IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 July 2022, accepted 11 August 2022, date of publication 17 August 2022, date of current version 24 August 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3199433*

RESEARCH ARTICLE

# Privacy Preserving Attribute-Focused Anonymization Scheme for Healthcare Data Publishing

**J. ANDREW ONESIMU[1], KARTHIKEYAN J[2], JENNIFER EUNICE[3], MARC POMPLUN[4], AND HIEN DANG[4,5], (Member, IEEE)**

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India
[2]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
[3]Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu 641114, India
[4]Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125, USA
[5]Faculty of Computer Science and Engineering, Thuyloi University, Hanoi 100000, Vietnam

Corresponding author: Hien Dang (hiendt@tlu.edu.vn)

**ABSTRACT** Advancements in Industry 4.0 brought tremendous improvements in the healthcare sector, such as better quality of treatment, enhanced communication, remote monitoring, and reduced cost. Sharing healthcare data with healthcare providers is crucial for harnessing the benefits of such improvements. In general, healthcare data holds sensitive information about individuals. Hence, sharing such data is challenging because of various security and privacy issues. According to privacy regulations and ethical requirements, it is essential to preserve the privacy of patients before sharing data for medical research. State-of-the-art literature on privacy preserving studies either uses cryptographic approaches to protect the privacy or uses anonymizing techniques regardless of the type of attributes, this results in poor protection and data utility. In this paper, we propose an attribute-focused privacy preserving data publishing scheme. The proposed scheme is two-fold, comprising a fixed-interval approach to protect numerical attributes and an improved $l$-diverse slicing approach to protect the categorical and sensitive attributes. In the fixed-interval approach, the original values of the healthcare data are replaced with an equivalent computed value. The improved $l$-diverse slicing approach partitions the data both horizontally and vertically to avoid privacy leaks. Extensive experiments with real-world datasets are conducted to evaluate the performance of the proposed scheme. The classification models built on anonymized dataset yields approximately 13% better accuracy than benchmarked algorithms. Experimental analyses show that the average information loss which is measured by normalized certainty penalty (NCP) is reduced by 12% compared to similar approaches. The attribute focused scheme not only provides data utility but also prevents the data from membership disclosures, attribute disclosures, and identity disclosures.

**INDEX TERMS** Anonymization, data privacy, data publishing, healthcare data, privacy-preserving.

## I. INTRODUCTION

In the current era of Industry 4.0, enormous amounts of data are generated through various digital activities. Information privacy[1] is at risk because the data are collected and indulged in various analyses. Data owners may not have proper control over their own data. Privacy is a conceptual integrity that connects the protection of personal data with information flow in specific contexts. The importance of contextual integrity is highlighted under specific contextual considerations such as networks, groups, and society [2]. Networked privacy is defined as an individual's loss of control over their personal data disclosure in social networks. Marginalized and socioeconomic groups of individuals lack knowledge about the Internet, and it is essential to examine risk factors. It is also important to align with the government

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

**TABLE 1.** Digital age privacy policies around the globe.

| Country | Privacy Policy | Year | Description |
|---|---|---|---|
| United States | Children's Online Privacy Protection Act - COPPA | 1998 | • Websites should get parents' consent while collecting information about a child under 13 years.<br>• Websites which collects personal information of the children under 13 years of age should abide by COPPA |
| | Health Insurance Portability and Accountability Act - HIPPA | 1996 | • Patient consent is required to collect and disclose the personal information<br>• HIPAA protects individual's health privacy information. |
| Canada | Personal Information Protection and Electronic Documents Act - PIPEDA | 1995 | • Collection of personal information is limited and only for reasonable purpose<br>• Consent should be garnered for data collection |
| European Union | General Data Protection Regulation - GDPR | 2016 | • Pseudonymization is required to store the personal data<br>• Privacy policies should be more clear, concise and transparency in data collection, storage, processing and disclosure |
| Australia | Privacy Act {Information Privacy Act 2000 (Victoria) Information Act 2002 (Northern Territory) Personal Information Protection Act 2004 (Tasmania)} | 1988 | • Upon data collection Australians have the right to know the why the information is acquired and who will use it.<br>• Australian has right to access information |
| India | Information Technology Act | 2000 | • Every organization should publish privacy policy on their website<br>• Privacy policy should describe what you collect, purpose of collection, third party disclosure, and security practices to protect data.<br>• Individual consent is required to collect information such as financial and personal. |

rules of privacy, as privacy rules differ for different societies. Table 1 shows a few digital-age privacy policies worldwide.

Recently, there has been explosive growth in healthcare big data generation with the development of Industry 4.0, which facilitates the Internet of Things (IoT) [3], mobile technologies, wearable devices, and artificial intelligence [4]. Electronic healthcare services offer several advantages, such as remote monitoring, telemedicine, e-health applications, and improved communication. Figure 1 shows some of the major advantages of e-healthcare. Research on healthcare big data is exceptionally captious in improving the accuracy of diagnoses and developing ingenious medicines [5], [6], [7]. Generally, healthcare organizations share health records with research organizations for medical discoveries [8], [9], [10]. Organizations collect Electronic Health Records (EHRs) through the healthcare data process [11], which refers to the medical treatment of a patient in healthcare information technology (IT) (e.g., Healthcare IoT Services). The EHR is a digitized version of a patient's paper chart. EHR are an important part of healthcare IT. It contains the patient's medical history, date of treatment, diagnoses, and other personal information [12]. Table 2 shows an example of an EHR collected by a healthcare organization. EHR is not merely a record management method that supports clinicians in various aspects of patient care through technological capabilities: (i) clinical documentation and health information display, (ii) results management, (iii) computerized provider order entry and management (CPOE), (iv) clinical decision support (CDS), (v) electronic communication and connectivity, (vi) patient support, (vii) administrative processes, and (viii) reporting and population health management [13]. Other benefits include improved patient care, diagnostics, easy access, and reduced costs.

EHR contain sensitive information about patients, and releasing it leads to numerous privacy issues. Although personal identifiers are removed from health records, they are still vulnerable to background knowledge attacks, where the attacker combines the released data with the available knowledge to identify an individual. Healthcare data sharing is vital for cutting-edge research in the medical field. However, data holders (e.g., healthcare organizations) can share their data only if they trust third-party cloud providers. Because privacy protection is an ethical and regulatory requirement, healthcare organizations are hesitant to share patient data despite several advantages. This demands a privacy-preserving data publishing (PPDP) [14] scheme that protects the patient's privacy and benefits the medical research community. Many PPDP approaches have recently been proposed, based on anonymization and cryptographic techniques. Cryptographic techniques include authentication, password management, access controls, and biometric schemes. Though they provide better protection they are less preferable for PPDP as they are computationally expensive for searching and manipulating

the data from huge datasets [15]. Techniques like differential privacy [16] and homomorphic encryptions are also being used for privacy preserving approaches however, they are not suitable for PPDP considering the data utility issues. On the other hand, anonymization techniques including generalization, suppression, randomization, and pseudonymization are specifically used for privacy preserving studies. During the process of anonymization, the records in the dataset are transformed into less specific and indistinguishable without changing the actual meaning of the data. Hence, anonymization techniques are preferable over cryptographic techniques to the protect privacy and provide better data utility in PPDP studies [17]. Sweeney first proposed an anonymization model, known as $k$-anonymity [18] for sharing personal data. This model makes personal records indistinguishable from at least $k$-1 records. The pitfalls of $k$-anonymity models are addressed in the $l$-diversity model, which introduces diversity among the sensitive attributes in the records. The $t$-closeness anonymity model brings closeness among *diverse* records to further protect sensitive attributes. However, they are still vulnerable to privacy attacks such as identity, membership, and attribute disclosure attacks [19]. Privacy attacks on healthcare data also lead to societal and psychological issues.


**FIGURE 1.** E-healthcare services.

### A. CONTRIBUTIONS

Numerous PPDP schemes based on $k$-anonymity have been proposed to protect the privacy of EHR, including *($\alpha$, $k$)-anonymity* [20], *($p$,$\alpha$)*-sensitive $k$-anonymity [21], and *($p$+, $\alpha$)* sensitive $k$-anonymity [22]. Nevertheless, an adversary can still ascertain personal information using sophisticated techniques [23], [24]. Especially for healthcare data publishing, different attributes contribute in their way to medical research. Generalizing or suppressing the data without considering their type would lead to information loss and privacy leaks. Majeed [17] proposed an attribute centric approach that focuses on protecting the numerical and categorical attributes of the healthcare data. The proposed approach comprises of an interval based approach to protect the numerical attributes and pseudonymization based approach for categorical attributes. However, the sensitive attributes are prone to disclosure. Hence, there is a need for an attribute-focused approach that protect different attributes of the healthcare data with specific anonymization approaches without privacy leak and better data utility. In this paper, we propose a novel attribute-focused anonymization scheme for healthcare data publishing. The major objective of our scheme is to provide maximum utility while preserving the privacy of the healthcare data. The healthcare records from each data holder are organized into an equivalent class with at least $k$ records. The proposed scheme is twofold. First, we perform attribute classification to identify the numerical attributes of EHR that need to be protected, and then, we delineate the fixed-interval anonymization approach that is efficient in generalizing the numerical attributes. Second, we propose an improved $l$-diverse slicing approach that efficiently generalizes the categorical attributes of the healthcare data. Consequently, the proposed PPDP scheme neutralizes identity disclosure, attribute disclosure, and membership disclosure attacks.

**TABLE 2.** Example personal electronic health record.

| Sex | Age | Zip code | Diagnosis |
|---|---|---|---|
| Female | 20 | 620706 | Dyspepsia |
| Female | 22 | 620709 | Flu |
| Female | 25 | 620712 | Flu |
| Male | 31 | 641008 | Gastritis |
| Male | 32 | 641014 | Cancer |
| Male | 33 | 641016 | Pneumonia |
| Female | 37 | 651406 | Cancer |
| Female | 38 | 651502 | Insomnia |
| Male | 39 | 651806 | Flu |

The summary of the contributions of our research work is as follows:

(1) We propose a novel attribute-focused anonymization scheme to protect healthcare data privacy during data publication.
(2) A novel fixed-interval-based anonymization approach is proposed to protect the numerical attributes of the EHR from disclosure.
(3) An improved $l$-diverse slicing approach is proposed to protect categorical and sensitive attributes from disclosure.
(4) Implemented and evaluated the proposed scheme with real-world datasets and compared with state-of-the-art anonymization approaches.

### B. ORGANIZATION OF THE PAPER

The remainder of this paper is organized as follows. Section II presents the state-of-the-art privacy-preserving approaches

for healthcare data. The preliminaries, data model, and system architecture are provided in Section III. Section IV describes the proposed attribute-focused anonymization scheme. Section V describes the experimental analysis and discusses the efficiency of the proposed scheme. Finally, Section VI concludes the study with a future scope.

## II. RELATED STUDY

Privacy-preserving approaches have gained significant attention in recent decades because of advancements in information technology that threaten the privacy of individuals. Generally, e-health services manage a large amount of sensitive personal information for various purposes. privacy-preserving approaches in the field of e-health services should provide a balance between data privacy and utility. Numerous privacy-preserving approaches have been proposed to protect individuals' privacy. Some of the popular privacy preserving techniques are $k$-anonymity [18], $l$-diversity [25], $t$-closeness [26], amplified randomization [27], *(a, d)* diversity [28], $p$-sensitive [29], $\beta$-likeness [30], and so on. However, they remain vulnerable to privacy attacks. In this section, we discuss the state-of-the-art literature on PPDP and highlight their shortcomings.

Outsourcing healthcare data to the cloud involves numerous security and privacy issues. Healthcare-sensitive data are vulnerable to attacks on public clouds. Wang *et al.* [31] proposed a framework for outsourcing high-dimensional healthcare data to a cloud. This framework first divides the data into sensitive and non-sensitive data; then, the sensitive data are stored in a private cloud, and the non-sensitive data are stored in a public cloud. Sensitive data are protected by injecting differential privacy noise into the data [32]. The other attributes were protected using partition-based anonymization techniques. However, the noise injected would impact the utility of the data, and partition-based anonymization are vulnerable to disclosure attacks. Attaullah *et al.* [33] proposed a fuzzy logic-based algorithm that is privacy-aware to protect healthcare data privacy from disclosures. In the proposed approach, the attributes are classified into quasi and sensitive attributes using fuzzy logic. Fuzzy classification is then used to anonymize the attributes. However, the fuzzy rules are generated in the static that determines the information loss and query accuracy. Kim and Chung [34] proposed a *k-anonymity-based* protocol to address identity-disclosure attacks. The author divides identity disclosure into internal and external disclosure. Internal identity disclosure occurs when the data collector discerns the identity of the data holder. External identity disclosure occurs when an identity is leaked through the network headers. To protect the data from such attacks, the *K_I* and *K_E* anonymity models were proposed. These models ensure that at least $k$ records share the same quasi-identifiers, and on the data collector side, each generalized group contains at least $k$ data holders that share similar quasi-identifiers.

However, $k$-anonymity ensures that at least $k$ records are similar in the dataset table. They are vulnerable to identity and attribute disclosure attacks. $l$-diversity and $t$-closeness bring diversity and closeness to the data; however, some of the sensitive attributes are left unanonymized, which leads to privacy leaks. Sei *et al.* [35] defined a new set of attributes, called sensitive quasi-identifiers. The proposed model comprises $l$-diversity, $t$-closeness, and differential-privacy techniques. First, the sensitive quasi-attributes are identified from the table and randomized. Then, the proposed anonymization algorithm is used to anonymize the table by applying frequency $l$-diversity and $t$-closeness to the data. This approach attempts to reduce information loss and improve data privacy; however, the selection of a sensitive QID remains crucial for data privacy. Conventional privacy-preserving techniques, such as $k$-anonymity, $l$-diversity, and $t$-closeness, cannot be applied directly to large datasets because of their unscalability. To address this issue, Mehta and Rao [36] proposed a scalable $l$-diversity approach. In this approach, the quasi-attributes are first $k$-anonymized and then $l$-diversity is applied to the dataset. This is scalable to increasing the size of the dataset. A comparative analysis showed that the scalable $l$-diversity approach provided minimum information loss and reduced time complexity. However, using this approach for publishing data streams has potential risks of privacy leaks. In [37] restricted sensitive attribute-based anonymization algorithm was proposed to preserve the privacy of data-stream publishing. This approach introduces two privacy constraints: sensitivity and semantic diversity. These constraints restrict the privacy breach of sensitive published data. Sensitive diversity and semantic diversity generalize the sensitive and semantic nature of the data; thus, they can be identified individually. The performance of the system was evaluated based on processing time and information loss.

Bucketization is a data anonymization technique used to publish sensitive attributes. In bucketization, the data records are grouped into smaller groups and the sensitive attributes are distributed among the groups, which weakens the relationship between quasi-attributes and sensitive attributes, thus preserving the privacy of sensitive attributes from disclosure. In [38] the bucket-setting problem was addressed through a flexible bucketization scheme. It allows every sensitive attribute to define its privacy setting and have a variable bucket size. The flexible nature of this scheme provides an option for adjusting the privacy and information loss of a dataset. Although generalization protects privacy, it also has some serious drawbacks. Generalization results in the loss of a considerable amount of information from the original data. Owing to the dimensionality of the data, the interval range can be extremely wide. This renders the generalized table useless for further analytics. An improved bucketization-based anonymization technique is proposed in [39]. The attributes are partitioned and clustered. The sensitive attributes are distributed among the clusters. Thus there is a possibility of sensitive attribute leakage. An atomizing algorithm was proposed by Xiao *et al.* to address the generalization issues [33]. The algorithm first partitions the database based on the $l$-diversity. It then generates a quasi-identifier table

and sensitive identifier table. Anatomy removes the direct correlation between quasi-identifiers and sensitive identifiers. This preserved the privacy of the database. The information loss is reduced through small diverse partitions. *K*-member clustering has been proposed [40] to reduce information loss by clustering similar records in the database. There are no restrictions on the number of clusters; however, every cluster should have $k$ records. This preserves privacy because individual records are indistinguishable from the clusters. However, $k$-anonymization is NP-hard. Bayardo and Agrawal [41] proposed an optimal k-anonymity algorithm to prune useless values and reorder the dataset tuples to reduce the complexity. Traditional anonymization and generalization algorithms fail miserably if the dataset contains multiple records for an individual. 1: M-anonymization [42] was proposed by adopting $k$-anonymity and $l$-diversity techniques. The model partitions attributes into quasi-sensitive attributes. The sensitive attributes are $k$-anonymized, and the equivalence class of quasi-attributes satisfies $l$-diversity. The model uses NCP metrics to calculate information loss. Wang and Li [43] proposed correlation-aware anonymization of high-dimensional data (CAHD) to anonymize transaction data through the greedy heuristic grouping technique. CAHD utilizes the $l$-diversity technique to group data and uses a band matrix to reconstruct the data. A fixed-interval approach for anonymizing electronic health records was proposed by Majeed [17]. The interval was calculated using the bin value for the quasi-attributes. The attribute set was divided into $n$ bins with a limited range of attribute values. The numerical attributes were anonymized by replacing the original value with the mean value. Categorical data were protected using an ID-based anonymization. In ID-based anonymization, categorical data are transformed into numerical IDs. The privacy-preserved data are then evaluated through classifiers, such as a support vector machine and random forest. Abbasi and Mohammadi [44] have proposed a $k$-means++ algorithm in which the healthcare attributes are first clustered and then anonymized using the $k$-anonymity technique. Though the results claim reduced information loss and execution time, individual attributes are not given appropriate importance. In [45] used pseudonymization techniques to preserve the privacy of the patient records. The pseudonyms are protected using cryptographic techniques thus it is not suitable for data publication. Another work by Arca and Hewett [46] focused to protect the privacy of the smart health data using attribute hierarchy is discussed. They have identified the potential attributes for privacy leakage through entropy measures. However, the hierarchy-based anonymity decreases the data utility. Chong and Malip [47] have addressed attribute disclosure and linkability issues. The attributes are classified as numerical and non-numerical. Permutation-based methods are used to anonymize the attributes. There may be information loss as the categorical data are also considered numerical data. The summary of privacy requirements fulfilled and possible attacks of various anonymization techniques are presented in Table 3. It is observed from the table that every anonymization technique has its own merits and demerits. Table 4 compares some of the popular state-of-the-art privacy preserving literature in the recent past. We compared the privacy model adopted by the literature, the anonymization technique, the performance metrics used, and its drawbacks. The major drawback we observed in the literature [48], [49], [50], [51], [52] is the information loss. This is due to the anonymization technique selection. It is noticed that these models utilized generalization and suppression techniques which transforms the data as per definition 1 and 2. For e.g., if the value of age attribute is '27' then it can be generalized as ''20-30'' or even ''20-40'' based on the privacy parameter selection. This results in huge information loss. Based on the analysis we understood that there is a need for a privacy preserving approach that protects the privacy of the patients with minimal information loss and execution time. Also we noticed lack of attribute focused privacy preserving approaches, which is essential for privacy preserving healthcare data publishing.

## III. PRELIMINARIES

This section defines the preliminaries used in this study, including the data model, definitions of various privacy notations, and a detailed description of the generalization technique.

### A. DATA MODEL

EHR generally follow a relational data model, where the attributes can be categorized into personal identifiers, quasi-identifiers, and sensitive information.

I. *Personal identifiers (ID)* are unique attributes of patients that distinguish the individual (e.g., name, SSN). To protect patients' privacy, they must be removed or replaced with dummy id's. Personal identifiers were not necessary for the data analysis.
II. *Quasi Identifiers (QI)* are attributes that identify an individual when combined with other published attributes (e.g., gender, age, and zip code). QI's are publicly available and useful for data analysis.
III. *Sensitive information (SI)*: Sensitive attributes of a patient that must be protected from the adversary (e.g., diagnosis and medication). In Table 2, the diagnosis is considered to be sensitive information.

The privacy-preserving data publishing (PPDP) problem can be modeled as follows: there are $n$ records on EHR $(1 \ldots n)$, where each record $r_i$ represents a personal healthcare record of an individual patient. Each record is a combination of personal identifiers (ID), quasi-identifiers (QI), and sensitive information (SI). There is a possibility of more than one sensitive piece of information; however, in this case, we consider a single sensitive attribute problem. IDs are generally removed from the EHR because they are not required for data analysis; they also lead to privacy breaches because they uniquely identify an individual. The QIs of EHR can be represented as

**TABLE 3.** Summary of privacy-preserving techniques-requirement fulfillments and vulnerabilities.

| Anonymization Technique | Description | Privacy Requirements Satisfied | Possible Privacy Attacks/Demerits |
|---|---|---|---|
| Pseudonymization | It is the process of replacing the personal attributes with pseudonyms | Anonymity, Consistency | Identity disclosure, attribute disclosure, membership disclosure, |
| Generalization | It is the process of replacing the specific values with generalized values | Non-disclosure agreement, sensitive attribute protection | Attribute disclosure, Membership disclosure, Heavy computational cost |
| Suppression | It is the process of replacing the original attributes with some special characters "*" | Patient consent, Insider attack, Spoofing | Poor data utility, Heavy computational cost |
| Bucketization | It is the process of dividing the original attributes into buckets to ensure protection | Non-disclosure of sensitive information, Diversity | Attribute disclosure, Membership disclosure |
| Randomization | It is the process adding noise to the original data to mask its actual behavior | Anonymity, decreasing the attacker success rate | Membership disclosure, Poor data utility |
| Slicing | It is the process of partitioning the dataset both horizontally and vertically to protect the actual attributes | Membership Non-disclosure, Anonymity, Audit | Identity disclosure, Membership disclosure |
| Cryptographic Approaches | It is generally used for secure the data | Trust, Authorization, Audit, non-repudiation, confidentiality | Heavy computational cost, Poor data utility |

**TABLE 4.** State-of-the-art anonymization algorithms.

| Literature | Privacy Model | Anonymization Technique | Environment Focus | Performance Metrics | Drawbacks |
|---|---|---|---|---|---|
| Mondrian [48] | *k*-anonymity | Multidimensional generalization | General | Discernibility metric, classification accuracy | High information loss and execution time |
| IACK [49] | *k*-anonymity | Generalization | General | Classification accuracy | High information loss with attribute disclosures |
| Incognito [50] | *k*-anonymity, *l*-diversity | Generalization and suppression | Classification model for healthcare data | Information Loss/Distortion and Classification Accuracy | High information loss and moderate execution time |
| Datafly [51] | *k*-anonymity, *l*-diversity | Generalization and slicing | Medical data | Distinctive attribute metric | Moderate information loss and not suitable for classification models |
| K,$K_m$ - Anonymity [52] | *k*-anonymity | Generalization | Transactional data | Normalized Certainty Penalty (NCP) | Moderate information loss and high execution time |
| KMDAE-DAC [53] | Clustering *k*-anonymity, Elliptic Curve Digital Signature | Generalization and Cryptographic techniques | Cloud environment | Classification Accuracy | Cluster size determines the information loss |
| Ac-FI Anonymity [17] | *k*-anonymity | Fixed interval generalization | Healthcare data - EHR | Classification Accuracy | Sensitive attributes are not protected |
| Clustering based K-Anonymity [10] | Clustering *k*-anonymity | Taxonomy based generalization | IoT healthcare data | NCP and discernibility metric | Information loss varies according to the level of taxonomy |
| HB-Anonymity [54] | Heap bucket anonymization | Anatomization and bucketization | Healthcare data - EHR | NCP and execution time | Determination of bucket size is challenging |

$r^q_{num,cat} = (r^1_{num}, r^2_{num}, r^3_{num}, r^1_{cat}, r^2_{cat}, r^3_{cat} \ldots, r^{m1}_{num}, r^{m2}_{cat})$ where $q \in QI$, $r^q_{num,cat}$ represent the types of QIs. Although an EHR consists of multiple datatypes of data in this work, we consider $r_{num}$ numerical and $r_{cat}$ categorical attributes only. Sensitive information (SI) can be represented as $r^s$ where $s \in SI$. The objective of the proposed PPDP scheme is to produce an anonymized dataset that is suitable for building classification models. Hence, in a PPDP scheme, the healthcare organization should publish $\left(r^*_{\varepsilon(1)}, r^*_{\varepsilon(2)}, r^*_{\varepsilon(3)}, \ldots, r^*_{\varepsilon(n)}\right)$ where $\varepsilon$ is a random permutation of QI with SI, and $r^*$ represents the anonymized version of the original records.

*Definition 1 (Generalization):* Consider the quasi-attribute sets $\{q_1, q_2, q_3, \ldots, q_m\}, q \in QI$ where $m$ denotes the number of QIs in each record. Generalization is the process of transforming the actual QI, which is more significant for QIs that represent the same QI with less significant values. For example, the sample EHR are shown in Table 2, and the generalized version of the table is presented in Table 5. The QI attribute age is generalized within a specific range of values rather than the original explicit values.

*Definition 2 (Suppression):* Suppression is the process of replacing original attribute values with the "∗" character. This technique partially masks the attributes, making them less significant. For example, in Table 4, the last three digits of the *QI* attribute *zip code* were suppressed.

**TABLE 5.** *k*-anonymized version of EHR table.

| Sex | Age | Zipcode | Diagnosis |
|---|---|---|---|
| Female | 20-25 | 620*** | Dyspepsia |
| Female | 20-25 | 620*** | Flu |
| Female | 20-25 | 620*** | Flu |
| Male | 31-33 | 641*** | Gastritis |
| Male | 31-33 | 641*** | Cancer |
| Male | 31-33 | 641*** | Pneumonia |
| Female | 34-37 | 651*** | Breast Cancer |
| Female | 34-37 | 651*** | Insomnia |
| Male | 34-37 | 651*** | Flu |

*Definition 3 (Anonymization):* Anonymization is the process of transforming the original attributes into insignificant attributes, which makes the individual record indistinguishable. Consider a table $T$ consists of *QI* and *SI* then the process of anonymization is mapping the original *QI* attributes through a function $f$ to generate the anonymized version of the table $T^*$ that consists of $Q^*$ where $f$ denotes the generalization or suppression operation and $Q^*$ denotes the anonymized $q$ attributes.

*Definition 4 (k-Anonymization):* An EHR table is said to be $k$-anonymized only if it satisfies the $k$-anonymity property that every record in the table is indistinguishable from at least $k$-1 records. The records in the $k$-anonymized table were anonymized using generalization or suppression techniques. Table 5 shows the $k$-anonymized ($k = 3$) version of the table, where each record is indistinguishable from at least $k-1$ (two records).

*Definition 5 (Equivalence Class):* Equivalence class $E$ is a subset of the anonymized table $T^*$. Every anonymized table has several equivalence classes that share similar generalized QI attributes. For example, Table 5 consists of 3 equivalence classes, where each class consists of indistinguishable QI attributes.

*Definition 6 (Bucketization):* Bucketization is the process of dividing table $T$ into equal-sized buckets. Each bucket shared *diverse* sensitive information. Therefore, sensitive attributes are protected from the adversary, even if they can discern the individual. Table 6 presents the bucketized table, which consists of three buckets and 2-diverse sensitive attributes, where each bucket contains at least two unique sensitive attributes. Let there be $b$ buckets $B_1, B_2, B_3, \ldots, B_b$ then $\bigcup_{i=1}^{b} B_i = T, B_i \cap B_j = \emptyset$.

**TABLE 6.** Bucketized table.

| Sex | Age | Zipcode | Diagnosis |
|---|---|---|---|
| Female | 20 | 620706 | Dyspepsia |
| Female | 22 | 620709 | Flu |
| Female | 25 | 620712 | Flu |
| Male | 31 | 641008 | Gastritis |
| Male | 32 | 641014 | Cancer |
| Male | 33 | 641016 | Pneumonia |
| Female | 37 | 651406 | Breast Cancer |
| Female | 38 | 651502 | Insomnia |
| Male | 39 | 651806 | Flu |

*Definition 7 (Slicing):* Slicing is the process of partitioning table $T$ horizontally (tuple partition) and vertically (attribute partition). The slicing also satisfies the $k$-anonymity property. Table 7 shows the sliced table that consists of tuple partitions $\{\{t_1, t_2, t_3\}, \{t_4, t_5, t_6\}, \{t_7, t_8, t_9\}\}$ and attribute partitions {Sex, Age}, and {Zipcode, Diagnosis}}.

**TABLE 7.** Sliced table.

| {Sex, Age} | {Zipcode, Diagnosis} |
|---|---|
| {Female, 20} | {620712, Flu} |
| {Female, 22} | {620706, Dyspepsia} |
| {Female, 25} | {620709, Flu} |
| {Male, 31} | {641016, Pneumonia} |
| {Male, 32} | {641008, Gastritis } |
| {Male, 33} | {641014, Cancer} |
| {Female, 37} | {651806, Flu } |
| {Female, 38} | {651502, Insomnia} |
| {Male, 39} | {651406, Breast Cancer} |

*Definition 8 (Bucket Matching):* Let $c$ be the columns of a sliced table $\{C_1, C_2, C_3, \ldots, C_c\}$. Let $t$ be the tuple of table $T$, and $t[C_i]$, where $C_i$ is the value of $t$. Let $B_i$ be a bucket in the sliced table, and $B[C_i]$ where $C_i$ is the value of multiset $B$. Bucket matching occurs when $(B == t) \iff \forall 1 \leq i \leq c, t[C_i] \in B[C_i]$. For example, as per Table 6, $t_1$ = {Female, 20, 620706,Dyspepsia}. Then, $t_1$ matches bucket $\{B_1\}$.

*Definition 9 (Distribution of Tuple and Bucket - Distr(t,B)):* To protect the sensitive attribute $s$ of tuple $t$ in bucket $B$, it is necessary to calculate the distribution of the sensitive attributes in the tuple and bucket. Let $Distr(t, B)$ be the

distribution of sensitive attributes in $B$. Sensitive attribute $s$ is said to be associated with $t$ in $B$ when $t[C_c - S]$ where $s \in S$. Let $Distr(t, B)$[s] be the probability of the sensitive attribute $s$ in the distribution. The probability of $t$ in bucket$B$ is denoted as $p(t, B)$ and the probability of $t$ taking the sensitive value $s$ is denoted as $p(s|t, B)$. Then $p(t, s)$ is then calculated based on the law of total probability as follows:

$$p(t, s) = \sum_B p(t, B)\, p(s|t, B) \tag{1}$$

Analysis of $p(t, Bu)$ is given as follows

$$t = \{t[C_1], t[C_2], t[C_3], \ldots, t[C_c]\}$$

and

$$B = \{B[C_1], B[C_2], B[C_3], \ldots, B[C_c]\} \times freq_i(t, B); \quad (1 \le i \le c-1)$$

where $freq_i(t, Bu)$ is the frequency of $t[C_i]$ in $Bu[C_i]$ then

$$freq_c(t, B) = t[C_c - \{S\}] \text{ in } B[C_c - \{S\}]$$

where $C_c - \{S\}$ denotes the set of QI in the sensitive attribute column. For example, in Table 6, $freq_1(t_1, B_1) = 1/3 = 0.33$ and $freq_2(t_1, B_1) = 2/3 = 0.67$. Correspondingly, $freq_1(t_1, B_2) = 0$ and $freq_2(t_1, B_2) = 0$. Intuitively, $freq_i(t, B)$ quantifies the matching degree in columns $C_i$, $t \le C_i \le Bu$. Subsequently, the value of $freq(t, B)$ is as follows:

$$freq(t, B) = \prod_{1 \le i \le c} freq(t, B)$$

We know that $freq(t, B) = 1$ when a tuple matches a bucket and $freq(t, B) = 0$ otherwise. Hence, for the matching bucket $\sum_t freq(t, B) = 1$. When multiple buckets match for a tuple $t$, the total matching degree for the entire data set is

$$freq(t) = \sum_B freq(t, B)$$

Then the probability of tuple $t$ in bucket $B$ is as follows

$$p(t, B) = freq(t, B) \big/ freq(t) \tag{2}$$

To compute $p(s|t, B)$ consider Table 6, $Distr(t_1, B_1) = (dyspepsia : 0.5, flu : 0.5)$ $Distr(t_1, B_1)[dyspepsia] = 0.5$. Then, the probability of determining the sensitive attribute $t$ is as follows:

$$p(s|t, B) = Distr(t_1, B_1)[s] \tag{3}$$

The probability of $p(t, s)$ can be computed using Equation (1). We can prove that tuple $t$ takes a sensitive attribute $s$ to sum up to 1.

*To Prove:* $\forall t \in D, \sum_s p(t, s) = 1$

*Proof:*

$$\sum_s p(t, s) = \sum_s \sum_{Bu} p(t, B) \cdot p(s|t, B) = \sum_{Bu} p(t, B) \cdot \sum_s p(s|t, B)$$

According to equation (3) $\sum_s p(s|t, B) = 1$ then

$$\sum_s p(t, s) = \sum_{Bu} p(t, B)$$

According to equation (2) $\sum_{Bu} p(t, B) = 1$ hence,

$$\sum_s p(t, s) = 1$$

*Definition 10 (l-Diverse Slicing):* The $l$-diversity property of slicing is defined using the probability of sensitive attribute $s$ in tuple $t$ *as follows:*

$$p(t, s) \le \frac{1}{l}$$

A sliced table is said to satisfy *the* $l$-diversity property if and only if every $t$ satisfies *the* $l$-diversity for any sensitive value $s$ in the bucket.

*Definition 11 (Correlation Measure):* Computing the correlations between attributes is a crucial step in preserving privacy and data utility. Grouping up correlated attributes provides better data utility as it preserves the correlation of the attributes. The attributes in the uncorrelated groups are more vulnerable to attribute disclosure, as the attributes are less frequent. Thus, it is vital to measure the correlation between the attributes. In this study, the correlation was calculated using the mean-square contingency coefficient [55]. The mean-square contingency coefficient is a chi-squared measure used to calculate the correlation between categorical attributes. Because our dataset is a combination of numerical and categorical attributes, we used this measure. The numerical attribute correlations were measured using a fixed-length approach (which will be discussed later in this paper).

Consider two attributes $R_1$ and $R_2$ and their domains $v_{11}, v_{12}, v_{13}, \ldots, v_{1d_1}$ and $v_{21}, v_{22}, v_{23}, \ldots, v_{2d_2}$ where $d_1$ and $d_2$ are the sizes of the domains. Then, the mean-square contingency coefficient for the attributes $R_1$and $R_2$ is measured as

$$\phi^2(R_1, R_2) = \frac{1}{\min\{d_1, d_2\} - 1} \times \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(freq_{ij} - freq_i.freq_j)^2}{freq_i.freq_j} \tag{4}$$

Here, $freq_i$ and $freq_j$ are the frequencies of occurrences in the data values $v_{1i}$ and $v_{2j}$ respectively. $freq_{ij}$ is the frequency of occurrence $v_{1i}$ and $v_{2j}$ in the data. Hence, $freq_{ij}$ is the total of $freq_i$ and $freq_j$.

$$freq_i = \sum_{j=1}^{d_2} freq_{ij}$$

$$freq_j = \sum_{i=1}^{d_1} freq_{ij}$$

The value of $\phi^2(R_1, R_2)$ can be $0 \le \phi^2(R_1, R_2) \le 1$

*Definition 12 (Attribute Clustering):* Attribute clustering was used to partition the columns after computing the correlation measure for every pair of attributes. We used the bottom-up clustering method, in which each attribute is considered

as a separate point in the cluster space. The distance between the points of the clusters is defined as

$$d(R_1, R_2) = 1 - \phi^2(R_1, R_2) \quad (5)$$

The value of $d(R_1, R_2)$ can be $0 \leq d(R_1, R_2) \leq 1$

### B. SYSTEM ARCHITECTURE

Figure 2 details the architecture of privacy preserving data publishing anonymization scheme that is considered in this paper.

The following are the details of the entities involved in the system.

- Users: In an e-healthcare system, users could be patients, doctors, healthcare professionals, pharmacists, and caregivers. Users are responsible for generating the electronic health data.
- Healthcare data–Healthcare data represent the patients' EHR. In general, healthcare data are generated by the users of e-health systems. It contains patients' identifiers, treatment history, diagnosis, and other health parameters.
- The data controller is an internal member of a healthcare organization who collects, processes, and stores EHR data. The data controller is responsible for publishing EHR for medical research without privacy breach.
- Dataset anonymization–Dataset anonymization is the process of anonymizing the EHR data to protect the privacy of the individual and provide sufficient data utility.
- Data Analyst/Attacker: The data analyst is the recipient of the anonymized EHR dataset. He performed various analyses of EHR data for numerous medical purposes. A data analyst can also be an adversary who tries to acquire more details of an individual or discern the attributes of a patient.

The system architecture of the PPDP anonymization scheme consists of three major processes: (1) data collection, (2) anonymization, and (3) data publishing. Data collection is an internal process in healthcare organizations. It involves patients, doctors, pharmacists, and other medical experts and records the personal and medical details of the patients in a computerized system. There must be a clear data collection policy available, and patient consent is required to collect and store the data. The role of the data controller in the data-collection process is to collect, store, and manage data. Second, the anonymization process removes or replaces the original values of EHR with less significant values that protect the privacy of the patients in the e-health system, also providing appropriate data utility for medical researchers. Data publishing is the final step in the PPDP scheme, in which anonymized data are shared or made publicly available to various medical researchers. The data analyst acquires an anonymized dataset and applies data-analysis algorithms. In this system, the adversary can be the data analyst when attempting to discern the details of an individual. The major objective of the PPDP scheme is to protect the privacy breach from the data analyst and to provide good utility to the data.

## IV. ATTRIBUTE FOCUSED ANONYMIZATION SCHEME

This section discusses the proposed privacy-preserving data anonymization scheme for healthcare records. Figure 3 shows the overall process of the proposed anonymization scheme. **Preprocessing** of data is an important step because data collection is often loosely controlled. Incorrect and redundant values result in incorrect results and increase the complexity of the scheme. Preprocessing includes data cleaning, feature extraction, and transformation.

**User Ranking** is important when grouping similar users. The cosine similarity measure [56] was used to rank similar users using common QIs. Grouping up similar users yields better data utility and privacy protection.

*Equivalence Classes:* The resultant microdata table is then divided into equivalence classes based on the privacy parameter $k$. This ensures that each record is indistinguishable from at least $k$-1 other records.

*Equivalence Class Evaluation:* In certain cases, the equivalence class may have a single value that can leak privacy during anonymization. Therefore, *range analysis* is performed to identify such values, and a constant value can be added also outliers detected are *pruned*.

*Attribute Classification:* In this step, the attributes were classified into numerical and categorical attributes, and data anonymization was performed as per the proposed schemes.

*Data Anonymization:* The data anonymization process uses fixed-interval anonymization and slicing approaches. The proposed fixed-interval anonymization scheme protects the numerical attributes of the EHR from disclosure. It is not only effective in protecting QI attributes but also provides better utility. This was made possible by calculating the interval width, mean, or median for the anonymization of numerical attribute

The categorical attributes of EHR data are protected through an improved $l$-diverse slicing approach. Sensitive attributes are also protected through the $l$-diverse property of slicing. Our proposed approach helps the data publishers to anonymize the EHR with ease. The complexity of the anonymization process is less compared to the state-of-the-art generalization techniques. The proposed approach is attribute-centric since it applies a different anonymization approach for numerical and categorical attributes.

Consider a healthcare organization that collects the original micro-data of an EHR. The microdata collected are vital for drug discovery and the early diagnosis of medical research. Hence, for research purposes, these micro data are published by research organizations. Releasing the original microdata leads to a privacy breach; therefore, it is essential to generate an anonymized version of the microdata. To anonymize the EHR microdata, we first identified different types of attributes in the table. Numerical and categorical attributes

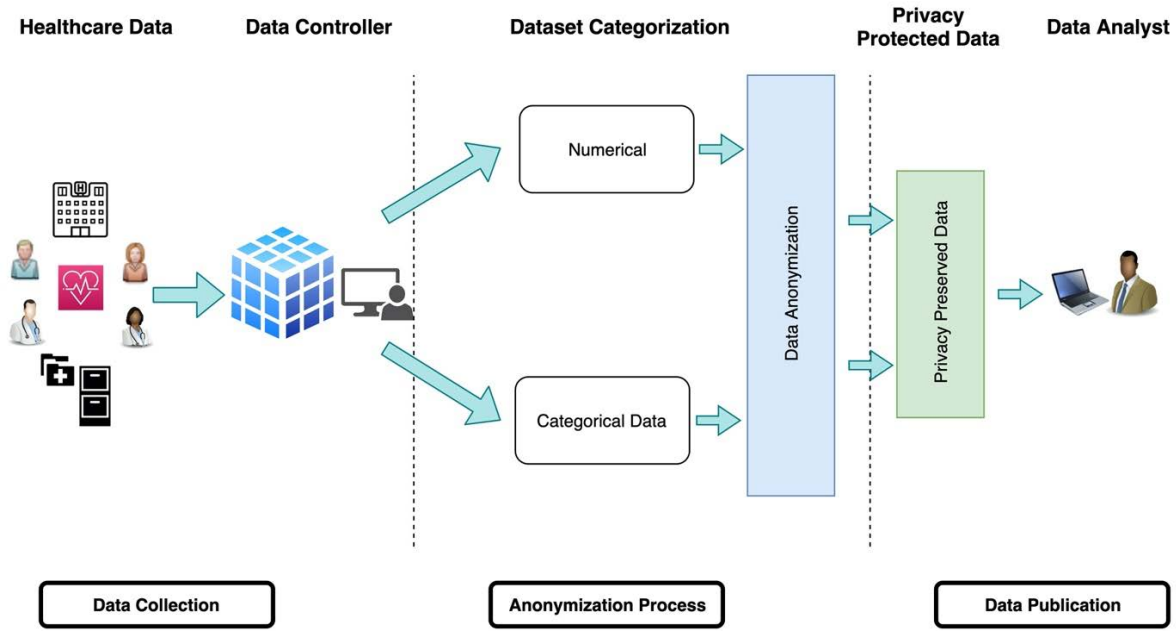


**FIGURE 2.** System architecture of PPDP anonymization scheme.

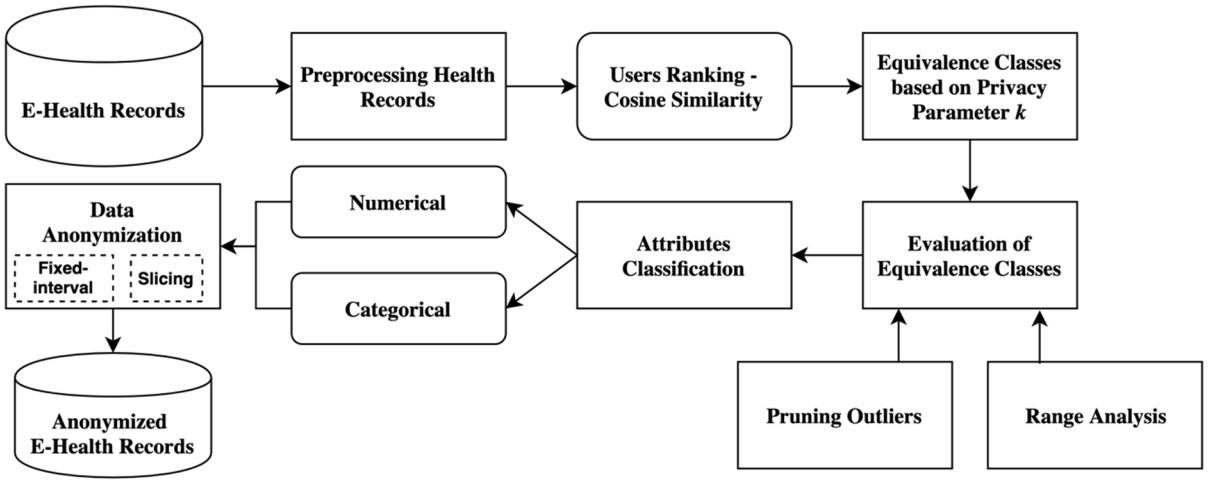


**FIGURE 3.** Attribute focused anonymization process.

were anonymized. Based on the $k$-anonymity principle, $k$ is a privacy parameter that is used to determine the level of privacy. The larger the $k$ value, the greater the protection to privacy information, but the poorer the data utility. The major objective of the PPDP scheme is to provide maximum data utility while preserving the privacy of an individual. Therefore, *k-value* selection is crucial in the *k-anonymity-based* PPDP scheme. The micro data table was then sorted based on the numerical attributes and divided into an equal-sized set of records based on the value of $k$. Subsequently, the numerical attributes were anonymized based on the fixed-interval generalization approach. Second, we identify the categorical attributes that need to be protected. Categorical attribute values were anonymized using the $l$-diverse slicing approach. The attribute partitioning method in the slicing approach disassociates sensitive and quasi-attributes to avoid sensitive attribute disclosure. The proposed approach consists of two algorithms: fixed-interval anonymization and *l-diverse* slicing.

### A. FIXED INTERVAL ANONYMIZATION

Fixed interval anonymization Algorithm 1 takes a privacy parameter k and numerical QI attributes of EHR microdata as inputs and outputs a numerical attribute anonymized version of the dataset. First, the input data are sorted to prepare and identify the interval width. To calculate the interval width, the largest and smallest values of the QI attributes were chosen (i.e., the first and last elements of the QI attributes) from the

sorted list. The interval width (IW) was then calculated to identify the equivalence class based on privacy parameter k (lines 5 and 6). After identifying the intervals, the attributes were anonymized by calculating the mean ($\mu$) for each interval. The QI attributes within the interval were then replaced with the calculated mean of the respective intervals. During the anonymization process, if all QI attributes in the interval are equal, a threshold ($\theta$) constant is used to increase the value to protect the original QI (lines 7 to 14).

For example, in Table 2, the numerical QI attributes "Age" and "Zipcode" (Record ID is not considered as it is not a QI attribute) are selected to be anonymized with fixed interval anonymization approach. To anonymize the "Age" attribute, the largest and smallest values (39 and 20, respectively) were selected. Privacy parameter $k$ was fixed at $k = 3$. The interval width (IW) was calculated as follows:

$$IW = \frac{39 - 20}{3} = 6$$

We have now divided the original table based on the calculated IW value of 6. We obtained the following groups:

i. 20 - 26
ii. 27 - 33
iii. 34 - 40

The numerical attributes in the interval were anonymized by replacing the original values with the mean.

$$\mu_{IW1} = \frac{20 + 21 + 25}{3} = 22$$

where $\boldsymbol{\mu_{IW_1}}$ is the mean value of the first interval ($\boldsymbol{IW_1}$). Similarly, calculate the mean for all intervals, and replace the original QI attributes in the interval with the calculated mean. The same steps were followed for the other numerical QI attributes. The anonymization of "Age" and "Zipcode" attribute is shown in Table 8.

**TABLE 8.** The anonymization of attribute

(A) AGE ATTRIBUTE ANONYMIZATION

| Intervals | Values |
|---|---|
| 20 - 26 | 22 |
| 27 - 33 | 32 |
| 34 - 40 | 38 |

(B) ZIP CODE ATTRIBUTE ANONYMIZATION

| Intervals | Values |
|---|---|
| 620706 - 631073 | 620709 |
| 631074 - 641441 | 641013 |
| 641442 - 651809 | 651571 |

Table 9 presents the output of the fixed-interval anonymization approach. The attributes "Age" and "Zipcode" are anonymized by replacing the original values with the interval mean values.

### B. *l*-DIVERSE SLICING

After anonymizing the numerical attributes through a fixed-interval anonymization approach, the microdata table is updated, as shown in Table 10, where the categorical attributes are not protected. The *l-diverse* slicing approach was utilized to protect categorical and sensitive attributes. It consists of three steps: cluster initialization, slicing, and *l*-diversity check.

**TABLE 9.** Fixed interval anonymization.

| Sex | Age | Zipcode | Diagnosis |
|---|---|---|---|
| Female | 22 | 620709 | Dyspepsia |
| Female | 22 | 620709 | Flu |
| Female | 22 | 620709 | Flu |
| Male | 32 | 641013 | Gastritis |
| Male | 32 | 641013 | Cancer |
| Male | 32 | 641013 | Pneumonia |
| Female | 38 | 651571 | Breast Cancer |
| Female | 38 | 651571 | Insomnia |
| Male | 38 | 651571 | Flu |

**TABLE 10.** Fixed interval + l-diverse slicing approach.

| {Sex, Age} | {Zipcode, Diagnosis} |
|---|---|
| {Female, 22} | {620709, Flu} |
| {Female, 22} | {620709, Flu} |
| {Female, 22} | {620709, Dyspepsia} |
| {Male, 32} | {641013, Cancer} |
| {Male, 32} | {641013, Gastritis} |
| {Male, 32} | {641013, Pneumonia} |
| {Female, 38} | 651571, Insomnia} |
| {Female, 38} | {651571, Breast Cancer} |
| {Male, 38} | {651571, Flu} |

**Algorithm 1** Fixed Interval Approach - Numerical Attribute

**Input:** Privacy parameter $k$, Numerical QI attributes - $QI_{num} \in Table(T)$
**Output:** Anonymized $QI_{num}$ - $T^*(QI_{num})$

1: Sort the numerical attributes in ascending order
2: Set threshold $\theta = constant_{num}$
3: **for each** $QI_{num} \in T$ **do**
4: Pick the largest $QI^L_{num}$ and smallest $QI^S_{num}$ values of $QI_{num}$
5: Calculate Interval Width $(IW) = \frac{QI^L_{num} - QI^S_{num}}{k}$
6: Divide $QI_{num}$ with respect to $IW$ to form equivalence class $E$
7: **for each** $E(QI_{num}) \in T$ **do**
8: calculate mean

$$\mu(QI_{num}) = \frac{QI^1_{num} + QI^2_{num} + \ldots + QI^n_{num}}{n}$$

9: **if** $\left(QI^1_{num} == QI^2_{num} == \ldots == QI^n_{num}\right)$ **then**
10: $QI_{num} \leftarrow \mu(QI_{num}) + \theta$
11: **else**
12: $QI_{num} = \mu(QI_{num})$
13: **end if**
14: **end for**
15: **end for**

#### 1) CLUSTER INITIALIZATION

In the *l-diverse* slicing approach, the cluster initialization step prepares the attributes for partitioning with respect to the categorical attributes. First, the correlation between attributes was calculated based on Equation (4), and the correlation

matrix was updated with the correlation values for every attribute and its domains. To partition the attributes, every tuple *t,* which is a combination of both anonymized numerical attributes and original categorical attributes, is considered as cluster point $c$. The distance between the cluster points was calculated using Equation (5). Then, the calculated distance between the attributes and privacy parameter $k$ is provided as input to the clustering algorithm to form the initial clusters. We identified partitioning around medoids (PAM) [57] which is a simple and robust $k$-medoids algorithm that creates clusters by choosing $k$ arbitrary points as medoids and subsequently swapping the medoids and non-medoids to satisfy cluster quality. Algorithm 2 describes cluster initialization.

Algorithm 2 clusters all the attributes into columns. The PAM clustering technique ensures that each cluster contains $k$ columns. This does not restrict the size of the sensitive attribute column $c_c$. The size of the sensitive attribute column can be determined using parameter $\alpha$. To protect sensitive attribute disclosure, the sensitive attribute column should have QI attributes ( if $\alpha > 1$ then $c_c > 1$).

**Algorithm 2** Cluster - Initialization

**Input:** Privacy parameter $k$, $T$
**Output:** $k$ clusters - $C = \{c_1, c_2, c_3, \ldots, c_k$

1: Tuple $t \in T^*$ $(QI_{num})$ , $T(QI_{cat})$
2: **for each** tuple $QI \in T$ **do**
3: create a cluster point $c$ for every QI
4: $C = \{c_1, c_2, c_3, \ldots, c_n$ where $n$ is the number QI
5: **end for**
6: calculate the distance between two clusters using equation (5)
7: formation of clusters using PAM algorithm [57]
8: return $k$ clustered columns

### 2) TUPLE PARTITIONING

Tuple partitioning is the primary component of the slicing approach. Tuple partitioning is a process of splitting tuples into buckets. Algorithm 3 describes the steps of tuple partitioning. Two data structures are maintained: 1) bucket queue ($QB$) and 2) sliced bucket set ($SB$). $QB$ is initialized with a single bucket that contains all the tuples belonging to table $T$, and $SB$ is empty (lines 1 and 2). Remove one bucket from $QB$, split the buckets into two (as per the splitting criteria [58]), and then check if the split satisfies $l$-diversity through Algorithm 4. This iteration is repeated until $QB$ is empty (Lines 3–11). If the split satisfies $l$-diversity, then the buckets are added to $QB$ (line 7) for further splitting; otherwise, they are added to the sliced buckets set $SB$ (line 9).

### 3) DIVERSITY CHECK

Diversity checks are pivotal for protecting sensitive attributes. It checks whether the sliced bucket satisfies the $l$-diversity. Algorithm 4 describes the steps for checking the $l$-diversity. The algorithm maintains a list of data structure *lists* $(t)$ to store the matching bucket statistics. Initially, *List(t)* is empty then for every tuple belongs to $T$ and for each bucket $B$ in sliced bucket, store the frequency *freq(v)* of column $v$ in bucket $B$. To find the matching bucket, for each tuple $t$ calculate *p(t, B)* for the tuple in the bucket $B$ and find *Distr(t, B)* for the distribution of sensitive values. Record the bucket matching and distribution statistics to $List[t]$. Finally, $p(t, s)$ is calculated for every sensitive attribute $s$ based on the *list* $(t)$. The sliced table is $l$-diverse if and only if, for all sensitive attributes $s$, $p(t, s) \leq 1/l$.

**Algorithm 3** Tuple Partitioning

**Input:** Privacy parameter $l$, $T$, Clusters $C$
**Output:** Sliced buckets - $SB$

1: $QB = t \in T$
2: $SB = \emptyset$
3: **while** $QB \neq \emptyset$ **do**
4: $QB = QB - B$
5: Split buckets $B$ into $B_1$ and $B_2$
6: **if** $Diversity(T, Q \cup \{B_1, B_2\} \cup SB, l)$ **then**
7: $QB = Q \cup \{B_1, B_2\}$
8: **else**
9: $SB = SB \cup \{B$
10: **end if**
11: **end while**
12: return $SB$

**Algorithm 4** Diversity (Check $l$-Diversity)

**Input:** $T$, $T^*$, $l$
**Output:**True/False

1: tuple $t \in T$
2: $List[t] = \emptyset$
3: **for each** $t \in T$ **do**
4: **for each** bucket $B$ in $T^*$ **do**
5: store $freq(v)$ in $B$
6: **for each** $t \in T$ **do**
7: calculate $p(t, B)$ and find $Distr(t, B)$
8: $List[t] = List[t] \cup \{p\,(t, B)\,, Distr(t, B)\}$
9: **end for**
10: **for each** $t \in T$ **do**
11: calculate $p(t, s)$ for each $s$ based on $List[t]$
12: **if** $p(t, s) \geq 1/l$ **then**
13: return False
14: **else**
15: return True
16: **end if**
17: **end for**
18: **end for**
19: **end for**

## V. EXPERIMENT & RESULTS

In this section, we demonstrate the concepts discussed through experiments. The primary objectives of the

experiment are to demonstrate the attribute focused anonymization approach and to compare the classification utility of the anonymized datasets obtained from the proposed approach to the benchmark anonymization approaches Mondrian [48], IACk [49], Attribute centric Fixed Interval (AcFI) anonymity [17], and Incognito [50], Datafly [51], and KMDAE-DAC [53]. We also assessed at how well the proposed approach performed by measuring the amount of data that was lost throughout the anonymization process. The Normalized Certainty Penalty (NCP) [59] was used as the standard information loss metric. The information loss is compared to Mondrian [48], ($k$,$k_m$)-anonymity [52], and clustering-based $k$-anonymity [10], which are all prominent anonymization methods.

### A. DATASET DETAILS

We used the Adult dataset [60] which is the de facto standard dataset for privacy-preserving studies, as in [17], [18], [36], [61], and [62]. After eliminating tuples with missing values, the dataset consisted of 45,222 with 15 attributes. The dataset contained both numerical and categorical attributes. The dataset is publicly accessible from the link [43] and it is presented in Table 11.

In our experiment, we acquired the Adult-7 and Adult-15 datasets from an original adult dataset. Adult-7 has seven attributes that includes {*Age, Workclass, Marital-Status, Race, Sex*} and "*Occupation*" as sensitive attribute. Adult-15 considers all 15 attributes with "Occupation" and "*Salary*" as the sensitive attributes for different experiments.

**TABLE 11. Dataset description.**

| Categorical | Distinct Values | Numerical | Distinct Values |
|---|---|---|---|
| Workclass | 8 | Age | 74 |
| Education | 16 | Final-Weight | NA |
| Marital-Status | 7 | Education-Num | 16 |
| Occupation | 14 | Capital-Gain | NA |
| Relationship | 6 | Capital-Loss | NA |
| Race | 5 | Hours-Per-Week | NA |
| Sex | 2 | | |
| Country | 41 | | |
| Salary | 2 | | |

### B. EXPERIMENTAL ANALYSIS

We built four classification models for our experiment: a decision tree (J48), naïve Bayes, SVM, RF, and ANN. The classification accuracy of the models was evaluated using Weka 3.0. A 10-fold cross-validation-based stratified sampling was used for all classification experiments.

Figures 4 - 6 present the classification accuracy of decision tree and Naïve Bayes algorithms with "*occupation*" as the sensitive attribute and other attributes are predictor attributes for the datasets Adult-7 and Adult-15.

Figure 4 and Figure 5 compare the classification accuracy of the anonymized data with the original data and other benchmark anonymization approaches, such as bucketization [63] and slicing [58]. Datasets Adult-7 and Adult-15 were used for the experiments. Figure 4(a) and 5 (a) represent the output decision tree (J48) model, and Figure 4(b) and 5(b) represent the output of the Naïve Bayes model. Here, the privacy parameter $k$ is set to five, and the number of columns $c$ is set to two. The $l$-diversity values vary between {5, 8, 10}. In the experiments, the proposed scheme outperformed the benchmark approaches.

Figure 6 present the effect of $c$ on the classification accuracy. The value of $c$ varies between {*2,3,5*}. Figure 6 (a) shows the variations in classification accuracy with $l = 5$ and Figure 6 (b) shows the variations in a classification accuracy with $l = 8$. The variations are very minimal in the accuracy because as the value of $c$ increases the correlated attributes are still belongs to the same column.

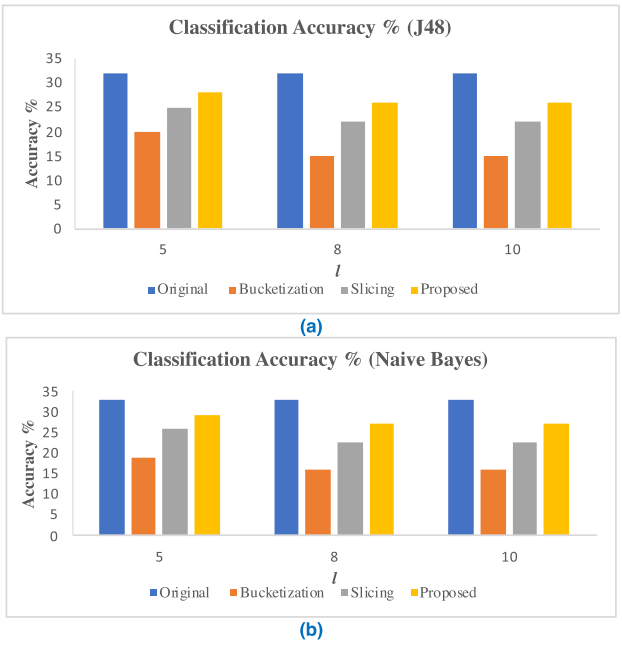


**FIGURE 4. Classification accuracy of sensitive attribute (*Occupation*) - Adult-7 dataset (a) Decision tree (J48) (b) Naïve bayes.**

Figure 7 presents the classification accuracy of SVM and RF for the Adult-15 dataset. Here, we consider "*salary*" as the sensitive attribute and all others are predictor attributes. In this experiment, we fixed the values $c$=*2* and $l = 2$ with varying $k$. Since the "salary" attribute has only 2 distinct values it is always possible to ensure $l$-diversity. So, small modifications have been implemented to neglect the $l$ value for the sake of the experiment. The classification accuracy of the anonymized data is compared with original data and benchmark approaches such as Mondrian [48], IACk [49], Attribute centric Fixed Interval (Ac-FI) anonymity [17], and Incognito [50], Datafly [51], and KMDAE-DAC [53]. According to the graph shown in Figure 7 (a) evince that the classification accuracy of the SVM model on the anonymized dataset is higher than state-of-the-art approaches. Similarly,

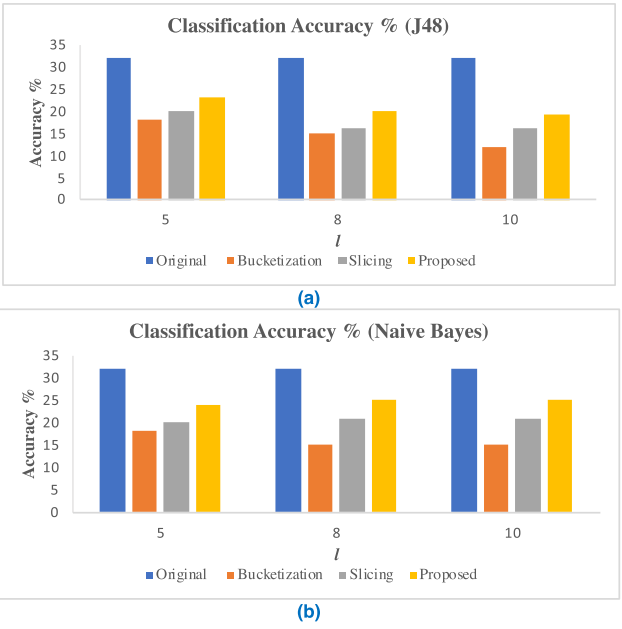


**FIGURE 5.** Classification accuracy of sensitive attribute (Occupation) - Adult-15 dataset (a) Decision tree (J48) (b) Naïve bayes.

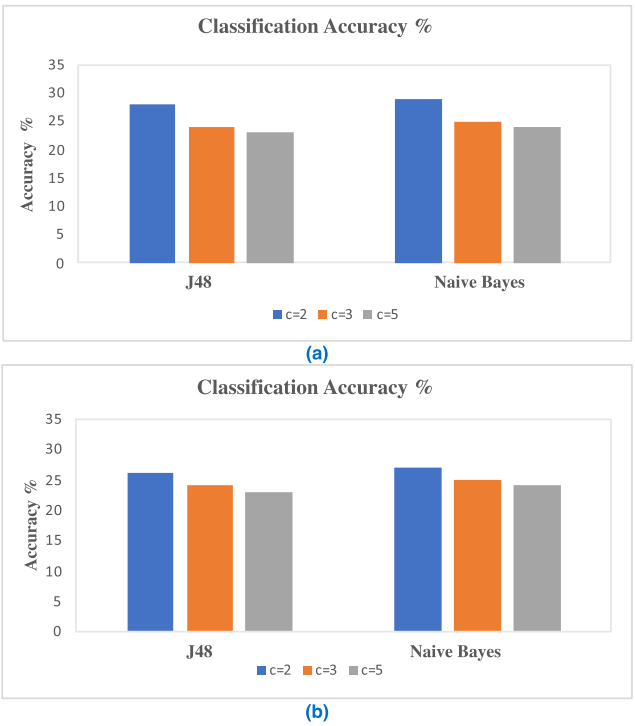


**FIGURE 6.** Classification accuracy for varying *c* {*2, 3, 5*} (a) Varying *c* with *l=5* (b) Varying *c* with *l=8*.

for Figure 7 (b) the classification accuracy of RF and Figure 7 (c) the classification accuracy of ANN are compared and found that the proposed approach yields higher classification accuracy. This shows that even with anonymization, the dataset has not lost essential information for classification and the information loss is minimal compared to the benchmarked approaches.

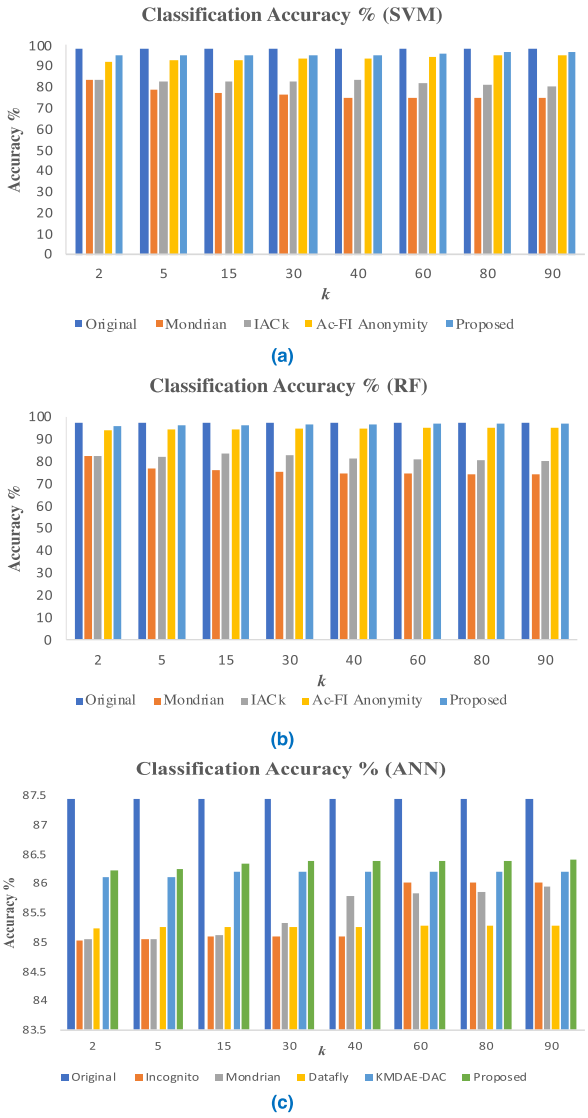


**FIGURE 7.** Classification accuracy of sensitive attribute (Salary) - Adult-15 dataset (a) Classification accuracy (SVM) (b) Classification accuracy (RF) (c) Classification accuracy (ANN).

Figure 8 shows the information loss of the proposed approach. We utilized the de facto standard information loss metric Normalized Certainty Penalty (NCP) [59] to measure the information loss. NCP calculates the information loss based on the equivalence class formed during the anonymization process. It gives the percentage of information loss after the anonymization compared to the original attributes. To evaluate the performance of the proposed approach, NCP values from state-of-the-art anonymization approaches such as Mondrian [48], $(k,k_m)$-anonymity[52], and clustering-based $k$-anonymity [10]. We observed an increase in the information loss for a higher $k$ value. This is because when the $k$ value is increasing the size of the bin (equivalence class) expands so more anonymized records are grouped under each bins. Hence, the data publisher can decide on the value of the privacy parameter. The graph shows that

the proposed approach has an acceptable range of information loss (as it yields better data utility with the classification models in Figure 7) and minimal compared to similar approaches.

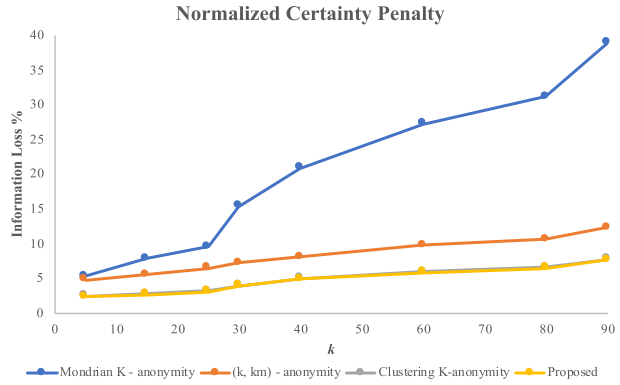


**FIGURE 8.** Normalized Certainty Penalty (NCP) - Information loss is calculated with respect to varying $k$ size.

Figure 9 presents the evaluation of the computation time of the proposed anonymization scheme with Mondrian, IACk, and HB-Anonymity [54] approaches. It shows that the proposed scheme has less computational complexity compared to the similar privacy preserving studies. The execution time is measured for the Adult-15 dataset with different $k$ values. It is observed that Mondrian approach complexity increases as the value of $k$ increases. Since, the value of $c$ and $l$ are fixed the variations in $k$ does not have much impact in the running time.

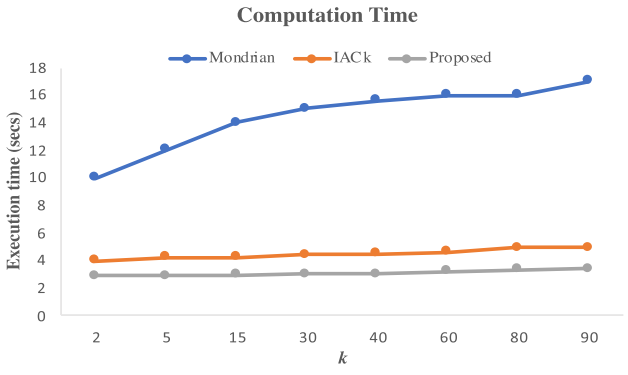


**FIGURE 9.** Computation time (secs) of proposed algorithm vs mondrian & IACk.

### C. DISCUSSION

In this section, we analyze the proposed anonymization scheme for membership, identity, and attribute disclosure protections.

#### 1) MEMBERSHIP DISCLOSURE PROTECTION

Membership disclosure protection protects against adversaries from discerning an individual's presence in published data. Membership disclosure can lead to identity and attribute disclosure. The proposed anonymization scheme offers membership disclosure protection through the $k$-anonymity property of the fixed-interval approach and the slicing technique. The anonymized microtable has equivalence classes consisting of $k$ tuples that are indistinguishable from at least $k$-1 other tuples. In addition, the original values of the tuples were replaced with the calculated mean for that specific interval. Further, slicing approach ensures multiple matching buckets for each tuple (i.e., Bucket $B == t \iff \forall 1 \leq i \leq c, t[C_i] \in B[C_i]$). Hence, the proposed anonymization scheme is efficient in protecting membership disclosures.

#### 2) IDENTITY DISCLOSURE

Identity disclosure occurs when an individual's QI attributes are linked to published data. An adversary may attempt to discern an individual with the background knowledge he/she possesses. Identity disclosure protection is ensured in the proposed scheme through the $k$-anonymity principle, where the probability of identity disclosure is less than $1/k$. Furthermore, identity disclosure is protected if membership disclosure is also protected.

In the slicing process, the columns are partitioned and the attribute values are permutated within each bucket to disassociate the links between columns. This process may create invalid tuples that can lead to attribute disclosure risk. For example, tuple $t_9$ in Table 7 describes a 39 year old Male is suffering from a breast cancer. Because a male cannot suffer from breast cancer, adversaries can use background knowledge and disclose the sensitive attributes of the individual. To reduce such risks, Hasan *et al.* proposed a value swapping method [64]. The value swapping method identifies invalid records in an anonymized microtable and swaps invalid records with relevant valid records in the bucket.

#### 3) ATTRIBUTE DISCLOSURE PROTECTION

Attribute disclosure occurs when an adversary attempts to acquire more knowledge about an individual (e.g., diagnosis). Identity disclosure can lead to attribute disclosure when the adversary can identify an individual's record in the published data and acquire their sensitive attributes. Sensitive attributes are at the risk of disclosure when all equivalence classes have a single sensitive attribute. The proposed anonymization scheme has a two-fold attribute disclosure protection. The fixed-interval approach protects numerical attributes by replacing the original attributes with the calculated mean value for specific intervals. Hence, every interval shares a common numerical QI attribute that makes the probability of an adversary discerning the individual attribute to $1/k$. The slicing approach protects categorical and sensitive attribute disclosures through tuple and attribute partitioning. When an adversary matches the background knowledge with the attributes in the bucket, the corresponding attribute column is permutated and cannot be discerned. Further, the sensitive attributes are protected through *l-diversity*, where the probability of discerning sensitive attributes is $1/l$.

For example, an adversary has background knowledge {Female, 22, 620709} and attempts to infer sensitive attributes from anonymized microdata in Table 9. First, the adversary finds the matching bucket for tuple $t$ by examining column values. The first column of Table 9 represents {Sex, Age} while the matching {Female, 22} adversary discerns $B_1$ as the matching bucket. So $p(t, B) = 1$. Now, the adversary attempts to compute $p(t, s)$ to discern the sensitive attribute. Table 9 is 2-diverse; hence, for a matching tuple, the probability of learning the correct attribute was less than $1/2$.

#### 4) COMPLEXITY ANALYSIS

The complexity of the proposed anonymization scheme was analyzed in three phases. In the fixed interval anonymization phase, the tuples of the table are divided into intervals, and an equivalence class is formed, similar to the $k$-anonymity property. The tuple anonymization problem is NP-hard [65] where the complexity is at $O(n^2)$. According to the optimal $k$-anonymity complexity, it can be reduced to $O(k \log m)$ where $m$ is the degree of the relation. In the clustering phase, the $k$-medoids clustering algorithm suffers from a high computational complexity, $O(k\,(n-k)^2)$. In our work, we utilized the PAM algorithm proposed by Park and Jun [57] which has a reduced complexity $O(nk)$ where $n$ is the number of records. Finally, the complexity of $l$-diverse slicing requires $O(n^2)$ to check the $l$-diversity for each bucket. The overall time complexity of the slicing phase is $O(n^2 \log n)$.

#### 5) DATA UTILITY ANALYSIS

Information loss is inevitable during the anonymization process. The primary goal of a PPDP scheme is to provide the maximum data utility while preserving privacy. To enhance the data utility, the anonymized values should be maintained as close as possible to the original values. However, this may leak private information. Therefore, the data utility of the PPDP scheme is measured by building quality models on the anonymized data values, as well as the model built on the original values. A possible argument here is that one needs a model, not the value released as a table. Different models are available and each model has unique model parameters. Therefore, the release of the built model may be suitable for all users.

Value consistency is important for model building. Models built on multiattribute domains and overlapping intervals produced inappropriate models. The proposed method maintains value consistency during the anonymization process through a fixed-interval approach. In this study, we evaluate the proposed PPDP scheme through classification models such as decision tree [66], naïve Bayes [67], support vector machine (SVM) [68], random forest (RF) [69], and artificial neural network (ANN) [70].

#### 6) VALIDITY OF WORK

Validity of work is to show how well the proposed scheme efficiently prevents privacy leaks and provides data utility. The validity of the work can be categorized as internal, external, statistical, and construct validity. The internal validity of the attribute-focused scheme can be discussed with the help of experimental analysis. To conduct the experiments we utilized a publicly available dataset that consists of 15 attributes. In order to show the variations of the dataset, we divided the dataset into two one with 7 attributes and another with 15 attributes considering one sensitive and the rest as quasi-attributes. We implemented Naïve Bayes and Decision Tree classification models to measure the accuracy of the proposed approach. The classification accuracy of the proposed approach is analyzed in terms of varying l values and observed slight variations in the accuracy. Also, we noticed that a change in the number of attributes does not impact the performance.

To external validity of the proposed work can be validated through the variations in the datasets. We modified the dataset by considering different parameters. At first, we considered 7 and 15 attributes of the dataset separately for different experiments. Then the privacy parameters k and l are varied according to measure the accuracy and information loss. Also, we compared our experimental results with benchmark algorithms to prove that our work is generalized and applicable to various situations and scenarios. The statistical validity of the proposed work is validated based on the statistical values of the state-of-the-art anonymization algorithms. The benchmarked algorithm results were extracted and compared with the proposed scheme's results. Based on the comparison we found that the proposed work is approximately 13% more accurate and incurs 12% less information loss. Finally, the construct validity is to ensure that we attained our objective. The major objective of this research is to anonymize the EHR data in such a way that it provides better classification utility, minimal information loss, and is computationally less expensive. Based on the experimental results we can validate that we achieved our primary goals.

## VI. CONCLUSION

Electronic Health Records (EHR) are vital for various medical researches. However, privacy concerns make it challenging to publish the data for research. Numerous privacy preserving approaches are proposed recently to preserve the privacy of healthcare data while publishing. But the privacy breach and improper data utility problems are still prevalent. In this paper, we proposed an attribute focused anonymization scheme to protect the privacy of EHR during data publishing. The proposed scheme is two-fold, it comprises a fixed interval approach to protect the numerical attributes of the EHR and an improved $l$-diverse slicing approach to protecting the categorical and sensitive attributes. The fixed interval approach is based on the $k$-anonymity technique where the attributes are generalized through a special operation proposed to increase the data utility. The proposed improved l-diverse slicing approach performs tuple partitioning and diversity checks to protect the categorical and sensitive attributes. Privacy requirement is determined by the privacy parameters $k$ for numerical and $l$ for categorical

and sensitive attributes. Experimental evaluations are conducted on real-world datasets and show that the classification utility of the proposed scheme is superior compared to benchmarked anonymization algorithms. Quantitatively, the proposed scheme's classification accuracy of SVM, RF, and ANN is at least ≈13% better. Normalized certainty penalty (NCP) is also utilized to measure the information loss during the anonymization process and found the proposed scheme incurred minimal information loss (≈12% less) compared to popular anonymization methods. Execution time is also evaluated in terms of the privacy parameter k and found that the proposed approach is computationally less expensive. Further, we discuss possible disclosure risks and theoretically demonstrate the resilience of our scheme. The future direction of this work is to consider the quasi-attributes as sensitive and semi-sensitive attributes. In this way, the protection of the attributes can be enhanced through sophisticated techniques combining *k*-anonymity, *l*-diversity, and *t*-closeness techniques. The work can be extended to applications other than healthcare.

## ACKNOWLEDGMENT

The authors would like to thank all our universities and institutes for facilitating time and support in this study.

## REFERENCES

[1] T. C. Clark and A. F. Westin, "Privacy and freedom," *California Law Rev.*, vol. 56, no. 3, p. 911, May 1968, doi: 10.2307/3479272.

[2] P. F. Wu, J. Vitak, and M. T. Zimmer, "A contextual approach to information privacy research," *J. Assoc. Inf. Sci. Technol.*, vol. 71, no. 4, pp. 485–490, Apr. 2020, doi: 10.1002/asi.24232.

[3] J. Andrew and J. Karthikeyan, "Privacy-preserving Internet of Things: Techniques and applications," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 3229–3234, Aug. 2019, doi: 10.35940/ijeat.F8830.088619.

[4] X. Wang, S. Garg, H. Lin, G. Kaddoum, J. Hu, and M. S. Hossain, "PPCS: An intelligent privacy-preserving mobile-edge crowdsensing strategy for industrial IoT," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10288–10298, Jul. 2021, doi: 10.1109/JIOT.2020.3032797.

[5] P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, "A novel hybrid K-means and GMM machine learning model for breast cancer detection," *IEEE Access*, vol. 9, pp. 146153–146162, 2021, doi: 10.1109/ACCESS.2021.3123425.

[6] A. Andrushia, K. Sagayam, H. Dang, M. Pomplun, and L. Quach, "Visual-saliency-based abnormality detection for MRI brain images—Alzheimer's disease analysis," *Appl. Sci.*, vol. 11, no. 19, p. 9199, Oct. 2021, doi: 10.3390/app11199199.

[7] G. N. Sundar, D. Narmadha, A. A. A. Jone, K. M. Sagayam, H. Dang, and M. Pomplun, "Automated sleep stage classification in sleep apnoea using convolutional neural networks," *Informat. Med. Unlocked*, vol. 26, Jan. 2021, Art. no. 100724.

[8] J. A. Onesimu and J. Karthikeyan, "An efficient privacy-preserving deep learning scheme for medical image analysis," *J. Inf. Technol. Manag.*, vol. 12, pp. 50–67, Dec. 2021, doi: 10.22059/jitm.2020.79191.

[9] J. Andrew, S. S. Mathew, and B. Mohit, "A comprehensive analysis of privacy-preserving techniques in deep learning based disease prediction systems," *J. Phys. Conf.*, vol. 1362, no. 1, pp. 1–9, 2019, doi: 10.1088/1742-6596/1362/1/012070.

[10] J. A. Onesimu, J. Karthikeyan, and Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services," *Peer Peer Netw. Appl.*, vol. 14, no. 3, pp. 1629–1649, Feb. 2021, doi: 10.1007/s12083-021-01077-7.

[11] A. Pika, M. T. Wynn, S. Budiono, A. H. M. ter Hofstede, W. M. P. van der Aalst, and H. A. Reijers, "Towards privacy-preserving process mining in healthcare," in *Business Process Management Workshops* (Lecture Notes in Business Information Processing), vol. 362. Switzerland: Springer, Sep. 2019, pp. 483–495, doi: 10.1007/978-3-030-37453-2_39.

[12] U. Department of Health and Human Services. (2019). *What is an Electronic Health Record (EHR)? | HealthIT.gov*. Official Website of The Office of the National Coordinator for Health Information Technology (ONC). Accessed: Aug. 25, 2020. [Online]. Available: https://www.healthit.gov/faq/what-electronic-health-record-ehr

[13] S. Hoffman, "EHR systems: Attributes, benefits, and shortcomings," in *Electronic Health Records and Medical Big Data*. Cambridge, U.K.: Cambridge Univ. Press, 2016, pp. 9–37.

[14] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surveys*, vol. 42, no. 4, pp. 1–53, Jun. 2010, doi: 10.1145/1749603.1749605.

[15] T. Craig and M. Ludloff, *Privacy & Big Data*. O'Reilly Media, Aug. 2011.

[16] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "A novel differential privacy approach that enhances classification accuracy," in *Proc. 9th Int. Conf. Comput. Sci. Softw. Eng.*, 2016, pp. 79–84, doi: 10.1145/2948992.2949027.

[17] A. Majeed, "Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 31, no. 4, pp. 426–435, Oct. 2019, doi: 10.1016/j.jksuci.2018.03.014.

[18] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.

[19] A. Zigomitros, F. Casino, A. Solanas, and C. Patsakis, "A survey on privacy properties for data publishing of relational data," *IEEE Access*, vol. 8, pp. 51071–51099, 2020, doi: 10.1109/ACCESS.2020.2980235.

[20] R. Chi-Wing, J. Li, A. W.-C. Fu, and K. Wang, "($\alpha$, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, 2006, pp. 754–759, doi: 10.1145/1150402.1150499.

[21] S. A. Abdelhameed, S. M. Moussa, and M. E. Khalifa, "Privacy-preserving tabular data publishing: A comprehensive evaluation from Web to cloud," *Comput. Secur.*, vol. 72, pp. 74–95, Jan. 2018, doi: 10.1016/j.cose.2017.09.002.

[22] X. Sun, H. Wang, J. Li, T. M. Truta, and P. Li, "($p^+$, $\alpha$)-sensitive k-anonymity: A new enhanced privacy protection model," in *Proc. 8th IEEE Int. Conf. Comput. Inf. Technol.*, Jul. 2008, pp. 59–64, doi: 10.1109/CIT.2008.4594650.

[23] J. Cheng, A. W.-C. Fu, and J. Liu, "K-isomorphism: Privacy preserving network publication against structural attacks," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2010, pp. 459–470, doi: 10.1145/1807167.1807218.

[24] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 163–175, Nov. 2014, doi: 10.1007/s00779-012-0633-z.

[25] 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, doi: 10.1109/ICDE.2006.1.

[26] 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, doi: 10.1109/ICDE.2007.367856.

[27] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*, 2003, pp. 211–222, doi: 10.1145/773153.773174.

[28] Q. Wang and X. Shi, "(a, d)-diversity: Privacy protection based on l-diversity," in *Proc. WRI World Congr. Softw. Eng.*, 2009, pp. 367–372, doi: 10.1109/WCSE.2009.362.

[29] B. K. Tripathy, A. Maity, B. Ranajit, and D. Chowdhuri, "A fast p-sensitive l-diversity anonymisation algorithm," in *Proc. IEEE Recent Adv. Intell. Comput. Syst.*, Sep. 2011, pp. 741–744, doi: 10.1109/RAICS.2011.6069408.

[30] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1388–1399, Jul. 2012, doi: 10.14778/2350229.2350255.

[31] W. Wang, L. Chen, and Q. Zhang, "Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation," *Comput. Netw.*, vol. 88, pp. 136–148, Sep. 2015, doi: 10.1016/j.comnet.2015.06.014.

[32] J. Andrew and J. Karthikeyan, "Privacy-preserving big data publication:(K, L) anonymity," in *Intelligence in Big Data Technologies-Beyond the Hype*. Cham, Switzerland: Springer, 2020, pp. 77–88.

[33] H. Attaullah, T. Kanwal, A. Anjum, G. Ahmed, S. Khan, D. B. Rawat, and R. Khan, "Fuzzy-logic-based privacy-aware dynamic release of IoT-enabled healthcare data," *IEEE Internet Things J.*, vol. 9, no. 6, pp. 4411–4420, Mar. 2022, doi: 10.1109/JIOT.2021.3103939.
[34] S. Kim and Y. D. Chung, "An anonymization protocol for continuous and dynamic privacy-preserving data collection," *Future Gener. Comput. Syst.*, vol. 93, pp. 1065–1073, Apr. 2019, doi: 10.1016/j.future.2017.09.009.
[35] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness," *IEEE Trans. Depend. Sec. Comput.*, vol. 16, no. 4, pp. 580–593, Jul. 2019, doi: 10.1109/TDSC.2017.2698472.
[36] B. B. Mehta and U. P. Rao, "Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1423–1430, Apr. 2022, doi: 10.1016/j.jksuci.2019.08.006.
[37] S. A. Abdelhameed, S. M. Moussa, and M. E. Khalifa, "Restricted sensitive attributes-based sequential anonymization (RSA-SA) approach for privacy-preserving data stream publishing," *Knowl.-Based Syst.*, vol. 164, pp. 1–20, Jan. 2019, doi: 10.1016/j.knosys.2018.08.017.
[38] K. Wang, P. Wang, A. W. Fu, and R. C.-W. Wong, "Generalized bucketization scheme for flexible privacy settings," *Inf. Sci.*, vol. 348, pp. 377–393, Jun. 2016, doi: 10.1016/j.ins.2016.01.100.
[39] R. Indhumathi and S. S. Devi, "Anonymization based on improved bucketization (AIB): A privacy-preserving data publishing technique for improving data utility in healthcare data," *J. Med. Imag. Health Informat.*, vol. 11, no. 12, pp. 3164–3173, Dec. 2021, doi: 10.1166/JMIHI.2021.3901.
[40] J. W. Byun, A. Kamra, E. Bertino, and N. Li, "*Efficient K-Anonymization Using Clustering Techniques* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4443. Berlin, Germany: Springer, 2007, pp. 188–200.
[41] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, 2005, pp. 217–228, doi: 10.1109/ICDE.2005.42.
[42] Q. Gong, J. Luo, M. Yang, W. Ni, and X.-B. Li, "Anonymizing 1: M microdata with high utility," *Knowl.-Based Syst.*, vol. 115, pp. 15–26, Jan. 2017, doi: 10.1016/j.knosys.2016.10.012.
[43] L.-E. Wang and X. Li, "A clustering-based bipartite graph privacy-preserving approach for sharing high-dimensional data," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 24, no. 7, pp. 1091–1111, Sep. 2014, doi: 10.1142/S0218194014500363.
[44] A. Abbasi and B. Mohammadi, "A clustering-based anonymization approach for privacy-preserving in the healthcare cloud," *Concurrency Comput. Pract. Exper.*, vol. 34, no. 1, 2021, Art. no. e6487, doi: 10.1002/cpe.6487.
[45] B. K. Rai, "Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records," *Int. J. Reliable Qual. E-Healthcare*, vol. 11, no. 1, pp. 1–15, Jan. 2022, doi: 10.4018/IJRQEH.297076.
[46] S. Arca and R. Hewett, "Analytics on anonymity for privacy retention in smart health data," *Future Internet*, vol. 13, no. 11, p. 274, Oct. 2021, doi: 10.3390/FI13110274.
[47] K. M. Chong and A. Malip, "Bridging unlinkability and data utility: Privacy preserving data publication schemes for healthcare informatics," *Comput. Commun.*, vol. 191, pp. 194–207, Jul. 2022, doi: 10.1016/J.COMCOM.2022.04.032.
[48] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE)*, 2006, p. 25, doi: 10.1109/ICDE.2006.101.
[49] J. Li, J. Liu, M. Baig, and R. C.-W. Wong, "Information based data anonymization for classification utility," *Data Knowl. Eng.*, vol. 70, no. 12, pp. 1030–1045, Dec. 2011, doi: 10.1016/j.datak.2011.07.001.
[50] R. C. W. Wong, A. W. C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 543–554.
[51] J. Sedayao, "Enhancing cloud security using data anonymization," Intel IT, IT Best Practices, Cloud Comput. Inf. Secur., IT@ Intel White Paper 17, 2012. [Online]. Available: https://www.intel.co.kr/content/dam/www/public/us/en/documents/best-practices/enhancing-cloud-security-using-data-anonymization.pdf
[52] G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, *Anonymizing Data With Relational and Transaction Attributes* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8190. Berlin, Germany: Springer, 2013, pp. 353–369.
[53] N. Yuvaraj, K. Praghash, and T. Karthikeyan, "Data privacy preservation and trade-off balance between privacy and utility using deep adaptive clustering and elliptic curve digital signature algorithm," *Wirel. Pers. Commun.*, vol. 2021, pp. 1–16, Jan. 2021, doi: 10.1007/s11277-021-09376-1.
[54] J. Jayapradha, M. Prakash, Y. Alotaibi, O. I. Khalaf, and S. A. Alghamdi, "Heap bucketization anonymity—An efficient privacy-preserving data publishing model for multiple sensitive attributes," *IEEE Access*, vol. 10, pp. 28773–28791, 2022, doi: 10.1109/ACCESS.2022.3158312.
[55] F. N. David and H. Cramer, *Mathematical Methods of Statistics*, vol. 34, nos. 3–4. Princeton, NJ, USA: Princeton Univ. Press, 1947.
[56] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Model.*, vol. 53, nos. 1–2, pp. 91–97, Jan. 2011, doi: 10.1016/j.mcm.2010.07.022.
[57] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009, doi: 10.1016/j.eswa.2008.01.039.
[58] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach for privacy preserving data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 3, pp. 561–574, Mar. 2012, doi: 10.1109/TKDE.2010.236.
[59] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *Proc. 33rd Int. Conf. Very Large Data Bases (VLDB)*, 2007, pp. 758–769. Accessed: Feb. 24, 2019. [Online]. Available: https://dl.acm.org/citation.cfm?id=1325938
[60] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository: Adult Data Set*. Accessed: Mar. 2, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult
[61] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 5, pp. 711–725, May 2007, doi: 10.1109/TKDE.2007.1015.
[62] J. Andrew, J. Karthikeyan, and J. Jebastin, "Privacy preserving big data publication on cloud using Mondrian anonymization techniques and deep neural networks," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2019, pp. 722–727, doi: 10.1109/ICACCS.2019.8728384.
[63] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. 32nd Int. Conf. Very Large Data Bases*, 2006, pp. 139–150. [Online]. Available: http://dl.acm.org/citation.cfm?id=1182635.1164141
[64] A. S. M. T. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Secur. Commun. Netw.*, vol. 9, no. 16, pp. 3219–3228, Nov. 2016, doi: 10.1002/sec.1527.
[65] 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, doi: 10.1145/1055558.1055591.
[66] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991, doi: 10.1109/21.97458.
[67] I. Rish, "IBM research report an empirical study of the naive Bayes classifier," *Science*, vol. 22230, no. 22, pp. 41–46, 2001.
[68] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999, doi: 10.1023/A:1018628609742.
[69] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.
[70] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, 2000, doi: 10.1016/S0731-7085(99)00272-1.

**J. ANDREW ONESIMU** received the B.E. degree in CSE and the M.E. degree from Anna University, Chennai, India, in 2011 and 2013, respectively, and the Ph.D. degree from the Vellore Institute of Technology (VIT), Vellore, India. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering (CSE), Manipal Institute of Technology (MIT), Manipal, India. He is an active researcher. He is having nine years of teaching experience at undergraduate (UG) and post graduate (PG) levels. He has also supervised number of projects at different levels in University. He has published more than 25 scientific research papers in reputed journals and conferences. His research interests include data privacy, healthcare data analysis, deep learning, machine learning, computer vision, and blockchain technologies.

**KARTHIKEYAN J** received the B.Sc., M.C.A., and Ph.D. degrees from the VIT University, India, in 2005, 2010, and 2013, respectively. He is currently serving as an Associate Professor with the Department of Software and Systems Engineering, School of Information Technology and Engineering, VIT University. He is currently guiding five research scholars. He has authored and coauthored several research articles, book chapters, and conference contributions. His research interests include machine learning, big data security, privacy-preserving data publishing, and deep learning.

**JENNIFER EUNICE** received the B.E. degree in ECE and the M.E. degree in VLSI design from Anna University, India, in 2012 and 2014, respectively. She is currently a full-time Research Scholar with the Department of Electronics and Communication Engineering (ECE), Karunya Institute of Technology and Sciences (KITS), Coimbatore, India. She had three years of teaching experience in UG and PG courses. She also supervised PG projects and guided various UG projects related to VLSI. Her research interests include VLSI design, VLSI analog circuits, signal processing, and deep learning.

**MARC POMPLUN** is currently a Professor of computer science with the University of Massachusetts Boston and the Director of the Visual Attention Laboratory. His current research interests include analysing, modeling, and simulating the aspects of human vision.

**HIEN DANG** (Member, IEEE) received the Ph.D. degree in computer science, in 2010. She is currently a Research Scholar with the University of Massachusetts Boston, USA. She is also working with the Faculty of Engineering and Computer Science, Thuyloi University, Vietnam. She is the author or editor of four books and about 40 articles. In addition, she is a team leader or member of many national or ministerial and corporate projects to solve real-world problems. Her research interests include artificial intelligence, artificial neural networks, deep learning, big data analytics, and some healthcare problems.