta\ (2^n)$). For exceptional cases, optimal efficient agglomerative methods, such as CLINK algorithm for complete-linkage clustering [83] and SLINK for single-linkage [84], has time complexity $\vartheta(n^2)$.

– Density-based methods: These methods are based on the concept of density, which is generally defined as the number of objects in some space [96]. They are effective for a combination of several distributions (clusters), which have a several connected objects affiliated with each cluster and are drawn from a specific probability distribution [97]. A cluster continues growing, provided that the density (i.e., number of objects) in the region surpasses some parameters. Hence, the output of density-based methods, such as the DBSCAN algorithm (one of the most widely applied clustering algorithms), is represented as a graph G. This graph G represents the relation $\acute{y}_i$, which is in the same cluster as $\acute{y}_j$ with the predicate. This framework differs from the partitional frameworks that depends on the iterative relocation of the specified points to a specific number of clusters [88], [91], [92], [96].
– Grid-based methods: These clustering methods are used to enhance the efficiency of clustering. In this class of methods, the object space is fragmented into several cells rather than the data that comprise a grid structure to form the base for all clustering operations. These methods are beneficial owing to the fast processing time, which is not often a function of volume of the data set but merely on the number of cells contained in each quantized space dimension. Among the examples of the grid-based technique are WaveCluster (wavelet transform-based objects clustering), STING (analysis and storage of statistical information in the grid cells), and CLIQUE (grid and density-based data clustering in high dimensional space).

C. DATA PERTURBATION TECHNIQUES

The perturbation of data involves the addition of noise to data to distort it prior to data mining [48]. This model is widely used for PPDM. Data perturbation has been commended as a more effective approach to data protection than the re-identification of individuals owing to the high probability that attacks could occur, thereby linking public data sets to QI or with another public database. The basic idea of perturbation is to create a copy individually by adding noise to distort the data before performing actual mining. The addition of noise to data makes the unconfined values inaccurate, thereby protecting the sensitive attributes of the disclosure. However, the two metrics (i.e., levels of privacy guarantee and data utility), which are often used to evaluate techniques of perturbation, are faced by poor trade-offs in several obtainable perturbation techniques [58], [64]. In general, several proposed data perturbation techniques could be classified into two: value-based approaches (i.e., value distortion approach), which focus on single-dimensional perturbation; and dimension-based approaches, which focuses on multi-dimensional data perturbation. We will describe this approaches in detail in a later section [12], [48], [58].

D. CLASSIFICATION

Classification is typically employed as a supervised learning method to construct a model for predicting the categorical labels (the class label attributes) by evaluating the relationship among the attributes and the classes of the objects in the training set [12]. Hence, different algorithms use diverse approaches to search for these relationships. Classification is a data mining function that determines the class of each object in a predefined set of classes or groups on the basis of the attributes [98], [99]. The simplest type of classification, which is known as binary classification, involves the prediction for any of two target classes. By contrast, multi-class targets have more than two target classes. Similarly, complex functions are a result of a classification process with more than two classes [99]. The data classification process includes two steps [12]. The first step is the training phase. In this phase, the model or classifier is constructed to describe a predetermined set of data classes and used for classification in the second step [99].

The frequently applied methods for the classification of data mining tasks are grouped into rule-based methods, decision tree induction methods, neural networks, memory-based learning, support vector machines, and Bayesian network.

By considering the aforementioned properties, many algorithms have been developed for diverse applications [6], [12].

VI. CONCLUSION

Privacy-preserving data mining (PPDM) is a subfield of the data mining research area. In this paper, the existing PPDM techniques and are intensively reviewed and classified based upon their methods that used data modification approaches, which then represented the main contribution of this study that will help researchers in this field having comprehensive understanding of PPDM. Furthermore, this study compared and analyzed the advantages and drawbacks of the various PPDM techniques. This research also elaborated on the existing challenges and unresolved issues in PPDM. The findings of this study show that PPDM continue to have potential challenges and open issues that would open the door for further research by scholars in the area of data privacy and protection. Thus, further technical studies are required to

propose an effective solution to address the challenges of PPDM raised in this research.

REFERENCES

[1] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.

[2] P. Rotella. *Is Data the New Oil?* Accessed: Oct. 25, 2018. [Online]. Available: https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#2042b4cd7db3

[3] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[4] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2018.

[5] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.

[6] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2011.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1999.

[8] C. Clifton, "Encyclopædia britannica: Definition of data mining," *Retrieved*, vol. 9, no. 12, p. 2010, 2010.

[9] J. C.-W. Lin, P. Fournier-Viger, L. Wu, W. Gan, Y. Djenouri, and J. Zhang, "PPSF: An open–source privacy–preserving and security mining framework," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2018, pp. 1459–1463.

[10] A. Pawar, S. Ahirrao, and P. P. Churi, "Anonymization techniques for protecting privacy: A survey," in *Proc. IEEE Punecon*, Nov. 2018, pp. 1–6.

[11] L. Cranor, T. Rabin, V. Shmatikov, S. Vadhan, and D. Weitzner, "Towards a privacy research roadmap for the computing community," Apr. 2016, *arXiv:1604.03160*. [Online]. Available: https://arxiv.org/abs/1604.03160

[12] R. Mendes and J. P. Vilela, "Privacy–preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017.

[13] G. Jagannathan and R. N. Wright, "Privacy-preserving imputation of missing data," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 40–56, Apr. 2008.

[14] O. Maimon and A. Browarnik, "NHECD—Nano health and environmental commented database," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2009, pp. 1221–1241.

[15] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[16] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *CSURACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, Jun. 2010.

[17] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explor. Newslett.*, vol. 4, no. 2, pp. 28–34, Dec. 2002.

[18] S. Banisar and S. G. Davies, "Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments," *John Marshall J. Comput. Inf. Law*, vol. 18, no. 1, p. 1, 1999.

[19] W. Gan, J. Chun-Wei, H.-C. Chao, S.-L. Wang, and P. S. Yu, "Privacy preserving utility mining: A survey," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2617–2626.

[20] Y. Xu, "New models and techniques on privacy-preserving information sharing," Ph.D. dissertation, School Comput. Sci., Simon Fraser Univ., Burnaby, BC, Canada, 2008. [Online]. Available: https://summit.sfu.ca/item/9259

[21] P. Bhaladhare and D. Jinwala, "A sensitive attribute based clustering method for k-anonymization," in *Proc. Int. Conf. Adv. Comput., Netw. Secur.*, 2011, pp. 163–170.

[22] N. Bairagi, "A survey on privacy preserving data mining," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 2015–2018, 2017. [Online]. Available: http://www.ijarcs.info

[23] K. P. Rao and A. Chaudhary, "Survey on privacy preserving data mining," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 3342–3343, 2014.

[24] R. Natarajan, R. Sugumar, M. Mahendran, and K. Anbazhagan, "A survey on privacy preserving data mining," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 1, no. 1, pp. 103–112, 2012.

[25] A. Shah and R. Gulati, "Privacy preserving data mining: Techniques, classification and implications—A survey," *Int. J. Comput. Appl.*, vol. 137, no. 12, pp. 40–46, Nov. 2016.

[26] Y. Ding and K. Klein, "Model–driven application–level encryption for the privacy of E-health data," in *Proc. Int. Conf. Availability, Rel. Secur.*, Feb. 2010, pp. 341–346.

[27] D. Nashik, "Novel approaches for privacy preserving data mining in k-anonymity model," *J. Inf. Sci. Eng.*, vol. 78, no. 1, pp. 63–78, 2016.

[28] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan, "Leveraging content sensitiveness and user trustworthiness to recommend fine–grained privacy settings for social image sharing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1317–1332, May 2018.

[29] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "IPrivacy: Image privacy protection by identifying sensitive objects via deep multi–task learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 5, pp. 1005–1016, May 2017.

[30] N. Zhang and W. Zhao, "Privacy-preserving data mining systems," *Computer*, vol. 40, no. 4, pp. 52–58, Apr. 2007.

[31] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy protection based on differential privacy strategy for big data in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3628–3636, Aug. 2018.

[32] E. Bertino, D. Lin, and W. Jiang, "A survey of quantification of privacy preserving data mining algorithms," in *Privacy-Preserving Data Mining: Models and Algorithms*. Boston, MA, USA: Springer, 2008, pp. 183–205.

[33] R. Conway and D. Strip, "Selective partial access to a database," in *Proc. of Annu. Conf.-ACM*, 1976, pp. 85–89.

[34] Z. Yang, S. Zhong, and R. N. Wright, "Privacy–preserving classification of customer data without loss of accuracy," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2005, pp. 92–102.

[35] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.

[36] S. Agrawal and J. Haritsa, "A framework for high–accuracy privacy–preserving mining," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 193–204.

[37] C. Clifton, "Privacy-preserving data integration and sharing," in *Proc. 9th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery*, 2004, pp. 19–26.

[38] N. Zhang, "Privacy-preserving data mining," Ph.D. dissertation, Texas A&M Univ., College Station, TX, USA, 2006. [Online]. Available: http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/ETD-TAMU-1080/ZHANG-DISSERTATION.pdf

[39] A. M. Ali, "Randomly encryption using genetic algorithm," *Int. J. Appl. Innov. Eng. Manage.*, vol. 2, no. 8, pp. 242–246, 2013.

[40] C. Perrin. (2018). *The CIA Triad*. Accessed: Jun. 29, 2018. [Online]. Available: https://www.techrepublic.com/blog/it-security/the-cia-triad/

[41] W. Stallings and M. P. Tahiliani, *Cryptography and Network Security: Principles and Practice*, vol. 6. London, U.K.: Pearson, 2014.

[42] H. Vaghashia and A. Ganatra, "A survey: Privacy preservation techniques in data mining," *Int. J. Comput. Appl.*, vol. 119, no. 4, pp. 20–26, Jun. 2015.

[43] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy–preserving data mining," *J. Priv. Confidentiality*, vol. 1, no. 1, p. 5, Feb. 2018.

[44] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 4, pp. 469–472, Jul. 1985.

[45] Z. Luo and C. Wen, "A chaos-based multiplicative perturbation scheme for privacy preserving data mining," in *Proc. 5th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Jun. 2014, pp. 941–944.

[46] M. Keyvanpour and S. S. Moradi, "Classification and evaluation the privacy preserving data mining techniques by using a data modification-based framework," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 2, pp. 862–870, 2011.

[47] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2001, pp. 247–255.

[48] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Rec.*, vol. 33, no. 1, pp. 50–57, 2004.

[49] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus*, vol. 4, no. 1, Dec. 2015, Art. no. 694.

[50] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[51] A. Sharma and N. Badal, "Literature survey of privacy preserving data publishing (PPDP) techniques," *Int. J. Eng. Comput. Sci.*, vol. 6, no. 5, pp. 1–12, Apr. 2017.
[52] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, 2006, p. 24.
[53] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 217–228..
[54] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information," in *Proc. PODS*, vol. 98, 1998, p. 188.
[55] L. Sweeney, "*k*-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Puzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
[56] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, Dec. 2013.
[57] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015.
[58] X. Li, Z. Yan, and P. Zhang, "A review on privacy-preserving data mining," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.*, Sep. 2014, pp. 769–774.
[59] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Unc. Fuzz. Knowl. Based Syst.*, vol. 10, no. 5, pp. 571–588, Oct. 2002.
[60] S. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.
[61] R. P. Priyadarsini, S. Sivakumari, and P. Amudha, "Enhanced ℓ—Diversity algorithm for privacy preserving data mining," in *Proc. Conf., Annu. Conv. Comput. Soc. India (CSI)*, Coimbatore, India. Singapore: Springer, 2016.
[62] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2004, pp. 223–228.
[63] A. Hasan, Q. Jiang, H. Chen, and S. Wang, "A new approach to privacy–preserving multiple independent data publishing," *Appl. Sci.*, vol. 8, no. 5, p. 783, May 2018.
[64] K. Chen and L. Liu, "A random rotation perturbation approach to privacy preserving data classification," Wright State Univ., Dayton, OH, USA, Tech. Rep. GIT-CC-05-12, 2005. [Online]. Available: https://corescholar.libraries.wright.edu/knoesis/916/
[65] O. H. Reddy and P. Singh, "Preserving privacy in data mining by data perturbation technique," *Int. J.*, vol. 4, no. 1, pp. 1–7, Jan. 2015.
[66] A. Charu and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*. Boston, MA, USA: ASPVU, 2008.
[67] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Apr. 2004, pp. 99–106.
[68] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 2005, pp. 37–48.
[69] W. Wang and Q. Zhang, *Location Privacy Preservation in Cognitive Radio Networks*. Cham, Switzerland: Springer, 2014.
[70] Y. Yin, W. Zhang, Y. Xu, H. Zhang, Z. Mai, and L. Yu, "QoS prediction for mobile edge service recommendation with auto–encoder," *IEEE Access*, vol. 7, pp. 62312–62324, 2019.
[71] C. C. Aggarwal and P. S. Yu, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-Preserving Data Mining*. Springer, 2008, pp. 11–52.
[72] A. Patel, K. Dodiya, and S. Pate, "A survey on geometric data perturbation in multiplicative data perturbation," *Int. J.*, vol. 1, no. 5, pp. 603–607, 2013.
[73] L. A. Sadun, *Applied Linear Algebra: The Decoupling Principle*. Providence, RI, USA: American Mathematical Society, 2007.
[74] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemp. Math.*, vol. 26, no. 1984, pp. 189–206, 1984.
[75] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 92–106, Jan. 2006.
[76] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Incognito: Efficient full-domain K-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2005, pp. 49–60.
[77] K. B. Agyapong, J. B. Hayfron-Acquah, and M. Asante, "An overview of data mining models (descriptive and predictive)," *Int. J. Softw. Hardw. Res. Eng.*, vol. 4, no. 5, pp. 53–60, 2016.
[78] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2011.
[79] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017.
[80] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.
[81] M.-S. Chen, J. Han, and P. Yu, "Data mining: An overview from a database perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, Dec. 1996.
[82] C. S. K. Selvi and A. Tamilarasi, "An automated association rule mining technique with cumulative support thresholds," *Int. J. Open Probl. Compt. Math*, vol. 2, no. 3, pp. 428–438, 2009.
[83] S. Darrab and B. Ergenç, "Frequent pattern mining under multiple support thresholds," *Methods*, vol. 10, no. 11, pp. 1–10, 2016.
[84] A. S. Sadiq, M. A. Tahir, A. A. Ahmed, and A. Alghushami, "Normal parameter reduction algorithm in soft set based on hybrid binary particle swarm and biogeography optimizer," *Neural Comput. Appl.*, 2019, doi: 10.1007/s00521-019-04423-2.
[85] A. S. Sadiq, H. Faris, A. M. Al-Zoubi, S. Mirjalili, and K. Z. Ghafoor, "Fraud Detection Model Based on Multi-Verse Features Extraction Approach for Smart City Applications," in *Smart Cities Cybersecurity and Privacy*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 241–251.
[86] M. A. T. Mohammed, A. S. Sadiq, R. A. Arshah, F. Ernawan, and S. Mirjalili, "Soft set decision/forecasting system based on hybrid parameter reduction algorithm," *J. Telecommun. Electron. Comput. Eng.*, vol. 9, nos. 2–7, pp. 143–148, 2017.
[87] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.
[88] V. Estivill-Castro, "Why so many clustering algorithms: A position paper," *SIGKDD Explor. Newslett.*, vol. 4, no. 1, pp. 65–75, Jun. 2002.
[89] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. New York, NY, USA: Springer, 2005.
[90] V. Estivill-Castro and J. Yang, "Fast and robust general purpose clustering algorithms," *Data Mining Knowl. Discovery*, vol. 8, no. 2, pp. 127–150, Mar. 2004.
[91] P. Andritsos, "Data clustering techniques," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2002. [Online]. Available: http://www.cs.toronto.edu/~periklis/pubs/depth.pdf
[92] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*, vol. 25. Berlin, Germany: Springer, 2006, pp. 25–71, doi: 10.1007/3-540-28349-8_2.
[93] R. Tamilselvi, B. Sivasakthi, and R. Kavitha, "A comparison of various clustering methods and algorithms in data mining," *Int. J. Multidisciplinary Res. Develop.*, vol. 2, pp. 32–36, May 2015.
[94] S. Firdaus and A. Uddin, "A survey on clustering algorithms and complexity analysis," *Int. J. Comput. Sci. Issues*, vol. 12, no. 2, pp. 62–85, 2015.
[95] K. Srivastava, R. Shah, D. Valia, and H. Swaminarayan, "Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment," *Int. J. Comput. Theory Eng.*, vol. 5, no. 3, p. 520, 2013.
[96] T. W. Liao, "Clustering of time series data-a survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.
[97] J. D. Banfield and A. E. Raftery, "Model–based Gaussian and non–Gaussian clustering," *Biometrics*, vol. 49, no. 3, p. 803, Sep. 1993.
[98] S. Sumathi and S. N. Sivanandam, "Data mining tasks, techniques, and applications," in *Introduction to Data Mining and Its Applications*. New York, NY, USA: Springer, 2006, pp. 195–216.
[99] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–7.

MOHAMMED BINJUBEIR received the master's degree in computer science from Universiti Sains Malaysia (USM). He is currently pursuing the Ph.D. degree with Universiti Malaysia Pahang (UMP).

ABDULGHANI ALI AHMED (Senior Member, IEEE) received the B.Sc. degree in computer science, in 2002, and the M.Sc. and Ph.D. degrees in cybersecurity and forensic investigation, in 2006, and 2014, respectively. He is currently a Senior Lecturer of cybersecurity with the Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang (UMP). He is also teaching security courses, such as information security, network security, ethical hacking, computer forensic & investigation, malware analysis, and cybercrime. He has been a long experience in working with a higher education as a Lecturer and a Senior Lecturer, since 2004. Prior to get his Ph.D., he was a Lecturer in the fields of cybersecurity, computer networks, information system and management, object oriented programming, and network administration. He serves as a Main Supervisor for many postgraduate (master's and Ph.D.) students who are studding and conducting researches in the area of cybersecurity, big data privacy, cloud computing security, and cybercrime investigation. He has managed to obtain several local and international grants to fund cybersecurity studies and research. He has published several researches and scientific articles in well-known international journal and conferences. His current research interests include cybersecurity, network security, cloud computing security, big data privacy, ethical hacking, malware analysis, incident response, digital forensic, and cybercrime investigation. He is a member of the International Association of Engineers (IAENG). He has excellent achievements in the track of invention and innovation. In terms of inventions, his record of Intellectual Properties achievements shows two patents and several copyrights. In terms of innovation, he has awarded many Gold, Silver, and Bronze medals from local, national, and international exhibitions. He is the Founder and the Leader of the Safecyber Systems Corporation for security solutions development. The current focus of Safecyber Corp is developing several systems, including safecyber, safeware, and safeapp. He acts as a Volunteer Reviewer in well-reputed journals, such as the IEEE Systems Journal, the *Journal of Network and Computer Applications* (JNCA), IEEE Access, *Wireless Networks*, *Neural Computing and Applications*, IETE *Technical Review*, KIIS, and JDCTA.

MOHD ARFIAN BIN ISMAIL received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Universiti Teknologi Malaysia (UTM), in 2008, 2011, and 2016, respectively. He is a Senior Lecturer with the Faculty of Computer Systems and Software Engineering, University Malaysia Pahang. His current research interests include machine learning methods and optimization method.

ALI SAFAA SADIQ (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science in 2004, 2011, and 2014 respectively. He is currently a Faculty Member of the Faculty of Science and Engineering, School of Mathematics and Computer Science, University of Wolverhampton, U.K. He is also an Adjunct Staff with Monash University, Malaysia. He has served as a Lecturer with the School of Information Technology, Monash University. Previously, he has also served as a Senior Lecturer with the Department of Computer Systems and Networking Department, Faculty of Computer Systems and Software Engineering, University Malaysia Pahang, Malaysia. His current research interests include wireless communications, network security, and AI applications in networking. He has been awarded the Pro-Chancellor Academic Award as the best student in his batch for both masters's and Ph.D. degrees. He has also been awarded the UTM International Doctoral Fellowship (IDF). He has published several scientific/research articles in well-known international journals and conferences. He was involved in conducting five research grants projects, whereby three of them are in network and security and the others in analyzing and forecasting floods in Malaysia. He has supervised three Ph.D. students and three master's degree as well as some other undergraduate final year projects.

MUHAMMAD KHURRAM KHAN (Senior Member, IEEE) is currently a Professor of cybersecurity with the Center of Excellence in Information Assurance (CoEIA), King Saud University, Saudi Arabia. He is one of the Founding Member of the CoEIA. He has served as the Manager of Research and Development from 2009 to 2012. He, along with his team, developed and successfully managed the Cybersecurity Research Program with CoEIA, which turned the center as one of the best centers of excellence in the region. He is the Founder and the CEO of the Global Foundation for Cyber Studies and Research, an independent, non-profit, and non-partisan cybersecurity think-tank in Washington, DC, USA, which explores and addresses global cyberspace challenges from the intersecting dimensions of policy and technology. He has also played the role of the guest editor of several international journals of the IEEE, Springer, Wiley. He was a Guest Editor of *Information Sciences* (Elsevier). Moreover, he is one of the organizing chairs of more than five dozen international conferences and a member of technical committees of more than ten dozen international conferences. In addition, he is an active Reviewer of many international journals as well as research foundations of Switzerland, Italy, Saudi Arabia, and Czech Republic. He was a recipient of the King Saud University Award for Scientific Excellence (Research Productivity), in May 2015. He was also a recipient of the King Saud University Award for Scientific Excellence (Inventions, Innovations, and Technology Licensing), in May 2016. He has secured the Outstanding Leadership Award from the IEEE International Conference on Networks and Systems Security, in 2009, Australia. Besides, he has received a certificate of appreciation for outstanding contributions in *Biometrics and Information Security Research* at the AIT international Conference, in June 2010, Japan. He has been awarded a Gold Medal of the Best Invention and Innovation Award from the Tenth Malaysian Technology Expo, in 2011, Malaysia. In addition, he was awarded the Best Paper Award from the *Journal of Network and Computer Applications* (Elsevier), in December 2015. He has published more than 350 research articles in the journals and conferences of international repute. In addition, he is an inventor of 10US/PCT patents. He has played a leading role in developing the BS Cybersecurity Degree Program and Higher Diploma in Cybersecurity with King Saud University. His research interests include cybersecurity, digital authentication, the IoT security, cyber policy, and technological innovation management. He is a distinguished lecturer of the IEEE. He is a fellow of the IET, U.K., the BCS, U.K., the FTRA, South Korea, a Senior Member of the IACSIT, Singapore, a member of the IEEE Consumer Electronics Society, the IEEE Communications Society the IEEE Technical Committee on Security and Privacy, the IEEE IoT Community, the IEEE Smart Cities Community, and the IEEE Cybersecurity Community. He is also the Vice Chair of the IEEE Communications Society Saudi Chapter. Moreover, in April 2013, his invention has got a Bronze Medal at 41st International Exhibition of Inventions at Geneva, Switzerland. He is the Editor-in-Chief of a well-reputed international journal including *Telecommunication Systems* published by Springer for over 26 years with its recent impact factor of 1.707 (JCR 2019). Furthermore, he is on the Editorial Board of several international journals including, the IEEE Communications Surveys and Tutorials, the *IEEE Communications Magazine*, the IEEE Internet of Things Journal, the IEEE Transactions on Consumer Electronics, the *Journal of Network and Computer Applications* (Elsevier), IEEE Access, the IEEE *Consumer Electronics Magazine*, *PLOS one*, *Electronic Commerce Research*, *IET Wireless Sensor Systems*, the *Journal of Information Hiding and Multimedia Signal Processing*, and the *International Journal of Biometrics*. He has edited seven books/proceedings published by Springer–Verlag and IEEE. He has secured several national and international competitive research grants in the domain of cybersecurity.