IM2P-Medical: Towards Individual Management Privacy Preferences for the Medical Web Apps

Nguyen Ngoc Phien[1], Nguyen Thi Hoang Phuong[2], Khiem G. Huynh[3]
Khanh H. Vo[4], Phuc T. Nguyen[5], Khoa D. Tran[6], Bao Q. Tran[7]
Loc C. P. Van[8], Duy T. Q. Nguyen[9], Hieu M. Doan[10], Bang K. Le[11]
Trong D. P. Nguyen[12], Ngan T. K. Nguyen[13], Huong H. Luong[14], Duong Hon Minh[15]
Center for Applied Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam[1]
Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam[1]
FPT University, Can Tho City, Viet Nam[3,4,5,6,7,8,9,10,11,12,14]
FPT Polytechnic, Can Tho City, Viet Nam[13]
Faculty of Information Technology, Pham Van Dong University, Quang Ngai Province, Viet Nam[2]
Faculty of Pharmacy, Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam[15]

***Abstract*—With the advancement of technology, people are now able to monitor their health more efficiently. Mobile phones and smartwatches are equipped with sensors that can measure real-time changes in blood pressure, SPO2, and other attributes and public them to service providers via web applications (called web apps) for health improvement suggestions. Moreover, users can share the collected health data with other people, such as doctors, relatives, or friends. However, using technology in healthcare has raised the issue of privacy. Some health web apps, by default, intrusively gather and share data. Additionally, smartwatches may monitor people's health status 24/7. Therefore, users want to control how their health is processed (e.g., collected and shared). This can be cumbersome as they would have to configure each device manually. To address this problem, we have developed a privacy-preference prediction mechanism in the web apps called IM2P-Medical: towards Individual Management Privacy Preferences for the Medical web apps. To capture individual privacy preferences in the web apps, our model learns users' privacy behavior based on their responses in different medical scenarios. In practice, we exploited several machine learning algorithms: SVM, Gradient Boosting Classifier, Ada Boost Classifier, and Gradient Boosting Regressor. To prove the effectiveness of the proposed model, we set up several scenarios to measure the accuracy as well as the satisfaction level in the two participant groups (i.e., expert and normal users). One key point in this research's selection of participants is its focus on those living in developing countries, where privacy violation issues are not a common topic. The main contribution of our model is that it allows users to preserve their privacy without configuring privacy settings themselves.**

***Keywords*—Letter-of-credit; cash-on-delivery; blockchain; smart contract; NFT; ethereum; fantom; polygon; binance smart chain**

I. Introduction

Monitoring health with technological devices and web apps has gained great popularity in this day and age. In fact, the market value of health monitoring devices in 2019 was $25,78.56 million and is predicted to soar to $44,861.56 million in 2027[1]. The devices can track multiple values like blood pressure, heart rate, and sleep quality and send them to a web app for visualization. Based on the collected data (called Evidence-based disease management [1], [2]) from these sensors, several approaches could detect the corresponding diseases. Furthermore, these devices can monitor users' metrics everywhere due to their portability. Thanks to these gadgets, hospitals can better support their patients, and people can take care of their health more efficiently. However, due to the vast amount of data that can be collected, the use of health-monitoring devices and web apps faces doubt from those value data privacy [3]. Thus, there is a demand for a solution that manages how personal data is collected by health devices and web apps.

Conventionally, users can manually adjust their privacy preferences via web apps' or devices' settings. However, this can be cumbersome and may not be effective. On the other hand, an automated solution that can suggest security settings for a user based on his/her personality or privacy preference can bring better results [4]. In this paper, we introduce our privacy preference prediction solution. Our system learns a user's security perspective and makes suitable suggestions for changing privacy settings.

Due to some economic barriers in developing countries, their citizens lack healthcare services and institutions. Thus, the privacy issues of the medical data are ignored. Currently, there are no medical data protection standards like the developed countries (e.g., European countries - GDPR [2]) to protect the user privacy issues. To address this drawback, our model focuses on individual privacy preferences w.r.t medical data - especially, in developing countries. Our dataset explores the feedback from the developing countries' citizens, e.g., Asia, Africa, and Latin America.

Moreover, most current health systems are focused on protecting users' medical data [5], [6]. For example, Son et al. [7] emphasizes the importance of user privacy preferences that are placed alongside the system's privacy policy. This means that the requestor must satisfy both privacy policy and privacy preferences in order to be able to access the patient's medical data. Besides, Hoang et al. [8] provide a mechanism to handle

[1]https://www.alliedmarketresearch.com/patient-monitoring-devices-market

[2]The GDPR document is available at https://edps.europa.eu/data-protection/data-protection/glossary/d_en#data_minimization

conflicts between the privacy policy and privacy preferences where it depends on prioritizing patient treatment or reducing the risk of personal information leakage. In addition to the above studies, the systems that build smart contract models for medical facilities using Blockchain technology also take care of users' privacy preferences, for example, Nghia et al. [9], [10], [11]. In these studies, the patient role was given full discretion in sharing their data with stakeholders. Moreover, the priority of treatment is also exploited in the approach of [12], [13], where patients allow access to their personal data in case of an emergency. Also, in the medical-related system, several studies deployed the individual privacy preference based on blockchain technology (e.g., blood donate [14], [15] or IoT medical sensors [16] via the transmission messages protocol [17]).

To make this suggestion function work, we introduce IM2P-Medical: Towards individual management privacy preferences w.r.t medical data based on a Machine Learning system, i.e., built based on Semi-Supervised Learning. The reason for choosing the Semi-Supervised Learning method is to be able to reduce the amount of data required while preserving the prediction accuracy. The data for this system was gathered via a questionnaire. This questionnaire aims to learn respondents' perspectives or attitudes toward privacy. We distribute the questions to two types of people. The former type was people who had a background in IT and privacy, such as IT students (i.e., expert participants), while the latter group included various types of people called average users (i.e., normal participants) - see Sect. IV-B. who responded to the questionnaire on the Internet.

The key contributions of this paper are three-fold, including i) designing the individual privacy preferences architecture w.r.t medical data (i.e., IM2P-Medical - Sect. II); ii) balancing the user burden and accuracy based on the semi-supervised learning approach (i.e., Implementation - Sect. III); and iii) proving the effectiveness of the IM2P-Medical based on the two participant dataset.

The remainder of this paper is organized as follows. Sections II and III introduce the IM2P-Medical architecture and implementation (i.e., training strategy, algorithm, and questionaire). Section IV presents experimental results, whereas related work are discussed in Section V. Finally, Section VI concludes the paper.

II. IM2P-Medical Architecture

Fig. 1 shows the interaction process among the parties in the proposed model, including i) Users; ii) Personal data; iii) People; iv) Service Providers, and v) IM2P-Medical. The main roles and responsibilities of these parties in the system are presented as follows:

User (also known as the data owner): they have the right to make a decision whether to share their personal data by responding with consent (Yes) or disagree (No). The data in their possession includes normal data (i.e., easy to share) and personal data (i.e., medical data).

Personal data: the data that needs to be protected because they are highly identifiable. As a result, a malicious user can obtain other types of user data based on the exploitation of this data pool. This study focused on grouping personal data in the medical environment which can be exploited by sensors or smartphones (e.g., heart rate, SPO2, calories burned).

People: Can act as a user (in some specific cases). **People** represent other users (in the same or not the same system) who have a relationship (e.g., relative, friend,) with the owner of the data. This party can have more than one relationship with the data owner.

Service providers:the party provides the necessary medical services to users (e.g., health monitoring, disease diagnosis, online doctor)[3]. This target group provides a specific type of health care service; in return, the user must provide the data requested by the service provider. In a traditional environment, users have to provide virtually any type of data to service providers, ignoring privacy risks [18]. Previous studies have shown that applications collect more data than what they need for the supported services [19]. In this paper, any data manipulation request must be accompanied by a corresponding purpose to eliminate this drawback.

IM2P-Medical: this party automatically identifies privacy preferences for each individual. Specifically, this model identifies users' privacy behaviors based on their responses to service providers' access requests. Besides, Fig. 1 also depicts five main components of IM2P-Medical including: data types; relationship(s); context; access request(s); and purpose(s). The relationship between IM2P-Medical and the remaining parties is presented as follows:

- **Data types**: the types of data (e.g., location, heart rate, etc). In fact, the identification of personal data also depends on the user's sense of privacy. Each individual will make a different decision depending on many factors. It is not possible to define all possible possibilities in this study. So we're targeting the kind of medical data that are being exploited by the user's sensors, wearable devices, and smartphones. To ensure that our survey achieves its stated objectives (i.e., risks of personal information disclosure), we also emphasized in our survey that these medical data can easily be exploited through their device without any notification to the user.
- **Relationship(s)**: This group of attributes also greatly affects the issue of sharing personal data. A user can easily share their walking record (e.g., steps, distance traveled, start and end locations) for a day with his/her friends or personal training etc. Users will share or not share their data depending on each specific relationship.
- **Context**: depending on the specific context, users can (not) share their data regardless of the same data type and relationship. For example, in healthcare scenarios (hospitals, clinics), heart rate data can be shared with **People** as healthcare workers (e.g., doctors, nurses); however, same data types and relationships but different contexts (e.g., sports participation) - users can opt-out of sharing. To be able to capture each user's data

[3]The service provider can reserve several services, but we target the medical environment

sharing behavior, we exploit sub-attributes (i.e., data types; relationship(s); access request(s); purpose(s)) on the context-specific (see Sect. III-C for more details).

- **Access request(s)**: This component is closely associated with the service provider. In contrast to **People**, where users voluntarily share their data with a specific purpose (i.e., decided by themselves), the data retrieval process for service providers is the opposite. Specifically, the structure must include the party of the request for access (e.g., medical or fitness apps) and the corresponding purpose (discussed in the next section). Users judge between the benefits and risks of privacy to make a decision.
- **Purpose(s)**: One of the important pieces of information to decide whether users share their data or not is the purpose of access. In particular, a series of analyses have shown that requests for supporting the application's service will be accepted more than advertising purposes. To clarify this, we also emphasize the importance of access intent in our survey scenarios (see Sect. III-C for more details).

III. Implementation

A. Self-training

Self-training or "Self-learning" is the most basic of pseudo-labeling approaches [20]. They consist of a single supervised classifier that is iteratively trained on both labeled data and unlabeled data that has been pseudo-labeled in previous iterations of the algorithm. At the initial procedure, a supervised classifier is trained on only the labeled data. The outcome of the classifier is used to obtain predictions for the unlabeled data. Then, the most confident of these predictions is added to the labeled data set, and the supervised classifier is retrained on both the original labeled data and the newly obtained pseudo-labeled data. This procedure is typically iterated until no more unlabeled data remain.

Several applications and variations of self-training have been put forward. For instance, Rosenberg et al. [21] applied self-training to object detection problems, and showed improved performance over a state-of-the-art (at that time) object detection model. Dopido et al. [22] developed a self-training approach for hyperspectral image classification. They used domain knowledge to select a set of candidate unlabeled samples, and pseudo-labeled the most informative of these samples with the predictions made by the trained classifier.

B. Algorithm

Algorithm 1 applies self-training model to label the *app*s in $UApp$ dataset. Specifically, it applies the SVM algorithm (several supervised methods) to pseudo-label the apps in the unlabeled data set ($UApp$) (see line 4). For the other supervised learning algorithm, we do the same idea.

Algorithm 1 selfTraining($LApp$,$UApp$,$supAlg$)

1: **input**:training apps $LApp$, target apps $UApp$, list of supervised algorithms $supAlg$.
2: **output**: label for $UApp$.
3: **for** each $app_i \in Uapp$ **do**
4: $label_{app_i}$ = SVM(app_i);
5: $UApp$ - $\{app_i\}$;
6: $LApp \cup \{(app_i, label_{app_i})\}$;
7: **end for**

C. Questionnaire

To make accurate suggestions, effectively learning users' behaviours is of paramount importance. We have meticulously designed a questionnaire for the learning purpose. Our questionnaire focuses on observing how users will adjust their privacy preferences in multiple contexts. To be specific, from our questions, we expect to answer three queries:

1) Given a specific context, how will participants share different types (e.g., heart rate, SPO2, burned calories) of data with other people (e.g., friends, relatives, doctors)?
2) Given a specific context, how will participants share the data for a certain purpose (e.g., analysis, education, ads)?
3) Given a specific context, how will participants share data with service providers (e.g., medical apps, fitness apps)?

For each query, we develop an appropriate type of question. In each question, there are a number of parameters whose values can be activities, individuals, permissions, etc. Participants have three options, they may completely agree, completely disagree or partly agree with the statement.

In the following parts of this section, we will explain the given questions and introduce the list of parameters.

1) Sharing data with others in a specific context: In this type of question, we attempt to understand how people share personal data with others when they are doing certain activities. We classify possible activities into two groups: indoor and outdoor. An example of an outdoor activity question is: "You are playing sports. Do you want to share your location with your doctors?". In this example, we want to know if the user is willing to share their location with a doctor while playing sports.

The general structure of the questions is: "You are *an activity*. Do you want to share your *information* with *a person*". Three parameters are required: the activity, the information to share and the person to share with. We have prepared a set of possible parameters' values as given below (Table I to III).

TABLE I. The Samples of Activities

ID	Activity name
1	playing sport
2	relaxing
3	doing daily activities
4	at home
5	at work
6	having treatment at home
7	at the hospital
8	under an emergency

Questionnaire participants have three options to choose: Yes (without restriction), No or Yes (with restriction). If they

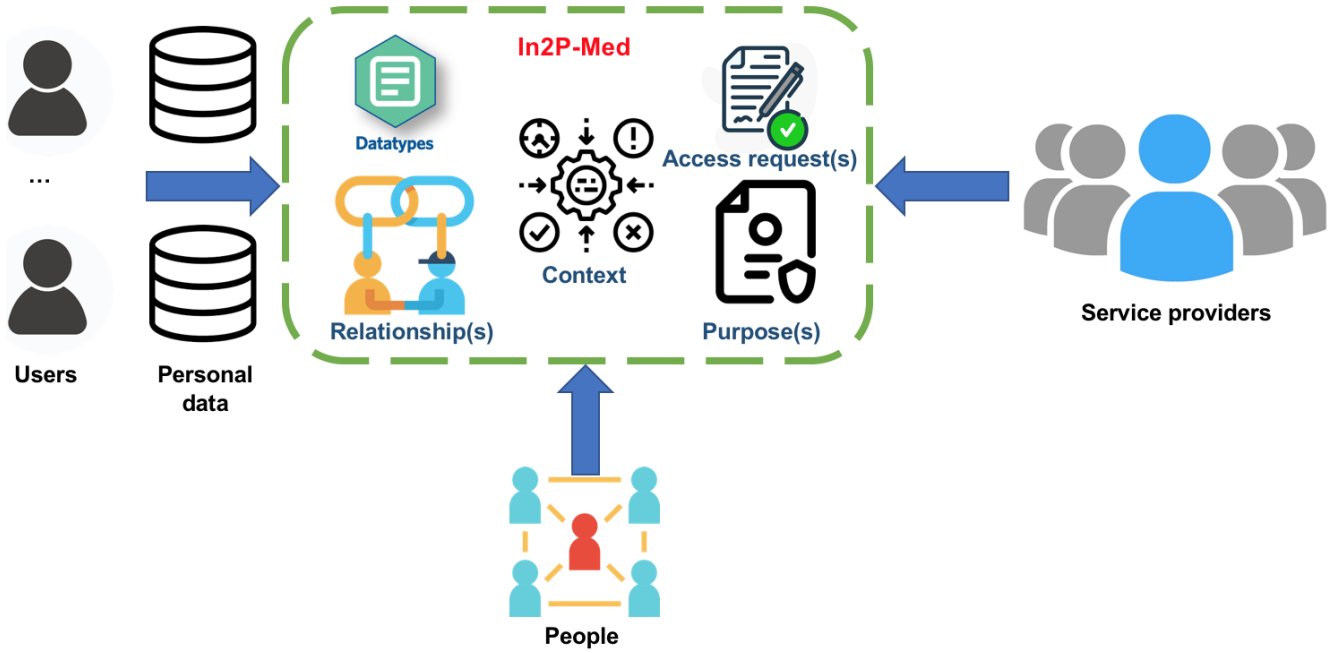

Fig. 1. IM2P-Medical architecture.

TABLE II. THE SAMPLE OF COLLECTIBLE INFORMATION

ID	Data item	ID	Data item	ID	Data item
1	name	7	steps	13	sugar level
2	phone number	8	heart rate	14	fat level
3	email	9	weight	15	travel distance
4	location	10	height	16	sleep
5	age	11	SPO2	17	health goal
6	gender	12	calories		

choose the first option, they are unconditionally willing to share data with an individual. However, if they choose to share with restrictions, they may want to implement some controls. For instance, they may share only once every 1 hour. Choosing "no" means that sharing under the given context is out of the question.

2) Granting permissions in a specific context: This type of question is intended to learn if users allows applications to access devices' components such as Bluetooth or Wifi.

The general structure of the question is: "You are *an activity*. Do you allow your device to access your *a component*?". There are two parameters required: the activity like in the first type of question and the device's components being granted access. We have prepared a set of possible parameters' values as given below (Table IV). Besides, we use the same set of answers in this type of question.

3) Sharing data with applications in a specific context: In this type of question, participants are surveyed if they allow applications in a specific category to collect data for a particular purpose. As an example, we may ask if they permit vendors making Health apps to collect usage information for marketing purposes.

The structure of the questions is: "Do you want your *data* to be collected by *an app category* for *the purpose* purposes ? (service provider x information collected)". Three parameters are required: the collected data, the app's category and the collecting purpose. We have prepared a set of possible parameters' values as given below (Table V).

We also use the same set of answers in this type of question (Table VI).

IV. EXPERIMENTAL RESULTS

A. Experiment Setting

In our tests, each participant had to take part in two phases i) they make their choice about whether to share medical data in each context different in the training period ii) they participate in the evaluation of the prediction results from our algorithms and give the satisfaction level of the corresponding algorithms. To achieve this goal, we have developed a web application to that requires interaction with participants through the two phases mentioned above. Specifically, participants label questions that share data during the learning phase (i.e. data set training data), then give their feedback on the labels generated by the models predict in the testing phase (i.e. test data set), and finally rate their satisfaction level on our predictive models.

More precisely, in the first phase, participants were asked to label each sentence (i.e. Yes (Y), No (N), or Maybe (M)) about sharing data in each term-specific scene as described in III-C. During the training phase, participants have to give answers to all 20 questions over a while. The minimum time is 10 minutes (an average of 30 seconds for an answer). After the labeling, the collected training dataset is built using algorithms learning is covered in Section III-B, specifically the -based approach to supervised learning and semi-supervised learning. To evaluate learning strategies, during the beta phase, the web app displays 20 new questions for those who participate.

TABLE III. THE ROLES OF REQUESTER

ID	Relationship	ID	Relationship	ID	Relationship	ID	Relationship
1	Friends	2	Family members	3	Doctors	4	Nurses

TABLE IV. PERMISSIONS

ID	Permission	ID	Permission	ID	Permission
1	Bluetooth	4	File Storage	7	Wifi
2	Camera	5	Microphone	8	Identity
3	Location	6	Sensors	9	Contacts

We randomly select prediction strategies instead of trying to define strategies according to the expected degree of accuracy. The main purpose is to remove all user prejudices about the algorithms in the back, which will be better than the previous algorithms. Specifically, four predictive models (five assessment questions/per model) are applied in this paper, including SVM, Gradient Boosting Classifier, Ada Boost Classifier, and Gradient Boosting Regressor.

For each new question in the experimental phase, the participants gave feedback on the labels, i.e. agree (Y) or disagree (N) - and in the case of disagreement, they must provide the correct label. For example, our forecast label is "Y", but their expected result is "M". They will give feedback on the label as no agree (N) and reselect the outcome they expected. In addition, in the 20 questions at a stage of the test phase, we reused four questions that appeared in the test phase corresponding to four predictive models. Specifically, each prediction model will have four sentences of new questions, and 1 question is randomly selected out of 20 questions during the experimental stage. The main purpose of this work is to divide into two groups of people based on their choices for that question for both periods, specifically, the selection group, the same selection, and the different selection group. The details of this comparison will be presented in section IV D 2. Finally, we collected the satisfaction level of the participants with the project guesses generated by each model. Participants can answer Yes (100%), No (0%), or Maybe (50%). The average time for answering each question at the test stage was 30 seconds. So each person participating in the survey must spend at least 20 minutes completing both learning and testing. To remove unsatisfactory answers, we have set the timer to track participants' time answering questions. If they spent less than the desired time (i.e. less than 30 seconds for a question), participants could not move on to the next question.

B. Participants

The primary purpose of this paper is to build an automatic medical data-sharing model that meets the privacy requirements of the users. We also want to explore the issue of sharing private data in developing countries where privacy is not widely aware of, especially in terms of sensitive data like medical. To achieve the above purposes, we conducted surveys. Our models are in countries in Asia, Africa, and Latin America. Besides that, there is a difference between the survey respondents on security and privacy, we categorized the differences between these two groups of users. Specifically, in the expert user group, we collected feedback from students as well as teachers who are studying and working at FPT University (Vietnam) majoring in Information Security at two campuses in Ho Chi Minh City. Ho Chi Minh City and Can Tho. For the normal user group, we used the tool Microworkers [4] to collect user feedback participants from developing countries.

1) Expert users: : For expert users (students and teachers of information security), we sent an email to the students who participated in the survey for four weeks (September 2021). There were a total of 20 qualified participants out of 32. The majority of participants were disqualified for not answering enough required questions. The average age of participants is 21.5, with the oldest and youngest being 29 and 18, respectively. Besides, about 15% of the participants were female (3/20).

2) Normal users: : The main purpose of this user group is to satisfy the requirement of diversity in terms of age, education, gender, and culture. We choose developing countries in two regional groups, including Latin America and Asia-Africa. We got 209 valid responses out of 296 participants. Each participant was paid $3. The number of participants belonging to the above two regional groups is 85 and 124, respectively. The mean age is 31.06 (minimum is 18 and oldest is 70) and 28.29 (the smallest age is 18 and the largest is 59). Out of a total of 85 responses, 42 were female (49.41%). The number for Asia-Africa is 25.6% (32 out of 125 participants).

C. Confusion Matrix

We used conventional measures to evaluate the accuracy of the proposed learning methods. Specifically, we exploited the 3X3 confusion matrix corresponding to the three labels (Y, N and M) (Table VII), where the columns represent the predicted labels (generated from approaches) and possible rows of values actual (participant opinion) and cells represent Error (E) or True Positive (TP). From the confusion matrix, we determined the evaluable metrics given in Table VIII.

D. Evaluation

In the evaluation, we performed a series of measurements to find the most appropriate algorithm in detecting the sharing behavior of personal medical data, given a specific context. Specifically, in the first test, we compared the accuracy obtained by different learning approaches (specifically between supervised learning and semi-supervised learning). We first compared the semi-supervised soft clustering method and the hard clustering techniques. This comparison aims to evaluate whether a semi-supervised system has good accuracy even with a reduced training set.

In the second test, we compared the accuracy of the proposed prediction models, namely SVM, Gradient Boosting Classifier, Ada Boost Classifier, and Gradient Boosting Regressor. As the results displayed in Section IV D 3, the semi-supervised-based approach gave better results than the supervised approach. Therefore, we apply the semi-supervised model to all four proposed algorithms.

[4]https://www.microworkers.com/

TABLE V. APP CATEGORIES

ID	Category	ID	Category	ID	Category
1	Communication	10	Libraries & Demo	19	Productivity
2	Dating	11	Lifestyle	20	Shopping
3	Education	12	Maps & Navigation	21	Social
4	Entertainment	13	Medical	22	Sports
5	Events	14	Music & Audio	23	Tools
6	Finance	15	News & Magazines	24	Travel & Local
7	Food & Drink	16	Parenting	25	Video Players & Editors
8	Health & Fitness	17	Personalization	26	Wear OS
9	House & Home	18	Photography	27	Weather

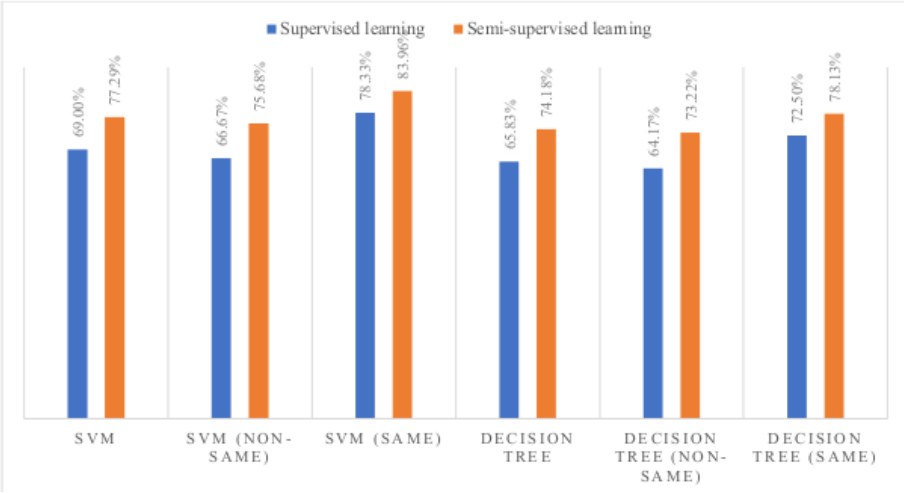

Fig. 2. The accuracy of SVM and Decision tree in supervised and semisupervised learning approaches.

TABLE VI. COLLECTING PURPOSES

ID	Purpose	ID	Purpose
1	Education	5	Scientific Research
2	Government	6	Treatment
3	Marketing/Advertising	7	Analytics
4	Product Development	8	Apps' functional

TABLE VII. CONFUSION MATRIX

	Predicted value: Y	Predicted value: N	Predicted value: M
Actual value: Y	TP_Y	$E_{Y,N}$	$E_{Y,M}$
Actual value: N	$E_{N,Y}$	TP_N	$E_{N,M}$
Actual value: M	$E_{M,Y}$	$E_{M,N}$	TP_M

1) Supervised learning and semi-supervised learning comparison: In this section, in addition to demonstrating which approaches (in particular supervised learning and semi-supervised learning) provide better prediction results with small data sets, we also wanted to test the difference between homogeneous and heterogeneous user groups in terms of data sharing decisions.

TABLE VIII. METRICS DEFINITION

$Accuracy$	$(TP_Y + TP_N + TP_M)/\#samples$
Pre_Y	$TP_Y/(TP_Y + E_{N,Y} + E_{M,Y})$
Pre_N	$TP_N/(TP_N + E_{Y,N} + E_{M,N})$
Pre_M	$TP_M/(TP_M + E_{Y,M} + E_{N,M})$
Re_Y	$TP_Y/(TP_Y + E_Y, N + E_{Y,M})$
Re_N	$TP_N/(TP_N + E_{N,Y} + E_{N,M})$
Re_M	$TP_M/(TP_M + E_{M,Y} + E_{M,N})$
$F1_X$	2* $(Pre_X * Re_X)/(Pre_X + Re_X)$, where $X \in \{Y, N, M\}$

To achieve the above goals, we first compare the accuracy between supervised learning and semi-supervised learning approaches by building a training set that is a subset of the original training set (with 10, instead of 20 questions). Specifically, in the new dataset, the number of questions in the shuttered train is ten and in the test set is 30 (including ten questions transferred from the train set). The main purpose for this reallocation of questions is that we wanted to aim for an approach that can balance the effort spent by the user to build the training set and the accuracy of the applied algorithm. A good approach that can satisfy the above criteria is the one that only has a small number of questions in the training set and ensures an acceptable accuracy. To meet the above requirements, we used a semi-supervised learning approach for the SVM algorithm, decision tree, and supervised learning for both SVM and Decision tree. We compared the accuracy of the SVM algorithm for both approaches as well as two different algorithms (SVM and decision tree) to get the most general view when choosing strategies to build predictive models.

Fig. 2 shows the accuracy between SVM and Decision tree on both approaches: semi- and supervised learning. It shows that the semi-supervised-based approach is always better than supervised learning for both SVM and Decision tree algorithms as well as groups of participants (same and non-same). SVM algorithm has higher accuracy than the Decision tree for all cases. Besides, the same answer group has the highest accuracy in each algorithm. This proves that the approach

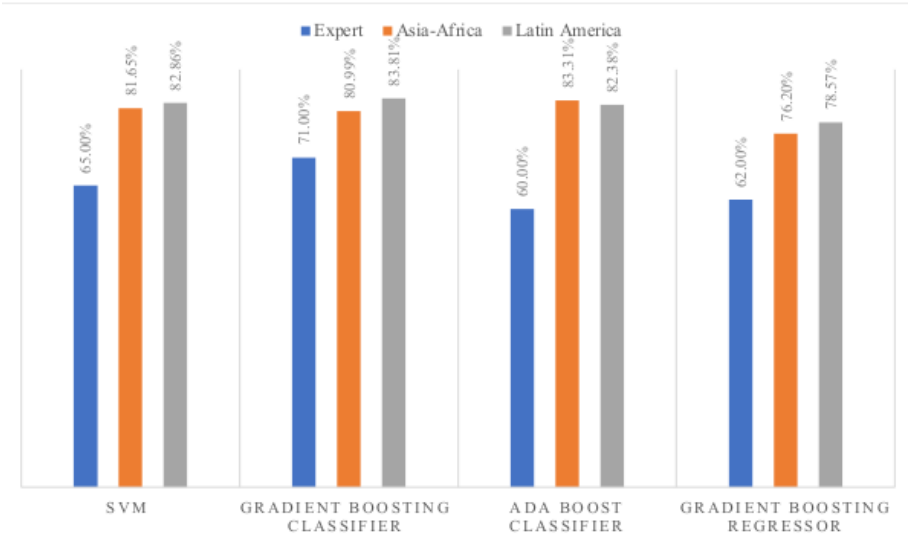

Fig. 3. Accuracy level of four classification models.

based on semi-supervised learning gives high accuracy even when trained on a small data set. In the following section, we apply a semi-supervised learning approach to delve into the analysis of accuracy, F1 score as well as the satisfaction of survey participants.

2) Answer prediction model accuracy: Fig. 3 depicts the accuracy of four predictive models for all three datasets (expert, Asia-Africa, and Latin America). The group with the lowest accuracy was experts from 60-71%, while the highest accuracy group was the group of participants from Asia-Africa and Latin America from 76.20% to 83.81%. This proves that the behavior of the user group is often easier to capture than the expert user group. Indeed, based on a manual analysis of users' comments on the reasons for their choice, we found that the difference between the two groups of users lies in the context of sharing personal data in the medical environment. In the case of ordinary users, they only care about the object to be shared or the type of data requested. Meanwhile, the experts evaluated all three groups of data, the shared audience, and especially their context. In particular, they are very careful when sharing high-risk data (motion data, location) with any individual (including relatives and friends). On the other hand, user groups are often more comfortable sharing personal data. They are willing to trade personal information to choose the services or utilities that the services or applications bring. Moreover, they trust the personal data protection mechanism of the service or application as well as the group of people who share it closely (relatives, friends).

3) Satisfaction level: This section evaluates the satisfaction level of four prediction algorithms for all three datasets (see Fig. 4). Specifically, all approaches have high satisfaction for all three datasets (from 90% to 98.81%). The algorithm with the highest satisfaction is the Gradient Boosting classifier (95% - 98.81%). Meanwhile, the algorithm with the lowest satisfaction level is SVM (95% - 98.81%). This experiment proves that the semi-supervised learning-based approach brings about a high level of satisfaction for all three groups of participants.

4) F1 score: Finally, we measured the score F_1 for each label (Y, N, M) in all three datasets (Expert, Asia-Africa, America Latin). The F_1 score considers both the precision and the recall aspect (see Table VII). There is a difference in approach four compared to the other approaches: this algorithm does not correctly predict any Yes or No answer options, but only predicts all possible answer options. This can be seen as a minus when applied to these models to identify user behavior in complex contexts.

Table IX shows the comparison of the four prediction models for the test dataset.

V. Related Work

Privacy problems have always been captivating researchers. To discover potential privacy infringement from browsing the Internet with a mobile phone, Collin Mulliner [23] tracked all HTTP headers sent to web services providers. From this activity, he could estimate the amount of covertly leaked personal information. Threats that come from unsecured applications have also been meticulously summarized by Jain et al. in their paper [24].

There have been multiple papers introducing various approaches to adjust privacy preferences dynamically. These proposed approaches do not only apply to applications but also to a wide range of other cases.

By introducing a Context-aware Privacy Policy Language (CPPL), Behrooz et al. [25] aimed to minimize the number of privacy policies that need analysis. In their work, the language filters policies that are relevant to the current scenario using context. The expectation of this research is to enhance the user experience of mobile users in general.

To cope with the issue of ever-changing contexts, Alom et al. [26] proposed a context-based privacy management system that utilizes machine learning algorithms. In their system, privacy preferences for a new context are automatically determined based on existing ones. To be specific, the authors

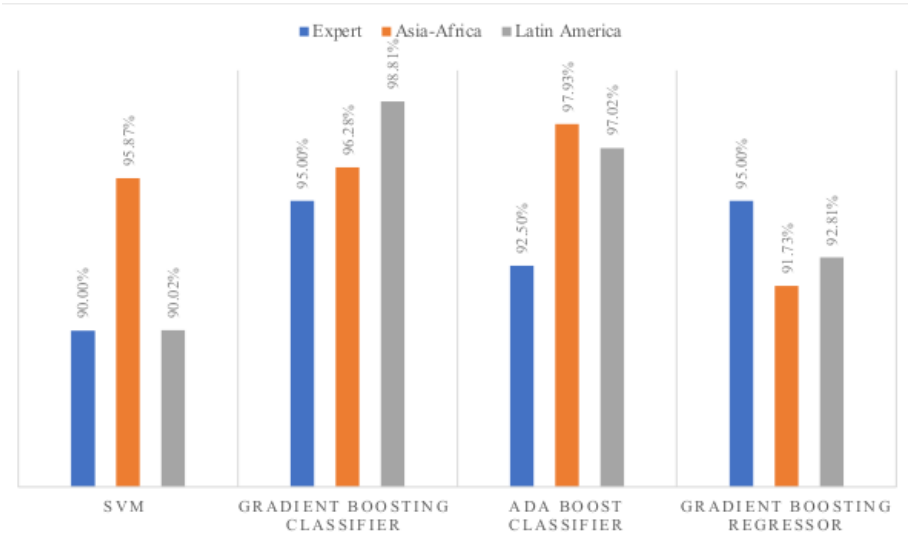

Fig. 4. Satisfaction level of four classification models.

TABLE IX. COMPARISON OF THE FOUR PREDICTION MODELS FOR THE TEST DATASET

		Approach 1			Approach 2			Approach 3			Approach 4		
		Y (%)	**N (%)**	**M (%)**	**Y (%)**	**N (%)**	**M (%)**	**Y (%)**	**N (%)**	**M (%)**	**Y (%)**	**N (%)**	**M (%)**
Expert-based	Precision	75.86%	55.10%	72.73%	88.46%	63.64%	66.67%	90.91%	41.18%	70.37%	NaN	62.99%	NaN
participants	Recall	62.86%	77.14%	53.33%	67.65%	75.68%	68.97%	60.61%	72.41%	50.00%	0.00%	100.00%	0.00%
(N=20)	F1	68.75%	64.29%	61.54%	76.67%	69.14%	67.80%	72.73%	52.50%	58.46%	NaN	76.54%	NaN
Crowd-based in Latin	Precision	85.00%	75.63%	87.65%	90.75%	73.17%	80.00%	86.29%	78.00%	75.00%	NaN	78.57%	NaN
American participants	Recall	92.12%	83.33%	65.14%	92.37%	83.33%	62.92%	95.54%	75.73%	58.06%	0.00%	100.00%	0.00%
(N=85)	F1	88.42%	79.30%	74.74%	91.56%	77.92%	70.44%	90.68%	76.85%	65.46%	NaN	88.00%	NaN
Crowd-based in Asia	Precision	89.07%	57.48%	84.47%	87.79%	62.81%	87.63%	90.15%	61.29%	83.53%	NaN	76.20%	NaN
-Africa participants	Recall	93.30%	71.57%	60.00%	93.50%	78.63%	56.29%	96.75%	72.38%	54.20%	0.00%	100.00%	0.00%
(N=124)	F1	91.13%	63.76%	70.16%	90.56%	69.83%	68.55%	93.33%	66.38%	65.74%	NaN	86.49%	NaN

build a classifier that can detect which users' preferences may be changed as well as the extent of that change. Therefore, given a new scenario, the classifier can predict users' choices. In their work, Lin et al. [27] also use machine learning to determine the most appropriate privacy preferences for the users. This is done after the system has created a collection of candidate configurations for a particular user.

Bahirat et al. [28] proposed a data-driven approach to designing privacy-setting profiles for IoT devices. Using scenario-based input, the system generates a collection of default privacy settings for the devices, and it is the responsibility of the users to pick one manually.

Knijnenburg et al. [29] introduced a system that supports privacy decisions by modeling privacy concerns. This approach is also known as user-tailor privacy, in which users are provided with private information and non-invasive controls. However, given the variance in people's perspectives on privacy, creating a general privacy model can be complicated.

There are also methods of adjusting privacy preferences specifically for smartphones. For instance, by taking contextual information into account, Yuan et al. [30] developed a machine learning-based privacy model for sharing photos. In their work, Sanchez et al. [31] developed a privacy preference recommendation system for personalized fitness apps. In their approach, the first profile users' traits and data permission preferences with machine learning clustering algorithms before designing privacy setting recommendation strategies. In their attempts to preserve users' privacy when using health applications, V. Koufi et al. [32] proposed an access control framework used in PHRManager. PHRManager is an Android app that gives authorized users access to Personal Health Records.

VI. CONCLUSION

This paper has introduced IM2P-Medical, a solution on how to learn users' privacy preferences and suggest appropriate settings for medical data (i.e., health-monitoring devices and apps scenarios). Specifically, semi-supervised learning can help understand people's perspectives while requiring fewer data to be explained in great detail. Moreover, a collection of questions for understanding users' thinking on privacy was also shown. The questions were then distributed to two types of participants (normal and expert).

At the end of the project, the satisfaction of users was gathered. Additionally, four models on how to minimize users' burdens were explained in the evaluation section. In this section, we also compare semi-supervised learning to prove the effectiveness of our model. The result indicates that semi-supervised learning can potentially conserve users' privacy.

The paper is the first attempt toward a user-centric model for healthcare systems, so it is extremely urgent to identify future development directions. Specifically, we plan to analyze user behavior to build a set of privacy settings recommen-

dations for new users to apply to the medical system. A blockchain-based solution is a potential option to validate service providers' claims about how much data mining is required. On the other hand, an extensive and in-depth study (e.g., increasing the number of participants, compared with users in developing countries) will also be launched soon.

ACKNOWLEDGEMENT

This work would not have been possible without the Mr. Le Thanh Tuan support in implementation and evaluation process. We also express our sincere gratitude to the students and crowd-workers who joined our survey.

REFERENCES

[1] H. H. Luong *et al.*, "Feature selection using correlation matrix on metagenomic data with pearson enhancing inflammatory bowel disease prediction," in *International Conference on Artificial Intelligence for Smart Community*. Springer, 2022, pp. 1073–1084.

[2] H. T. Nguyen *et al.*, "Enhancing inflammatory bowel disease diagnosis performance using chi-squared algorithm on metagenomic data," in *Intelligent Systems and Networks*. Springer, 2022, pp. 669–678.

[3] S. Lim, T. H. Oh, Y. B. Choi, and T. Lakshman, "Security issues on wireless body area network for remote healthcare monitoring," in *International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*. IEEE, 2010, pp. 327–332.

[4] H. X. Son *et al.*, "Priapp-install: Learning user privacy preferences on mobile apps' installation," in *Information Security Practice and Experience: 17th International Conference*. Springer, 2022, pp. 306–323.

[5] H. X. Son and E. Chen, "Towards a fine-grained access control mechanism for privacy protection and policy conflict resolution," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019.

[6] H. X. Son *et al.*, "Toward an privacy protection based on access control model in hybrid cloud for healthcare systems," in *12th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2019)*. Springer, 2019, pp. 77–86.

[7] H. X. Son and N. M. Hoang, "A novel attribute-based access control system for fine-grained privacy protection," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 76–80.

[8] N. M. Hoang and H. X. Son, "A dynamic solution for fine-grained policy conflict resolution," in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy*, 2019, pp. 116–120.

[9] N. Duong-Trung, H. X. Son, H. T. Le, and T. T. Phan, "Smart care: Integrating blockchain technology into the design of patient-centered healthcare systems," in *Proceedings of the 4th International Conference on Cryptography, Security and Privacy*, ser. ICCSP 2020, 2020, p. 105–109.

[10] ——, "On components of a patient-centered healthcare system using smart contract," in *Proceedings of International Conference on Cryptography, Security and Privacy*, 2020, p. 31–35.

[11] N. Duong-Trung, N. Quynh, T. Tang, and X. Ha, "Interpretation of machine learning models for medical diagnosis," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 5, pp. 469–477, 2020.

[12] H. X. Son, T. H. Le, N. T. T. Quynh, H. N. D. Huy, N. Duong-Trung, and H. H. Luong, "Toward a blockchain-based technology in dealing with emergencies in patient-centered healthcare systems," in *International Conference on Mobile, Secure, and Programmable Networking*. Springer, 2020, pp. 44–56.

[13] H. T. Le *et al.*, "Patient-chain: Patient-centered healthcare system a blockchain-based technology in dealing with emergencies," in *International Conference on Parallel and Distributed Computing: Applications and Technologies*. Springer, 2022, pp. 576–583.

[14] N. T. T. Quynh, H. X. Son, T. H. Le, H. N. D. Huy, K. H. Vo, H. H. Luong, K. N. H. Tuan, T. D. Anh, N. Duong-Trung *et al.*, "Toward a design of blood donation management by blockchain technologies," in *International Conference on Computational Science and Its Applications*. Springer, 2021, pp. 78–90.

[15] H. T. Le, T. T. L. Nguyen, T. A. Nguyen, X. S. Ha, and N. Duong-Trung, "Bloodchain: A blood donation network managed by blockchain technologies," *Network*, vol. 2, no. 1, pp. 21–35, 2022.

[16] L. N. T. Thanh, N. N. Phien, H. K. Vo, H. H. Luong, T. D. Anh, K. N. H. Tuan, H. X. Son *et al.*, "Ioht-mba: an internet of healthcare things (ioht) platform based on microservice and brokerless architecture," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[17] L. T. T. Nguyen *et al.*, "Bmdd: a novel approach for iot platform (broker-less and microservice architecture, decentralized identity, and dynamic transmission messages)," *PeerJ Computer Science*, vol. 8, p. e950, 2022.

[18] H. X. Son *et al.*, "A risk estimation mechanism for android apps based on hybrid analysis," *Data Science and Engineering*, vol. 7, no. 3, pp. 242–252, 2022.

[19] ——, "A risk assessment mechanism for android apps," in *International Conference on Smart Internet of Things*. IEEE, 2021, pp. 237–244.

[20] I. Triguero *et al.*, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information systems*, vol. 42, no. 2, pp. 245–284, 2015.

[21] C. Rosenberg *et al.*, "Semi-supervised self-training of object detection models," 2005.

[22] I. Dópido, J. Li, P. R. Marpu, A. Plaza, J. M. B. Dias, and J. A. Benediktsson, "Semisupervised self-learning for hyperspectral image classification," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 7, pp. 4032–4044, 2013.

[23] C. Mulliner, "Privacy leaks in mobile phone internet access," in *2010 14th International Conference on Intelligence in Next Generation Networks*. IEEE, 2010, pp. 1–6.

[24] A. K. Jain and othes, "Addressing security and privacy risks in mobile applications," *IT Professional*, vol. 14, no. 5, pp. 28–33, 2012.

[25] A. Behrooz and A. Devlic, "A context-aware privacy policy language for controlling access to context information of mobile users," in *International Conference on Security and Privacy in Mobile Information and Communication Systems*. Springer, 2011, pp. 25–39.

[26] M. Z. Alom *et al.*, "Helping users managing context-based privacy preferences," in *2019 IEEE International Conference on Services Computing (SCC)*. IEEE, 2019, pp. 100–107.

[27] J. Lin, B. Liu, N. Sadeh, and J. I. Hong, "Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings," in *10th Symposium On Usable Privacy and Security ({SOUPS} 2014)*, 2014, pp. 199–212.

[28] P. Bahirat, Y. He, A. Menon, and B. Knijnenburg, "A data-driven approach to developing iot privacy-setting interfaces," in *23rd International Conference on Intelligent User Interfaces*, 2018, pp. 165–176.

[29] B. P. Knijnenburg, "Privacy? i can't even! making a case for user-tailored privacy," *IEEE Security & Privacy*, vol. 15, no. 4, pp. 62–67, 2017.

[30] L. Yuan *et al.*, "Context-dependent privacy-aware photo sharing based on machine learning," in *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 2017, pp. 93–107.

[31] O. R. Sanchez, I. Torre, Y. He, and B. P. Knijnenburg, "A recommendation approach for user privacy preferences in the fitness domain," *User Modeling and User-Adapted Interaction*, pp. 1–53, 2019.

[32] V. Koufi *et al.*, "Privacy-preserving mobile access to personal health records through google's android," in *2014 MOBIHEALTH*. IEEE, 2014, pp. 347–347.