Best-Practice-Based Framework for User-Centric Privacy-Preserving Solutions in Smart Home Environments

Chathurangi Ishara Wickramasinghe(✉)

Georg-August-University Göttingen, Göttingen, Germany
c.wickramasinghe@stud.uni-goettingen.de

Abstract. The rapid technological progress causes smart environments, such as smart homes, cities, etc., to become more ubiquitous in our daily lives. Privacy issues arise when the smart objects in those smart environments collect and disclose sensitive data without users' consent. Therefore, existing works and the European General Data Protection Regulation (GDPR) are still calling for privacy-preserving solutions with more user involvement and automated decision-making. Existing works show research gaps regarding context-aware privacy-preference modellings. They do not present best-practice-based frameworks for user-centric privacy-preserving approaches allowing context-aware adapting of users' privacy and data disclosure preferences while considering their past activities. Hence, this paper proposes a best-practice-based framework for user-centric privacy-preserving solutions with automation options. The proposed approach supplies users data sharing recommendations with minimum human interference while considering (1) GDPR requirements, (2) context-sensitive factors and (3) users' past activities. The paper also outlines how the proposed framework can be integrated in an existing user-centric privacy-preserving approach in the future. In this way, the proposed approach can be integrated in the existing IoT architecture systems, which allow users to control the entire data collection, storage and disclosure process in smart home environments.

Keywords: Machine learning · Privacy preserving · Smart homes · Sensitivity · Data protection · Smart environments · Smart objects · Ubiquitous computing · Pervasive systems

1 Introduction

The technological progress in the context of pervasive systems leads to the fact that more and more smart objects are integrated into our personal spaces, such as in homes [11]. The integrated smart objects in such smart home environments are, for instance, smart bulbs, fridges, door locks, etc. [11]. Although these smart objects improve our lives, they also collect and disclose a vast amount of sensitive and non-sensitive data without users' consent [49,50]. Especially in smart home

S. Longfei and P. Bodhi (Eds.): MobiQuitous 2022, LNICST 492, pp. 101–120, 2023.
https://doi.org/10.1007/978-3-031-34776-4_6

environments, this privacy issue gains more importance because the integrated smart objects in smart homes collect data in a personal space [15]. In order to address the arising privacy issues in this context, several privacy-preserving solutions, including different machine learning and automated approaches, have been proposed, such as [3,22,23,26]. Note that most of them do not (1) allow the entire control over the data collection, storage and disclosure process [45] and also do not (2) supply users with data disclosure recommendations based on best practices, context-sensitive factors and users' past activities [25].

GDPR (Art. 4, 5, 9, 12, 15, 17, 19, 22 and 23) and existing works are still calling for privacy-preserving solutions with more user-centricity and possibility to consider context-aware user privacy preferences [1,25,38].

This paper proposes a best-practice-based framework for privacy-preserving in smart home environments to address these open issues. The proposed approach delivers user data disclosure recommendations while considering GDPR-based best practices, users' context-aware sensitive factors and their past activities. In order to allow users to control the entire data collection, storage and disclosure, the proposed framework is integrated within an existing user-centric privacy-preserving approach from [45]. The approach from [45] include four **User-Centric-Control-Points** (UCCPs), which can be integrated in existing IoT architecture systems. The four UCCPs from [45] include the following features:

- UCCP 1 - Data Object Tagging: Allows users to set their general sensitivity awareness by assigning themselves to one of the described profiles.
- UCCP 2 - Data Minimization and Aggregation: Allows users to minimize the collected data by the smart object sensors and set the aggregation period for the review regarding the collected data before data sharing.
- UCCP 3 - Data Sharing: Allows users to assess the sensitivity of the collected data types and associated privacy risks and advantages in the data disclosure context. Moreover, it also allows users to set their risk aversion or risk affinity.
- UCCP 4 - Data Access Limitations: Allows users to limit the data sharing while setting the data consumers and usage purposes of the shared data after considering the model's recommendations.

Further details regarding the integration of the proposed framework from this paper in the approach from [45] are described in Sect. 2. However, the implementation of the proposed approach is out of this manuscript. To sum up, the proposed approach in this paper contributes to the following points compared to previous work: (1) Providing users with best-practice-based and context-aware recommendations regarding data disclosure and (2) allowing them to consider users' privacy preferences from past activities with minimum human interference. Additionally, the integration of the proposed framework in [45] allows user-centric privacy-preserving in smart home environments.

This paper is structured as follows. Firstly, the derived best-practice-based framework and its integration in [45] are presented in Sect. 2 and its qualitative evaluation in Sect. 3. In Sect. 4, the proposed approach is discussed, and in

Sect. 5 the related work is presented. Closing remarks conclude this paper in Sect. 6, respectively.

2 The Framework for User-Centric Privacy-Preserving Approaches in Smart Homes

2.1 Proposed Framework

The Fig. 1 presents the proposed best-practice-based framework for user-centric privacy-preserving solutions. The proposed framework includes a supervised learning method, including the decision tree, which is an essential, efficient and significant way to find logical connections between learned and predicted items [46]. It also contains an active learning method, Support Vector Machines (SVM), which is a successful method for real-world learning [22,27]. Integrating the decision tree and SVM algorithm allows the proposed approach to run its technique in a less time-, cost-, and energy-consuming way [48]. This deployment allows the proposed approach to work in a privacy-preserved way since the data does not need to leave its smart home environment in order to be processed. In this way, data leakage can be prevented by applying the proposed approach. The proposed framework supplies users with data disclosure recommendations (1) based on GDPR-related best practices while (2) considering the impacts of users' past activities on their context-sensitive privacy preferences in those data disclosure recommendations. The integration of this best-practice-based framework in an existing user-centric privacy-preserving approach from [45] allows (1) users to control the entire data collection and disclosure process with minimum human interference and (2) the integration in existing IoT architecture systems. In the following, the necessary inputs, as well as the proposed framework, are described.

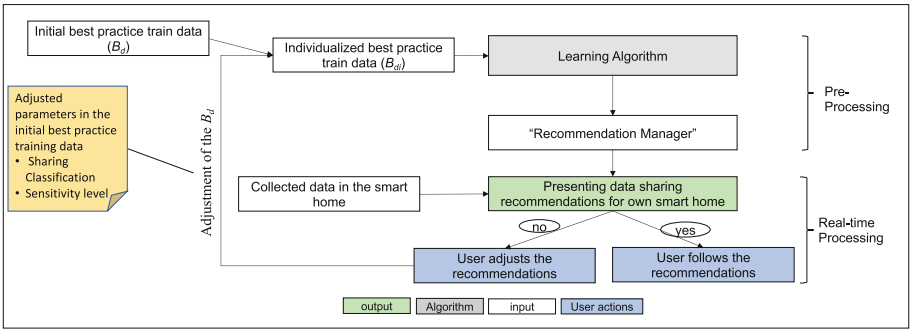

Fig. 1. Overview of the proposed best-practice-based automated framework

Initial Inputs for the Framework: In order to train the algorithm of the proposed framework to supply users with data disclosure recommendations, an initial best-practice train data (B_d) is derived and used for supervised learning. The B_d includes different information types, such as collected data type, smart object, user category, age, country, etc., which influence sensitivity perception and which then in turn influence users' privacy and data sharing attitude [6, 19,21,31,42]. The Table 1 presents an extract of the B_d, which is derived based on GDPR specifications, especially Art. 4, 5, 9, 12, 15, 17 and 19 [1,12,38] and literature review [4,17,28,31,39,41,45]. According to the GDPR demands and literature review the following information types are defined as most relevant variables of B_d and include the following definitions:

- **Data Type**: Collected data type by each smart home object [4,17,31,39,41, 45]
- **Smart Object**: Respective smart home object collecting the specific data type
- **Sensitivity Level**: Sensitivity Level of the respective collected data, according to GDPR demands [12]. The sensitivity level is set to 5 with "sensitive". In a later iteration, the proposed approach adjusts the sensitivity level of the corresponding data type based on users' past activities, which is described in the next section (scale: 1 = "non-sensitive" to 10 = "highly sensitive", with 5 = "sensitive")
- **Sharing Classification**: Sharing recommendations for the respective data collected based on the sensitive level derived from GDPR requirements [12] and literature analysis [45] (options: 0 = *do not share the data type* and 1 = *share the data type*). This variable can also be changed in a later iteration by the proposed approach according to users' preferences, which is also described in the next section
- **Age**: Age is considered as one of the influencing factors of sensitivity level [4,28,41] (groups: 1 = 18 - 30 years, 2 = 30 - 45 years and 3 = > 45 years)
- **User Category**: User category is considered as a further influencing factor of sensitivity level [28,41] (scale: 1 = unfamiliar users, who own smart home objects but not familiar with their usage, 2 = less familiar user, who own smart objects and are very little familiar with their usage and 3 = familiar user, who own smart objects and are very trusted with their usage). This clustering of the user categories also allows us to consider the technical skills of the users [4,39,41]
- **Country**: Countries are considered as influencing factors of the sensitivity level [28,41], and in this way, the culture as an influencing factor for sensitivity level can be considered [4,39,41]

Mechanisms of the Proposed Framework (A_f): In the first iteration to train the learning algorithm of A_f, the B_d is used so that the *Recommendation Manager* of A_f (see Fig. 1) can deliver data sharing recommendations while applying A_f in privacy-preserving approaches in the smart home context. This

Table 1. Extract of the initial best practice train data B_d [4,12,17,28,31,39,41,45]

Data Type	Smart Object	Sensitivity Level	Sharing Classification	Age	User Category	Country
Fingerprint	Smart door locks	5	0	3	1	Germany
Voice print	Smart speakers	5	0	2	1	UK
Medical history data	Smart wearable	5	0	2	3	Switzerland
Availability at home	Smart smoke detectors	5	0	3	2	Germany
Body images	Smart cameras	5	0	2	1	Austria

step is defined as "Pre-Processing" in Fig. 1 and in the first training iteration B_d and *individualized best practice train data* (B_{di}) contain the same data. After the first iteration, the automated solution can be applied in any smart home environment, defined as "Real-time Processing" in Fig. 1. The collected data types in that respective smart home environment are imported into the model in the second iteration. Based on the previous learning, the *Recommendation Manager* supplies users with data sharing recommendations. Based on the delivered recommendations, users have the opportunity to follow the recommendations or adjust them. In case, the users decide to adjust the recommendations, then two variables of B_d, **sharing classification** and **sensitivity level**, are adjusted and those adjustments are included in B_{di}[1]. The **sharing classification** is adopted according to users' settings, and the value can be changed between 1 and 0. The **sensitivity level** is adjusted according to the following scheme: In the first step, users are asked to set their perception regarding the dependencies between different collected data types in their smart home environment. The framework only asks users to indicate the dependencies for some of the collected data. In this step, the users will also be supplied with an "i" icon next to each collected data type, giving users some background information regarding the data type. An example for the supplied information ("i") regarding collected fingerprints or health data could be: "Biometrical data, such as fingerprints, voice prints, face IDs, describe specific characteristics of a natural human." or "Medical or health data, such as lifestyle data, wellness data, diagnoses, describe the way of corresponding human's lifestyle clearly." These details regarding "i" of corresponding data types are derived based on the data type clustering of previous surveys on data sensitivity, for instance [31,39,41]. An active learning method is used to derive the dependencies for the rest of the unassigned data based on users' indicated dependencies. As already mentioned, in this framework, the active learning method, SVM, is applied, which includes significant success in real-world learning functions for various reasons, such as the reduced need for labelling instances, good performance on unlabeled data [22,27]. When indicating the dependencies for a data sample, the users are also asked to set a *Weight Coefficient* between 0 (dependency is weak) to 1 (dependency is strong) for the indicated dependency. This *Weight Coefficient* is later used to adjust the sensi-

[1] From the second iteration the B_d and B_{di} do not contain the same data in case the users decide to adjust the data sharing recommendations.

tivity level of the dependent data types in case users decide to share one data type according to their preferences. One example in this context is presented in Fig. 2, and its results in Table 2. In this example, the user, Tim, owns four smart objects: (1) smart door locks collecting fingerprint and face ID, (2) smart speakers collecting voice print, (3) smart scale collecting weight and height data and (4) smart fridge collecting purchase data.

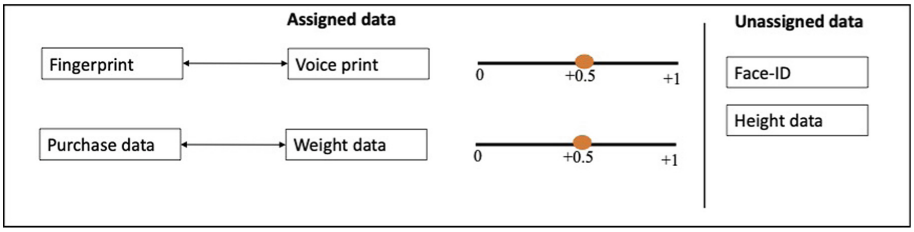

Fig. 2. Example scenario of Tim indicating the dependencies between collected data types

Table 2. Example scenario Tim: A possible result after assigning the data

Data Type	Dependency with other data types	Weight Coefficient for the Sensitivity Level	Assigned By
Fingerprint	Voice print	+ 0.5	user
Purchase Data	Weight data	+ 0.5	user
Fingerprint	Face ID	+ 0.5	framework
Weight Data	Height	+ 0.5	framework

Once the users decide to share an assigned data type, for example, fingerprint, twice in a row, which leads to the interpretation that the user does it deliberately, then the sensitivity level of the dependent data types, in Tim's scenario, voice print and face ID, will increase. In order to adjust the sensitivity level of the dependent data, in the proposed framework, the proportional-integral-derivative (PID) controller is used, a widely used control system in the industrial world and a variety of applications [7]. The technique of the PID controller includes the calculation of an error value *E(t)*, which is the difference between the expected set point *SP* and a measured process variable *PV* [7]. In the proposed framework, the PID controller is adapted as follows. The set point of a dependent data type SP_{dpn} is twice of *Weight Coefficient* of SP_{dpn} and PV_{dpn} is set to the corresponding *Weight Coefficient* of the dependent data type. In every iteration, the users decide to disclose a specific data type, the *Weight Coefficient* of the dependent data type will incrementally increase according to the set value of *Weight Coefficient* of the dependent data type (PV_{dpn}). In case the users decide to share a data type, the following calculation is performed:

$$E(t) = SP_{dpn} - PV_{dpn} \tag{1}$$

In case $E(t) \geq 0$, the sensitivity level of the dependent data type increases by one. The maximum value of the sensitivity level is ten ("highly sensitive"), and the values of the sensitivity level can be increased to ten. When the users decide to disclose dependent data, which has already been increased in sensitivity level, they are explicitly asked whether they are sure about this decision because of the already shared data in the past. In case the users decide not to disclose an already disclosed data type twice in a row, which leads to the interpretation that the user does it deliberately, then the sensitivity level of the corresponding dependent data will be set back to the sensitivity level from B_d. Increasing the sensitivity level of a data type will always cause the sharing classification to be set to 0 if this is not the case. Integrating this mechanism helps to consider users' past activities and context-sensitive perceptions while supplying them with data disclosure recommendations according to their privacy preferences. Figure 3 presents the above-described process of the PID controller in the proposed framework.

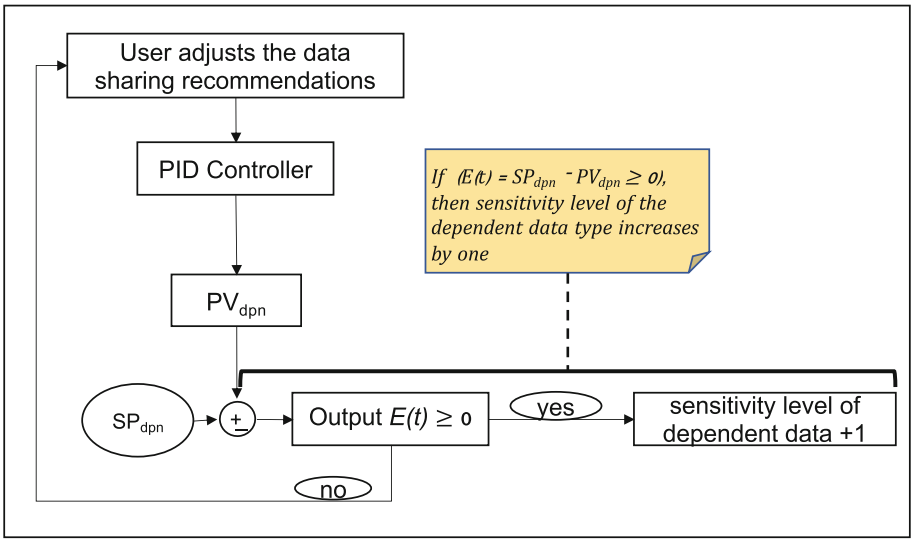

Fig. 3. Process of the adjusted PID controller in the proposed framework

2.2 Integration of the Proposed Framework in User-Centric Privacy-Preserving Approach

Integrating the described framework from the previous section in the user-centric privacy-preserving approach from [45] simplifies the original model for smart home environments. After the integration of the proposed framework in this paper, the modified user-centric privacy-preserving approach includes only two

User-Centric-Control-Points (UCCPs): (A) **UCCP 1: Data Aggregation** and (B) **UCCP 2: Data Sharing and Access Limitations**. The modified user-centric privacy-preserving approach is implemented in the Data Storage and Processing Node (*DSPN*) of the IoT device layer of the IoT system architecture, as recommended in [45]. Figure 4 illustrates the UCCPs of the modified model and their interrelationships.

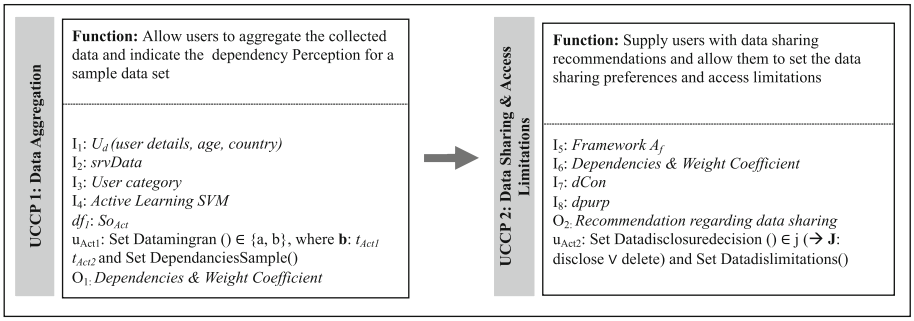

Fig. 4. Modified user-centric privacy-preserving approach for Smart Home Environments

UCCP 1: Data Aggregation: This UCCP allows users to set the aggregation period for their review *(Set Datamingran())*[2] from [45] and users are asked to set the settings during the **registration process**. In order to capture user details and their perception regarding the dependency between the collected data, users are asked to supply the inputs U_d[3], *User category* and to carry out the user action *Set DependanciesSample()*. Users are asked to indicate the dependencies for some of the collected data in this user action. This will help the active learning method I_4 to derive the dependencies for the unassigned data and deliver the output O_1: Dependencies & Weight Coefficient for UCCP 2, as described in Sect. 2.1. Furthermore, UCCP 1 also includes a further input I_2: srvData and default settings So_{Act}. I_2: srvData must be supplied by the smart object provider by the time when users install the specific smart home object in their smart home and include the collected data types by the smart home object service. The So_{Act} is set as in the original model[4] from [45].

[2] "The setting options regarding data aggregation allow end users to choose between two options. The two options are (1) the exact time of each action of the smart object for daily review (t_{Act1}) or (2) the time period users want to aggregate and review the collected data by their smart objects (t_{Act2}), for example, weekly, monthly. An example for t_{Act1} could be that the smart object owner is absent at 07:30 am on the 5th of February and present again at 8 pm in the living room. He gets up at 06:30 am and switches on his smart bulbs in two rooms, namely the bathroom and sleeping room. In contrast to this, an example for t_{Act2} could be that the smart object owner is available at home at various times per month and switches on his smart bulbs 200 times per month." [45].

[3] Examples for user details are age, country.

[4] "The default settings for So_{Act} regarding data aggregation layer is assigned to So_{Act1}, which means that the granularity of the data is set at the layer of sensors." [45].

UCCP 2: Data Sharing and Access Limitations: Complementary to UCCP 1, UCCP 2 supplies users data sharing recommendations while considering users' context-aware privacy preferences and past activities. Additionally, this UCCP also allows users to set their data sharing preferences and access limitations with the user actions, *Set Datadisclosuredecision ()* and *Set Datadislimitations()*. The UCCP 2 includes four inputs: (1) I_5: Framework A_f, (2) I_6: Dependencies & Weight Coefficient, (3) I_7: dCon[5] and (4) I_8: dpurp[6]. The proposed Framework (A_f) is integrated in this UCCP (I_5), as described in Sect. 2.1 and the O_1: Dependencies & Weight Coefficient from UCCP 1 is considered as the I_6. The data sharing recommendations are supplied during the **aggregation period**, and the users are also asked to carry out the user actions during this period. In case users decide not to follow the data sharing recommendations, supplied based on the A_f (O_2: Recommendation regarding data sharing), then the users are asked to set the *dCon* and *dpurp*. At this point the user preferences will be captured and used to update the two variables of B_d, **sharing classification** and **sensitivity level** according to the mechanisms of the A_f, as described in Sect. 2.1. In this UCCP, users can access the shared data with data consumers for different usage purposes. They can also disclose those already shared data with further data consumers for other usage purposes. The integration of the A_f in the approach from [45] allows supplying users with data sharing recommendations with minimum human interference while considering users' past activities and context-sensitive preferences. Moreover, integrating the proposed framework allows users to control the entire data collection, storage and disclosure process in smart home environments.

3 Evaluation

The proposed best-practice-based framework and its integration in an existing user-centric privacy-preserving approach address existing gaps, as mentioned in [25] as well as in the demands from GDPR, especially Art. 9, 12, 15, 17, 19 and 22 [1,12,38], and supply a solution regarding context-aware privacy-preference modelling. In the following sections, the evaluation of functional and non-functional requirements regarding the proposed framework is presented since, as mentioned at the beginning, the implementation of the framework is not a part of this manuscript.

3.1 Evaluation of Non-Functional Requirements

In this section, the non-functional requirements regarding the proposed framework are evaluated. The security-, privacy-, and performance-related non-functional requirements are derived based on existing literature [9,20] in the

[5] "*dCon* include third parties getting access to disclosed data, such as doctors, insurance company, government agencies, etc." [45].

[6] "... usage purposes informs end users for which purpose, such as personal health plan, statistical purposes, etc., the shared data are used by the *dCon*..." [45].

context of software and machine learning development. The considered non-functional requirements are: (1) performance, (2) maintainability, (3) legal, (4) portability, (5) deployability, (6) interoperability, (7) data integrity, (8) efficiency and capacity, (9) scalability, (10) availability. Since the real-world implementation of the proposed approach is not in the scope of the paper, the evaluation of non-functional requirements allows us to analyze the feasibility and the performance of the proposed framework.

Performance: According to [37], the user interfaces of software programs should be fast and deliver results within 250 ms. In order to address this performance issue in the proposed framework, several measurements are taken: (1) applied machine learning approaches and (2) technical equipment. Regarding the first point, two machine learning approaches are applied in the proposed approach: Decision tree and SVM. Decision trees are intuitive, more effortless, and not time-consuming in their implementation [48]. Furthermore, the decision trees include high flexibility and deliver high accuracy results regarding logical connections between learned and predicted items [48]. Additionally, SVM is considered as a powerful learning method allowing to achieve an efficient data classification with high accuracy [43] as well as to utilize less energy per second in comparison to other techniques, such as random forest, multi-layer perception, for instance, as outlined in a real-world IoT device experiment from [48]. Striking is that SVM includes complex training structures, which require more training time compared to other learning algorithms, as shown in the reference example from [48]. In the reference example [48], the training and inference speed of a data set with 100 instances (after data pre-cleansing) is tracked in seconds on an IoT device and outlines that training run time per instance is 0.75 s and inference run time per instance is 0.04 s. In the IoT and smart home context, where data sets contain millions of data, due to the second/minute measurement, it is useful to sample the valid training data set (after data cleansing according to [8,48]) while maintaining the representatives of the different values of various smart objects in own home environment [14] and considering user preferences regarding the aggregation period in order to allow less time-consuming learning for both applied algorithms, decision tree and SVM. Regarding the technical equipment, it is recommended to integrate a server element, for instance, an element with an i7 kernel and quad-core processor, in the technical component *DSPN* from [45] in order to integrate the both modified UCCPs with the learning algorithms of A_f in an energy-efficient and fast way. The real-world experiment [48] in the IoT context outlines based on different real-world data sets, such as energy efficiency data of different buildings, disease diagnostic data, etc., that running a classification algorithm on a computer element, such as a Personal Computer takes 0.017–0.029 s, compared to a Rasberry Pi model B integrated IoT device with 4.99–5.72 s run-time [48]. In their work [48], Yazici et al. considered the average time in their speed measurements after running each algorithm 20 times with each data set to deliver reliable speed measurements. This approach for the speed measurement can also be applied in the future real-life experiment of the proposed approach in this publication. Based on these previous observa-

tions, in order to achieve high accuracy in a real-life scenario of the proposed approach in this paper, the B_d should contain at least 100 instances with representative data sets. The model training for the classification based on SVM (with c hyperparameter between 0.1 and 1.0) and decision tree (with default hyperparameters) can be executed [32] after data pre-cleansing and -sampling on the *DSPN* from [45], which will cost about under 75 s. run time according to the previous reference example, such as [48]. Additionally, in the real-life experiment of the proposed approach, the SVM is considered in UCCP 1, carried out during the registration process. Its results will be presented later in the UCCP 2 during the aggregation period. In this way, sufficient time is available for the SVM to perform the elaborate training and produce the input for the UCCP 2 (I_6: Dependencies & Weight Coefficient). Furthermore, in a real-life experiment, the data cleansing and sampling methods are considered in the *DSPN* server element. They are already applied after the registration process before supplying users with data sharing recommendations in UCCP 2. In the real-world experiment, it could be possible to implement the *DSPN* server element in a smart home hub, such as Almond+, or Google OnHub, which ensures communication between the smart objects and allow their easy operations via a (mobile or web) application for smart home owners [5]. However, those existing smart home hubs must be expanded in terms of (privacy-preserving) functionality and hardware (non-functional requirements) in order to implement the modified user-centric privacy-preserving solution from Sect. 2.2. To sum up, implementing data pre-cleansing, -sampling, and the algorithms decision tree and SVM in the *DSPN* server element enables the proposed approach to address the performance issues and deliver the results, such as data sharing recommendations, to users within 250 ms on the corresponding user interface of the proposed privacy-preserving solution.

Maintainability: According to [30], maintainability includes the ease of customizing and modifying software to fix bugs, improve performance and make it adaptable. For the proposed modified user-centric privacy-preserving approach in this paper, no maintenance is needed after the initial installation of the approach. The updates for modification and customization are automatically installed during the nighttime not to influence the users and the use of the framework system.

Legal: The proposed automated approach with its algorithm is integrated into users' smart home environment at the *DSPN* in the IoT architecture system, as recommended in [45]. This implementation allows collecting and processing the collected data in a privacy-preserving way in users' environments according to the GDPR requirements from [12,38].

Portability: Users can have encrypted access to the data collected and processed in the *DSPN* of their smart environment via a mobile application. The main focus of the user-centric privacy-preserving approach is to supply users solutions with a minimum number of external accesses to allow control of the entire data collection, storage, and disclosure process [38,45]. According to Art.

20 of the GDPR [12], the proposed approach allows users to have an overview of the shared data with data consumers for different usage purposes, and it also allows users to supply additional data consumers with the already shared data.

Deployability: The proposed approach can be implemented using the Java-based open-source tool called WEKA 3[7]. This tool is recommended because it is portable and allows the integration of the framework with other Java-based user interfaces [2]. In addition, Java is also adaptable to different operating systems and devices [2]. Using WEKA 3 will allow the implementation of the entire proposed solution from Sect. 2.2 in the future.

Interoperability: According to [44], interoperability means that different software components can interact and cooperate despite different languages, user interfaces and platforms. The proposed approach is designed to integrate all types of smart home objects into the proposed user-centric privacy-preserving solution with the presented framework and to allow users control over the entire data collection, storage and disclosure process in their smart home environment with different smart home objects.

Data Integrity: The proposed framework and its integration into the existing framework from [45] allow the consideration and integration of all the collected data by users' smart home objects in their smart home environments. By integrating a server element with sufficient memory management in the *DSPN*, there will be hardly any technical limitations regarding the amount of data which can be stored.

Efficiency and Capacity: The proposed framework works in a less time-, cost- and energy-consuming and effective way because the proposed approach's machine learning algorithms are integrated in the server-based *DSPN* [48]. Additionally, considering the machine learning algorithms, decision tree and SVM in the proposed framework, integrated in the *DSPN*, help efficient data classification with high accuracy [43,48] in the IoT context. Furthermore, the consideration of the data cleansing and sampling methods in *DSPN* supports less time and efficient data processing [8,36].

Scalability: According to [10], scalability means when a system can perform tasks under growing work volumes and allows its enlargement. Integrating an efficient server element in the *DSPN* and the efficient and energy-saving integration of the machine learning algorithms allows the proposed framework to perform the corresponding tasks, such as supplying users with data sharing recommendations, in a growing environment of collected data. In case the initial supplied server capacity is insufficient, the server capacity of the *DSPN* can also be expanded accordingly by upgrading its capacity based on the supplied services of the corresponding provider.

Availability: According to [18], there are different levels of availability when it comes to the server and computer systems. The proposed framework recommends

[7] WEKA 3 is considered a very highly ranked top detection tool and data mining tool [35].

integrating at least a high-availability server element or better in the *DSPN*. Integrating such a system in the proposed framework could cause at most 5 min of service interruption per year [18]. If the integrated server element in *DSPN* is down, then it is conceivable to enable data access to the *DSPN* data on a private cloud, which is also installed in one's smart home environment and can be accessed by the smart home owner. This private cloud provides a backup of the *DSPN*.

To sum up, the evaluation outlines that considering the abovementioned requirements in the proposed framework allows fulfilling non-functional requirements from security-, privacy-, and performance-related categories.

3.2 Evaluation of Functional Requirements

This section evaluates the proposed solution with existing approaches, [3,22,23,26], qualitatively in order to outline the added value of the proposed framework and to evaluate functional requirements. These works [3,22,23,26] are relevant works in this area and partially supply the basics for the proposed solution. While [22,23] propose standalone machine learning approaches allowing users to express their data sharing and data sensitivity preferences by labelling a sample of collected data, [26] presents a technique to present users derived conclusions based on collected data about users' daily routines and activities. Additionally, [3] presents an architecture allowing users to understand and control their smart home network. However, the detailed and qualitative evaluation of the previous solutions outlines that those approaches do not supply users with best-practice-based data sharing recommendations with minimum human interference while considering context-sensitive factors and users' preferences based on past activities. Furthermore, integrating the proposed framework in an existing user-centric privacy-preserving approach from [45] allows users to control the entire data collection, storage and disclosure process and integrate the proposed solution in existing IoT architecture systems.
The evaluation metrics are categorized into three clusters: (1) privacy-preserving features, (2) user-friendliness and (3) GDPR requirements for automated decision making (Art. 22). The metrics of each category are derived from [12,25,33,38,45]. Table 3 presents the results of the qualitative evaluation.

4 Discussion and Limitations

4.1 Discussion

The proposed framework allows addressing (1) existing gaps from [25] and (2) demands from GDPR, especially Art. 9, 12, 15, 17, 19 and 22 [1,12,38], in context-aware privacy-preference modelling research. The proposed best-practice-based framework for user-centric privacy-preserving approaches in the smart home context supplies users with data sharing recommendations while considering (1) context-sensitive factors and (2) users' preferences based on users'

Table 3. Qualitative evaluation of the proposed solution: Each metric is evaluated by using the following rating scale: ○ = no possibility; ◑ = partially possible; and ● = possible.

Evaluation category	Metrics	Proposed Model	Model 2 from [22]	Model 3 from [23]	Model 4 from [26]	Model 5 from [3]
Privacy-preserving features	Allowing privacy-preserving data storage [45]	●	◑	◑	○	◑
	Allowing privacy-preserving and data protection of the users [25,45]	●	◑	◑	◑	○
	Limiting data access by limiting data consumers and usage purposes [33]	●	●	○	○	○
	Supplying best-practice-based data sharing recommendations [25]	●	○	○	○	○
	Considering context-sensitive factors while deriving data sharing recommendations [25,45]	●	○	○	◑	○
	Adjusting sensitivity of the data types according to already disclosed data [25]	●	○	○	○	○
User-friendliness	Minimum human interference (user inputs) [25,45]	●	◑	◑	◑	◑
	Consideration of users' data sharing preferences based on past activities [25]	●	○	○	○	○
GDPR (Art. 22) requirements for automated decision making	Users have the opportunity to intervene the automated processing [12,38]	●	◑	◑	◑	◑
	Automated approach includes suitable measurements to safeguard users' rights and privacy [12,38]	●	●	◑	◑	◑
	All the data types are considered in the automated approach [12,38]	●	◑	●	●	◑

past activities. The proposed framework with automation options, which can be integrated within an existing privacy-preserving approach with user-centricity from [45], allows users to control the entire data collection and disclosure process with minimum human interference. It also allows the integration of the proposed solution in existing IoT architecture systems. Addressing the mentioned gap in [25] and therefore including supervised and active machine learning methods allow supplying users with best-practice-based data sharing recommendations

derived based on GDPR specifications according to Art. 4, 5, 9, 12, 15, 17 and 19 [12] while considering users' context-sensitive factors as well as past activities. Additionally, the related work from Sect. 5 with already proposed machine learning privacy-preserving solutions in this context do not include mechanisms which facilitate the process of preference specification based on users' past activities [13,25,29]. Furthermore, the analysis of previous works also outlines that other solutions must be introduced, which allow the presentation and control of the context-sensitive factors related to users' privacy preferences [25]. With the proposed approach in this paper, these research gaps are addressed, and it allows supplying data sharing recommendations based on GDPR-based best practices, which can be adjusted according to users' past activities and context-sensitive factors. In this way, the GDPR requirements in the context of user-centric privacy-preserving approaches and automated decision-making [12,25,38] are also addressed.

As mentioned, the included initial input B_d for the automated framework, A_f, is derived based on the GDPR specification in Art. 4, 5, 9, 12, 15, 17 and 19 [1,12,38]. Additionally, the users are asked to indicate the dependencies between the collected data types according to their perception. However, it must be investigated with user studies whether there is a user-friendly way to capture users' dependency perceptions in this respective context. Moreover, the initial input (B_d) must be validated and completed based on interviews with GDPR experts to cover all the cases. Furthermore, in the modified user-centric privacy-preserving approach, the users are supplied with different inputs, such as $srvData$, $dCon$, $dpurp$, in the integrated user-centric privacy-preserving approach. These inputs must be validated with users within a user study to find out in which way these inputs can be adjusted and whether those inputs are sufficient for their decision-making process.

4.2 Limitations

The findings of this paper are mainly based on a literature review and previously derived approach from [45]. The proposed approach must be validated with user studies, and an additional real-world experiment to (1) validate the initial input, (2) find out a user-friendly way for deriving users' perceptions regarding data type dependencies and (3) investigate its acceptance and applicability, which will be addressed in the near future.

5 Related Work

Existing works can be clustered into two categories: (1) technical solutions for disclosure behaviour prediction in different contexts and (2) machine learning solutions in IoT and smart environments.

In the first category, solutions are presented, which supply users with disclosure

recommendations based on predicting disclosure behaviour models. While [40] presents a machine learning mechanism to help users specify disclosure preferences in a location-sharing system, [16] proposes a privacy wizard in a social network context, which configures users' privacy settings based on machine learning mechanisms automatically after asking users different questions. Additionally, Knijnenburg and Jin outline in their work that users are willing to receive privacy recommendations by an assisted system and that the input for those recommendations will influence users' satisfaction positively [24]. In [47], Xie et al. present a prediction algorithm which allows users to configure privacy settings in location sharing context. In their work, Xie et al. also outline that the context and data consumers (audience) influence users' location privacy preferences, and the observations show that few users also share similar sharing preferences [47]. Moreover, Pallapa et al. present another context-aware privacy-preserving solution in the mobile context, which derive users' privacy preferences based on the interaction history between the users and apply those in new situations [34].

The second category includes machine learning solutions in IoT and smart environment contexts, which support users in the automatic configuration of privacy settings and supply users with recommendations for privacy settings. Several machine learning solutions are presented in the smart home context, such as [3,22,23,26]. While [22] presents a machine learning-based framework allowing users to express their data sharing preferences by labelling some collected data in their smart home environments while considering the data consumers, usage purposes and information granularity, Keshavarz and Anwar propose in another work an active machine learning approach helping users to classify between sensitive and non-sensitive data according to users' privacy preferences [23]. Also, in this approach, Keshavarz and Anwar ask users to label some amount of data as sensitive or non-sensitive so that the model can learn users' privacy concerns and apply it while labelling the rest of the collected data [23]. Furthermore, Aïvodji et al. present in their work [3] an architecture called IOTFLA for data security and privacy in smart home environments. This approach allows users to improve the efficiency of the smart home systems, the understanding and control over the smart home networks [3]. Additionally, Kounoudes et al. present in their work [26] a data inference technique which derives conclusions about users' routines and activities based on the collected data to present those conclusions to the users and uses those conclusions to improve smart object services.

In comparison to all the above-mentioned previous works, the contribution of the proposed approach to this body of literature is three-fold: Best-practice-based framework supplying users data disclosure recommendations while considering (1) context-sensitive factors, (2) users' preferences in data sharing recommendations after learning from users' past activities and (3) its integration in an exiting solution allow users to control the entire data storage, collection and disclosure process with minimum human interference. Furthermore, the entire proposed approach in this paper allows addressing a few existing gaps in context-aware privacy-preference modelling with user focus, as mentioned in [25] and the demands from GDPR, especially Art. 9, 12, 15, 17, 19 and 22 [1,12,38].

6 Conclusions and Future Work

In this paper, a best-practice-based automated framework is proposed, which (1) supplies users with data sharing recommendations based on GDPR-related best practices and (2) allows to consider users' past activities, privacy preferences and context-sensitive factors. Integrating the proposed approach in an existing user-centric privacy-preserving approach allow users to control the entire data collection, storage and disclosure process with minimum human interference. Furthermore, this integration also allows us to implement the proposed automated approach in existing IoT architecture systems.

In order to investigate the performance and user acceptance, we plan (1) to implement the proposed approach in a real-world smart home environment and (2) to conduct user studies in the future.

Acknowledgments. We thank the anonymous reviewers for their feedback, and special thanks to Lindrit Kqiku, Alexandr Railean, Patrick Kühtreiber and Alexander Richter for the exchange and feedback.

References

1. GDPR Art. 9 Processing of Special Categories of Personal Data. https://gdpr-info.eu/art-9-gdpr/. Accessed May 2022
2. Aher, S.B., Lobo, L.: Data mining in educational system using Weka. In: International Conference on Emerging Technology Trends (ICETT), vol. 3, pp. 20–25 (2011)
3. Aïvodji, U.M., Gambs, S., Martin, A.: IOTFLA : a secured and privacy-preserving smart home architecture implementing federated learning: a secured and privacy-preserving smart home architecture implementing federated learning. In: Proceedings of 2019 IEEE Security and Privacy Workshops (SPW), pp. 175–180 (2019)
4. Al-Ameen, M.N., Tamanna, T., Nandy, S., Ahsan, M.M., Chandra, P., Ahmed, S.I.: We Don't Give a Second Thought Before Providing our Information: Understanding Users' Perceptions of Information Collection by Apps in Urban Bangladesh, pp. 32–43 (2020)
5. Awasthi, A., Read, H.O., Xynos, K., Sutherland, I.: Welcome PWN: almond smart home hub forensics. Digit. Investig. **26**, 38-S46 (2018)
6. Balapour, A., Nikkhah, H.R., Sabherwal, R.: Mobile application security: role of perceived privacy as the predictor of security perceptions. Int. J. Inf. Manage. **52**, 102063 (2020)
7. Bennett, S.: Development of the PID controller. IEEE Control Syst. Mag. **13**(6), 58–62 (1993)
8. Bermingham, M.L., et al.: Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. **5**(1), 1–12 (2015)
9. Binkhonain, M., Zhao, L.: A review of machine learning algorithms for identification and classification of non-functional requirements. Expert Syst. Appl. X. **1**, 100001 (2019)
10. Bondi, A.B.: Characteristics of scalability and their impact on performance. In: Proceedings of the 2nd International Workshop on Software and Performance, pp. 195–203 (2000)

11. Carretero, J., García, J.D.: The internet of things: connecting the world. Personal Ubiquit. Comput. **18**(2), 445–447 (2014)
12. Consulting, I.: Art. 22 GDPR Automated Individual Decision-Making, Including Profiling.https://gdpr-info.eu/art-22-gdpr/. Accessed July 2022
13. Das, A., Degeling, M., Wang, X., Wang, J., Sadeh, N., Satyanarayanan, M.: Assisting users in a world full of cameras: a privacy-aware infrastructure for computer vision applications. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1387–1396 (2017)
14. De Choudhury, M., Lin, Y.R., Sundaram, H., Candan, K.S., Xie, L., Kelliher, A.: How does the data sampling strategy impact the discovery of information diffusion in social media? In: Fourth International AAAI Conference on Weblogs and Social Media (2010)
15. Dutta, S., Chukkapalli, S.S.L., Sulgekar, M., Krithivasan, S., Das, P.K., Joshi, A.: Context sensitive access control in smart home environments. In: IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), pp. 35–41 (2020)
16. Fang, L., LeFevre, K.: Privacy Wizards For Social Networking Sites. In: Proceedings of the 19th International Conference on World Wide Web, pp. 351–360 (2010)
17. Fietkiewicz, K., Ilhan, A.: Fitness tracking technologies: data privacy doesn't matter? The (Un)Concerns of users, former users, and non-users. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, pp. 1–10 (2020)
18. Gray, J., Siewiorek, D.P.: High-availability computer systems. Computer **24**(9), 39–48 (1991)
19. Guhr, N., Werth, O., Blacha, P.P.H., Breitner, M.H.: Privacy concerns in the smart home context. SN Appl. Sci. **2**(2), 1–12 (2020)
20. Jahan, N., Ghani, T., Rasheduzzaman, M., Marzan, Y., Ridoy, S.H., Khan, M.M.: Design and feasibility analysis of nsugt a machine learning-based mobile application for education. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0926–0929. IEEE (2021)
21. Jozani, M., Ayaburi, E., Ko, M., Choo, K.K.R.: Privacy concerns and benefits of engagement with social media-enabled apps: a privacy calculus perspective. Comput. Human Behav. **107**, 106–260 (2020)
22. Keshavarz, M., Anwar, M.: Towards improving privacy control for smart homes: a privacy decision framework. In: 2018 16th Annual Conference on Privacy, Security and Trust (PST), pp. 1–3 (2018)
23. Keshavarz, M., Anwar, M.: The automatic detection of sensitive data in smart homes. In: International Conference on Human-Computer Interaction, pp. 404–416 (2019)
24. Knijnenburg, B., Jin, H.: The persuasive effect of privacy recommendations for location sharing services. SSRN Electron. J. 2399725 (2013)
25. Kounoudes, A.D., Kapitsaki, G.M.: A mapping of IoT user-centric privacy preserving approaches to the GDPR. Internet Things **11**, 100179 (2020)
26. Kounoudes, A.D., Kapitsaki, G.M., Katakis, I., Milis, M.: User-centred privacy inference detection for smart home devices. In: 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), pp. 210–218 (2021)
27. Kremer, J., Steenstrup Pedersen, K., Igel, C.: Active learning with support vector machines. Wiley Interdiscipl. Rev. Data Mining Knowl. Disc. **4**(4), 313–326 (2014)

28. Kulyk, O., Reinheimer, B., Aldag, L., Mayer, P., Gerber, N., Volkamer, M.: Security and privacy awareness in smart environments-a cross-country investigation. In: International Conference on Financial Cryptography and Data Security, pp. 84–101 (2020)
29. Liu, B., et al.: Follow my recommendations: a personalized privacy assistant for mobile app permissions. In: Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), pp. 27–41 (2016)
30. Malhotra, R., Chug, A.: Software maintainability prediction using machine learning algorithms. Softw. Eng. Int. J. (SeiJ). **2**(2) (2012)
31. Milne, G., Pettinico, G., Hajjat, F., Markos, E.: Information sensitivity typology: mapping the degree and type of risk consumers perceive in personal data sharing. J. Consum. Affairs **51**(1), 133–161 (2016)
32. Mohammed, R., Rawashdeh, J., Abdullah, M.: Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 243–248. IEEE (2020)
33. Oetzel, M.C., Spiekermann, S.: A systematic methodology for privacy impact assessments: a design science approach. Eur. J. Inf. Syst. **23**(2), 126–150 (2014)
34. Pallapa, G., Das, S.K., Di Francesco, M., Aura, T.: Adaptive and context-aware privacy preservation exploiting user interactions in smart environments. Pervas. Mob. Comput. **12**, 232–243 (2014)
35. Peerspot: WEKA Review. https://www.peerspot.com/products/weka-reviews. Accessed July 2022
36. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
37. Raskin, J.: The Human Interface: New Directions for Designing Interactive Systems. Addison-Wesley Professional (2000)
38. Regulation (EU): 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L119/1, pp. 1–88 (2016)
39. Rumbold, J., Pierscionek, B.: What are data? A categorization of the data sensitivity spectrum. Big Data Res. **12**, 49–59 (2018)
40. Sadeh, N., et al.: Understanding and capturing people's privacy policies in a mobile social networking application. Pers. Ubiquit. Comput. **13**(6), 401–412 (2009)
41. Schomakers, E.M., Lidynia, C., Müllmann, D., Ziefle, M.: Internet users' perceptions of information sensitivity-insights from Germany. Int. J. Inf. Manage. **46**, 142–150 (2019)
42. Sheehan, K.B., Hoy, M.G.: Dimensions of privacy concern among online consumers. J. Publ. Policy Mark. **19**(1), 62–73 (2000)
43. Shen, M., Tang, X., Zhu, L., Du, X., Guizani, M.: Privacy-preserving support vector machine training over blockchain-based encrypted IoT data in smart cities. IEEE Internet Things J. **6**(5), 7702–7712 (2019)
44. Wegner, P.: Interoperability. ACM Comput. Surv. (CSUR) **28**(1), 285–287 (1996)
45. Wickramasinghe, C.I., Reinhardt, D.: A user-centric privacy-preserving approach to control data collection, storage, and disclosure in own smart home environments. In: International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 190–206 (2021)

46. Wu, H., Knijnenburg, B.P., Kobsa, A.: Improving the prediction of users' disclosure behavior by making them disclose more predictably? In: Symposium on Usable Privacy and Security (SOUPS) (2014)
47. Xie, J., Knijnenburg, B.P., Jin, H.: Location sharing privacy preference: analysis and personalized recommendation. In: Proceedings of the 19th international conference on Intelligent User Interfaces, pp. 189–198 (2014)
48. Yazici, M.T., Basurra, S., Gaber, M.M.: Edge machine learning: enabling smart internet of things applications. Big Data Cogn. Comput. **2**(3), 26 (2018)
49. Zeng, E., Mare, S., Roesner, F.: End user security and privacy concerns with smart homes. In: Proceedings of SOUPS 2013, Symposium on Usable Privacy and Security, pp. 65–80 (2017)
50. Zhou, W., Jia, Y., Peng, A., Zhang, Y., Liu, P.: The effect of IoT new features on security and privacy: new threats, existing solutions, and challenges yet to be solved. IEEE Internet Things J. **6**(2), 1606–1616 (2019)