A Survey of Privacy Vulnerabilities of Mobile Device Sensors

PAULA DELGADO-SANTOS*, School of Engineering and Digital Arts, University of Kent, UK
GIUSEPPE STRAGAPEDE*, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain
RUBEN TOLOSANA, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain
RICHARD GUEST, School of Engineering and Digital Arts, University of Kent, UK
FARZIN DERAVI, School of Engineering and Digital Arts, University of Kent, UK
RUBEN VERA-RODRIGUEZ, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain

The number of mobile devices, such as smartphones and smartwatches, is relentlessly increasing to almost 6.8 billion by 2022, and along with it, the amount of personal and sensitive data captured by them. This survey overviews the state of the art of what personal and sensitive user attributes can be extracted from mobile device sensors, emphasising critical aspects such as demographics, health and body features, activity and behaviour recognition, etc. In addition, we review popular metrics in the literature to quantify the degree of privacy, and discuss powerful privacy methods to protect the sensitive data while preserving data utility for analysis. Finally, open research questions are presented for further advancements in the field.

Keywords: Background Sensors, Mobile Devices, Sensitive Data, Privacy, Data Protection

1 INTRODUCTION

Mobile devices such as smartphones, tablets, and wearables are provided with several sensors that are able to acquire a vast amount of personal information in different forms and for different purposes. This aspect, in combination with the significant advancements of the computational and communication capabilities of mobile devices over the last years, has shown the high potential of mobile devices in many application fields [69, 114, 165]. The large availability of personal data generated on mobile devices, in combination with their ubiquity (with 3.9 billion smartphones globally in 2016, estimated to rise to 6.8 billion by 2022 [81]) and their *always-on* nature has turned this technology into a potential source of major invasion of persona privacy.

The European Union has provided the General Data Protection Regulation (GDPR), defining personal data as any information related to an identified or identifiable natural person [2]. Moreover, the GDPR also defines sensitive data as a subset of personal information, that includes: *i)* personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs; *ii)* trade-union membership; *iii)* genetic data, biometric data processed solely to identify a human being; *iv)* health-related data; and *v)* data concerning a personâĂŹs sex

*These authors contributed equally to this research.

Authors' addresses: Paula Delgado-Santos, p.delgado-de-santos@kent.ac.uk, School of Engineering and Digital Arts, University of Kent, UK; Giuseppe Stragapede, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain, giuseppe.stragapede@uam.es; Ruben Tolosana, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain, ruben.tolosana@uam.es; Richard Guest, School of Engineering and Digital Arts, University of Kent, UK, r.m.guest@kent.ac.uk; Farzin Deravi, School of Engineering and Digital Arts, University of Kent, UK, f.deravi@kent.ac.uk; Ruben Vera-Rodriguez, Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Spain, ruben.vera@uam.es.

0360-0300/2022/1-ART1
https://doi.org/10.1145/3510579

life or sexual orientation [2]. Automated processing of user data, also known as user profiling [2], can easily reveal such attributes from data acquired through mobile user interaction by requesting irrelevant permissions, lax definition of permissions or misuse of permissions, combined with the aggregation of highly personalised data, reducing the privacy and security experience of the final users [26, 36]. Consequently, many works in the literature have focused on preventing potential misuse. This is the motivation of recent EU-funded, Innovative Training Networks (ITN) such as PriMa [5] and TReSPAsS [6].

In this context, we distinguish between privacy protection and sensitive data protection. They both aim to de-identify the user data and avoid re-identification [76] of direct identifiers, such as names, social security numbers, addresses, etc. [3], and indirect identifiers. The latter are not capable of identifying a particular individual but can be used in conjunction with other information to identify data subjects [58]. However, privacy protection refers to the security of the personal data and it borrows terminology, definitions and methods from cybersecurity, whereas sensitive data protection focuses on data modification techniques that account for the sensitive data while maximising the residual data utility for analysis. The idea of selective sensitive data protection was conceived with the development of the first large databases [23]. Early works in this field led to the concept of data sanitisation [34], as a database transformation before its release to a third party, and to the concept of Privacy Preserving Data Mining (PPDM) [24], as the development of models about aggregated data without access to precise information in individual data records. Furthermore, the term de-identification was coined to define the operation of Personal Identifiable Information[1] (PII) removal from data collected, used, archived, and shared by organisations [76].

Privacy is a multifaceted concept which has received a plethora of formulations and definitions [35, 43, 133, 174]. A profound discussion of the concept of privacy is, however, not in the scope of the present work. We will adopt the perspective of Article 21 of the GDPR, which states that the subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her. From this perspective, the main contributions of the present article are:

- An overview of the sensors and the raw data commonly available in modern mobile devices, paying special attention to the background sensors as they may be considered innocuous by the end users.
- A description of the typical application scenarios and purposes of collected data for mobile scenarios.
- An in-depth analysis of the personal and sensitive data extracted from mobile background sensors and the corresponding automated methods, focusing on: demographics, activity and behaviour, health parameters and body features, mood and emotion, location tracking, and keystroke logging.
- A summary of the metrics proposed in the literature for privacy quantification from the perspective of sensitive data, including also a review of the methods to achieve sensitive data protection.

For completeness, we would like to highlight other recent surveys in the field focusing on other privacy aspects. In [80], the authors focused on privacy protection in the context of authentication. In [78], a broad survey was presented about privacy leakage in mobile computing with especial interest in mobile applications, advertising libraries, and connectivity. A comprehensive overview is provided, but without a specific focus on the sensitive data. Finally, an analysis of privacy in the context of soft biometrics[2] is considered in [59], focusing on the extraction of demographic information such as gender from image and video data, not from mobile background sensors as in the present survey. Similarly, privacy was studied in the context of audio data in [73]. In contrast to previous work, we pay special attention to sensitive data and provide, to the best of our knowledge, the first survey that focuses directly on sensitive data and their privacy protection metrics and methods.

[1]Personal Identifiable Information (PII) is defined as information sufficient to distinguish or trace an individual's identity. This information may be used on its own or in conjunction with other information relating to an individual [103].

[2]*Soft biometrics* is defined as the characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals [89].

The rest of this survey is organised as follows. We first provide in Sec. 2 an overview of the sensors and the raw data commonly available in modern mobile devices. In Sec. 3 the typical application scenarios and purposes of collected data are described. Sensitive user information extraction is addressed in Sec. 4. In addition, the methods used to achieve the goal of extracting information are systematically discussed. Sec. 5 focuses on metrics whereas Sec. 6 focuses on methods for privacy protection techniques from the perspective of the sensitive data. In Sec. 7, the general conclusions of the present study are drawn and some open research questions that have emerged through the present survey are outlined for further investigation.

2 MOBILE ACQUISITION OF SENSITIVE DATA

Mobile devices offer a rich ground for data collection and processing. Smartphones, to begin with, are in fact distinguished from previous generation cellular phones by their stronger hardware capabilities (e.g., equipped with multi-core processors, GPUs, hardware acceleration units and gigabytes of memory) and powerful mobile operating systems, which facilitate wider sensing, software, internet, and multimedia functionalities, alongside core phone functions.

Mobile device built-in sensors, known as background sensors, are capable of providing frequent measures of physical quantities in an unobtrusive and transparent way. However, these data can be easily utilised to extract sensitive information of the user such as gender, age, emotion, ethnic group, etc.

This is also the case of other popular wearable devices such as smartwatches. Wearables might be considered under the broad definition of Internet of Things (IoT) devices since they are connected to the internet to collect and exchange data to perform automated decision making [61]. Their popularity among consumer electronics is rapidly increasing and they are progressively becoming capable of more specialised measurements and analyses [91]. In general, wearable manufacturers often provide users with mobile applications to install on their smartphones for communication and computing purposes, together with a more complete user interface. For example, smartwatches or fitness tracker bracelets can provide measurements of walked or run distances (based on data from motion sensors and Global Positioning System (GPS)) but also physiological parameters such as heart rate, Electrocardiogram (ECG), stress, sleep quality, etc.

Table 1 provides a description of the sensors and the raw data commonly available in modern mobile devices, grouped according to their sensing domain. In general, sensors can be classified into two categories based on the process adopted to produce the output signal: *i)* hardware sensors, on one hand, are physically installed components that perform a transduction of the physical quantity they measure to an electrical signal, which is converted into the digital domain for further processing; *ii)* software sensors, on the other hand, rely on data already made available by hardware sensor and/or calculate them to produce a measurement.

Motion sensors are responsible for measuring the acceleration and rotational forces in the three axes of the device. Hardware-based motion sensors will register continuous quantities as in the case of acceleration or angular velocity, whereas, when software-based, their output could be either continuous or event-driven as in the case of a step detector. Position sensors range from measuring changes in the Earth's magnetic field for orientation in space to proximity sensors, whereas environmental sensors are generally triggered by an event and return a single scalar value measurement. When designed to return continuous measurements, the sampling rate of these sensors can reach up to around 200Hz. Nevertheless, their power consumption is low [17].

Specific physiological/biological parameter measurements are also available on many mobile devices thanks to dedicated health sensors. For example, most smartphones and smartwatches include built-in optical sensors to capture changes in blood volume in the arteries under the skin, from which heart-related as well as polysomnographic parameters can be obtained [54, 161].

Touchscreen data can be in the form of keystrokes acquired from the keyboard [122], or in the form of touch data acquired throughout the user interaction [166]. In the former case, the virtual keys pressed are logged

Table 1. Description of the sensors and the raw data commonly available in modern mobile devices. BPM- Beats Per Minute, ECG- Electrocardiogram, SpO_2- Saturation of Peripheral Oxygen, GPS- Global Positioning System, SSID- Service Set IDentifier, RSSI- Receiver Signal Strength Indicator.

Sensor Type	Sensor / Data Source	Measured / Logged Quantity	Scope / Purpose	Sensor Type
Motion	Accelerometer / Linear Accelerometer	Acceleration Force	Device Traslation	Hardware
	Gyroscope	Angular Velocity	Device Rotation	Hardware
	Rotation Vector	Angle	Device Orientation	Hardware, Software
	Gravity	Magnitude of Gravity	Device Orientation	Hardware, Software
	Significant Motion	Change of user movement	Walking or Riding Vehicle	Software
	Step Counter	Number of Steps	Physical Activity Tracking	Software
	Step Detector	Step	Physical Activity Tracking	Software
Position	Geomagnetic Field	Earth's Magnetic Field	Device Orientation	Hardware
	Proximity	Distance	Device Distance from Surface	Hardware
	Magnetometer	Earth's Magnetic Field	Device Orientation	Hardware
	Geomagnetic Rotation Vector	Earth's Magnetic Field	Device Orientation	Hardware, Software
	Game Rotation Vector	Angle	Device Rotation	Hardware, Software
Environmental	Light	Illuminance	Screen Luminosity Regulation	Hardware
	Pressure	Ambient Pressure	Contextual Information	Hardware
	Temperature	Ambient Temperature	Contextual Information	Hardware
	Humidity	Ambient Humidity	Contextual information	Hardware
Health	BPM	Number of Beats	Physical Activity Monitoring	Hardware
	ECG	Sinus Rhythm Graph	Physical Activity Monitoring	Hardware
	SpO_2	Arterial Blood Oxygen Saturation Percentage Level	Physical Activity Monitoring	Hardware
	Blood Pressure	Systolic and Diastolic Average Pressure	Physical Activity Monitoring	Software
	Stress	Percentage based on Heart Beat Variability	Physical Activity Monitoring	Software
	Sleep / Wake Amount	Time	Physical Activity Monitoring	Hardware, Software
	Sleep Phase Transitions	Time	Physical Activity Monitoring	Hardware, Software
	Caloric Consumption	Step Counter	Physical Activity Monitoring	Software
Touchscreen	Keystroke	Keys Presses and Releases	Key Input	Hardware
	Touch Data	Screen Coordinates, Pressure of Touch	Complex Touch Gestures	Hardware
Network, Location and Application	Wi-Fi	SSID, RSSI, Encryption Protocol, Frequency, Channel	Connectivity	Hardware
	Bluetooth	SSID, RSSI, Encryption Protocol, Frequency, Channel	Connectivity	Hardware
	Cell Tower	ID	Connectivity	Hardware
	GPS	Latitude, Longitude, Altitude, Bearing, Accuracy	Navigation	Hardware
	App Usage	Name and Time of Used Apps	System Log	Software

together with pressure and timestamps, for each key press and release. From these raw data, it is possible to extract more complex features, such as the hold time, inter-press time, inter-release time, etc. [18]. In addition to keystroke, touchscreen panels significantly enlarged the input data space including touch data. In fact, it is possible to track the touch position in terms of X and Y coordinates in the screen reference system, but also pressure information and complex multi-touch gestures such as swipe, pinch, tap and scroll. Other complex features that can be extracted from touch data are velocity, acceleration, angle and trajectory [167].

Connectivity is yet another fundamental aspect of mobile devices. Their usefulness and ubiquity stem from the vast spectrum of functionalities they support thanks to many installed network protocols. Network connection data retains information about users' routine patterns, therefore it can be used for behavioural profiling and sensitive information extraction [108]. With the fifth-generation standard for cellular networks (5G) being commercialised and the sixth generation (6G) in development, significant improvement in terms of bit rates and latency will allow for extensive machine-to-machine communications, thus increasing the vast spectrum of functionalities already supported by mobile devices [60].

3 SENSOR APPLICATION SCENARIOS

In 2008, the two most common mobile operative systems, Android and iOS, had less than 500 apps available for download. To date, Android users are able to download over 2.87 million apps, followed by the Apple App Store with almost 1.96 million apps [7]. The possible application scenarios are wide ranging. Here we describe some popular application scenarios using mobile sensors.

3.1 User Authentication

In traditional authentication schemes, the legitimate user is expected to have knowledge of a secret such as a PIN code, a password or a pattern to gain access (authentication based on "what-you-know"), or an object, such as a card reader (authentication based on "what-you-have"), whereas recent authentication schemes largely deployed on mobile devices are based on the "what-you-are" paradigm: some traits of the user are acquired and processed in order to verify their identity [131]. With regard to mobile user authentication, a common approach is based on biometrics (both physiological and behavioural) [90], as in the case of entry-point fingerprint or face-based identification. A severe limitation of these processes consists in the fact that once the device is unlocked, as long as it remains active, an intruder would have unlimited time at their disposal. To provide prolonged protection, several studies have investigated and proved the feasibility of continuous authentication schemes for mobile devices based on behavioural biometrics [170]. In this case, biometric data would be continuously acquired in a passive way throughout normal device usage to constantly verify the user's traits. Different aspects such as modality, scenarios or environment, among others, can lead to alterations in the performance of mobile biometric systems [39]. Often combined, background sensors [10, 176], touchscreen [150], and network information [108] are among the most frequent modalities explored to develop behavioural biometric continuous authentication systems.

3.2 Healthcare and Fitness

Healthcare is a major field of study for mobile applications. The term "mHealth" was coined to indicate a sub-set of eHealth that includes medical and public health practice supported by mobile devices. Mobile apps help improve healthcare delivery processes and patients could benefit in terms of monitoring and treatment of diseases and chronic conditions, among many other healthcare purposes[130]. Examples of mobile apps include those that provide measurements of postures, report on mental disorders [77], and assess symptoms of conditions such as Parkinson disease, stress, dementia, etc. [71, 143]. Moreover, mobile health apps can be essential in sustaining a healthy lifestyle among people by monitoring and recommending behaviour corrections. From this perspective,

mobile devices such as smartwatches are largely used for fitness tracking. Physical exercise monitoring takes place by acquiring and processing background and GPS sensor data in a explicit and transparent way for the user [30, 31, 98].

3.3 Location-based Services

GPS and geolocation data are used by applications to present information related to the environment and the position of the users, for purposes such as targeted advertising, navigation, and recommendations [78]. These location-aware applications are under the context awareness paradigm [147]. Additionally, besides their native scope of communication, short range protocols such as Bluetooth and Wi-Fi allow mobile devices to exploit the information of nearby devices for purposes similar to the ones described. This concept can be particularly useful defining a semantic context of immediate surrounding, especially in the case of indoor environments. For example, in [113], the authors explored the feasibility of creating virtual tours in museums or expositions to deliver information about the items in the proximity of the users, who can receive this information on their mobile devices.

3.4 Other Applications

Traditionally, background sensors contribute to improving the mobile device user experience in several ways. For instance, position sensors are useful for recognising the orientation of the device in order to switch from portrait to landscape modality, and vice versa. Light sensor information about the illuminance is used to automatically adjust the screen brightness. The proximity sensor will lock the screen and activate a different speaker when the user is placing a call. Mobile device background sensors are also widely employed for Augmented Reality (AR) applications in several fields, such as education, entertainment, commerce, and navigation, among others [100]. AR-based apps heavily rely on the information provided by the background sensors to deliver information.

In addition, the sophisticated sensing capabilities of mobile devices, combined with their vast diffusion, have led to the idea of accomplishing large-scale sensing through them, known in literature as *mobile participatory sensing* [42]. Individuals with sensing and computing devices volunteer to collectively share data to measure and map phenomena of common interest, in a crowd-sourced fashion [78]. Applications where mobile participatory sensing has been used include noise pollution monitoring, litter monitoring, monitoring of traffic and road conditions, among others [117].

4 PRIVACY SENSITIVE DATA

The automated processing of user data acquired by mobile device sensors can reveal a significant amount of personal and sensitive information. In particular, while sensors such as cameras, GPS, or microphone are privacy-sensitive and require explicit user permission, many other sources such as accelerometer, touchscreen or network connection logs are less protected in terms of privacy. However, these data can also become crucial in obtaining private user information, since they can be processed to ascertain attributes that allow to re-identify a person, to extract demographic information or data related to their activity and health, among others.

Processing data from which it is possible to extract personal and sensitive information can lead to problems arising from the nature of these data. A common characteristic of sensitive data is in fact its uniqueness for each individual and its strict association to their owner. These implications are particularly relevant with regard to biometric data. In the biometric scenario, additional risk factors include: the modalities used to store personal data, the owner of the system, the used recognition modality (authentication or identification in a biometric database), the durability and class of the used traits, depending on which the severity of the consequences can vary [107]. An outline of the different sensitive aspects of mobile device users that can be extracted from the different mobile device sensors, with some of the most important work in each field, is shown in Table 2. In the

Table 2. Comparison of different state-of-the-art sensitive data acquisition approaches. *k*-NN- *k*-Nearest Neighbours, RF- Random Forest, SVM- Support Vector Machines, LSTM- Long-Short Term Memory, HMM- Hidden Markov Model, AUD- Active User Detection, DBSCAN- Density-based Spatial Clustering of Applications, CNN- Convolutional Neural Network, RNN- Recurrent Neural Network, DTW- Dynamic Time Warping, Acc- Accuracy, ERR-Equal Error Rate, AUROC- Area Under the Receiver Operating Characteristic, KL-Score- Kullback-Leibler Score

Sensitive Data	Sensors	Study	Classifier	Best Performance
Demographics	Motion	Jain and Kanhangad (2016) [12]	SVM	Acc. = 76.83%
		Davarci *et al.* (2017) [67]	*k*-NN	Acc. = 85.3%
		Nguyen *et al.* (2019) [129]	RF	Acc. = 96%
		Singh *et al.* (2019) [144]	4 Classifiers	Acc. = 80%
		Sabir *et al.* (2019) [145]	LSTM + Leave One Out	Acc. = 94.11%
		Ngo *et al.* (2019) [160]	HMM	ERR = 5.39%
		Meena and Saeawadekar (2020) [116]	Ensemble Boosted Tree	Acc. = 96.3%
	Touchscreen	Miguel-Hurtado *et al.* (2016) [119]	Decision Tree	Acc. = 78%
		Acien *et al.* (2019) [9]	AUD	Acc. = 97%
		Nguyen *et al.* (2019) [129]	RF	Acc. = 99%
		Jain and Kanhangad (2019) [88]	*k*-NN	Acc. = 93.65%
	Network, Location and Application	Riederer *et al.* (2015) [141]	Logistic Regresion	Acc. = 72%
		Neal and Woodard (2018) [127]	RF + Naïve Bayes	Acc. = 91.8%
		Wu *et al.* (2019) [180]	XGBoost	Acc. = 80%
Activity and Behaviour	Motion	Sun *et al.* (2010) [105]	SVM	Acc. = 93.2%
		Anjum and Ilyas (2013) [29]	Decision Tree	AUROC = 99%
		Thomaz *et al.* (2015) [164]	DBSCAN	Acc. = 76.1%
		Arnold *et al.* (2015) [33]	RF	Acc. = 70%
		Chang *et al.* (2018) [104]	*k*-NN	Acc. = 71%
	Network, Location and Application	Wan and Lin (2016) [177]	Fuzzy Classification	Acc. = 96%
		Chen *et al.* (2018) [55]	CNN	Acc. = 97.7%
		Ma *et al.* (2021) [182]	2D CNN + RNN	Acc. = 83%
Health Parameters and Body Features	Motion	Yao *et al.* (2020) [183]	CNN+ LSTM	Acc. = 94.8%
		Hussain *et al.* (2021) [84]	Naïve Bayes	Acc. = 71%
	Touchscreen	Arroyo-Gallego *et al.* (2017) [158]	SVM	AUROC = 88%
	Network, Location and Application	Palmius *et al.* (2016)[124]	Linear Regression	Acc. = 85%
Mood and Emotion	Motion	Quiroz *et al.* (2018) [138]	RF	AUROC >81%
		Neal and Canavan (2020) [159]	RF	F1-Score >95%
	Touchscreen	Gao *et al.* (2012) [74]	SVM	Acc. = 69%
		Shah *et al.* (2015) [153]	Lienar Regressin	Acc. 90.47%
	Network, Location and Application	Zhang *et al.* (2018) [188]	Factor Graph	Acc. = 62.9%
Location Tracking	Motion	Hua *et al.* (2017) [87]	Naïve Bayes + Decision Tree	Acc. = 92%
		Nguyen *et al.* (2019) [93]	DTW	KL-Score = 0.057%
	Network, Location and Application	Singh *et al.* (2018) [156]	RF	Acc. = 85.7%
Keystroke Logging and Text Inferring	Motion	Cain and Chen *et al.* (2011) [45]	Gaussian Distribution	Acc. = 70%
		Aviv *et al.* (2012) [11]	HMM	Acc. = 73%
		Owusu *et al.* (2012) [68]	Hierarchical Classifier	Acc. = 93%

remainder of this Section, examples of the personal and sensitive information extracted from the mobile device sensor data are presented, grouped in several categories depending on the nature of the extracted information and arranged by the particular data acquisition sensor.

4.1 Demographics

Probably the largest share of personal and sensitive information extracted from mobile user interaction data consists of attributes such as age, gender and ethnicity, which can all be ascribed to the category of demographics.

4.1.1 Motion Sensors. In [67] the user age range was extracted from the accelerometer data, while performing a task based on tapping on a predetermined series of different spots appearing on the device screen. The authors exploited the *k* Nearest Neighbours (*k*-NN) algorithm, obtaining an accuracy of 85.3%. Similarly, Nguyen *et al.* [129] developed a method to distinguish an adult from a child exploiting the behavioural differences captured by the motion sensors. Based on the hypothesis that children, with smaller hands, will tend to be shakier, they achieved an accuracy of 96% using the Random Forest (RF) method. In [12], the gender of the users was determined from their walking patterns acquired by smartphone motion sensors. The authors achieved an accuracy of 76.8% by processing with Support Vector Machines (SVM) and bagging algorithms. Meena and Saeawadekar [116] presented an approach for gender recognition based on the gait data extracted from smartphone sensors. The authors achieved an accuracy of 96.3% using the bagged tree classifier. The authors in [144] also focused on gender recognition from the data extracted by the accelerometer and gyroscope, obtaining an accuracy of 80% through Principal Component Analysis (PCA). Ngo *et al.* [160] focused on extracting gender and age with Hidden Markov Models (HMMs). The authors organised a competition based on accelerometer and gyroscope data acquired by wearable devices, which lead to a percentage error rate of 24.23% for gender and 5.39% for age. With the development of deep learning techniques, it has been possible to achieve enhanced results, as in the case of Sabir *et al.* [145], who obtained an accuracy of 94.11% analysing gait for gender classification by the means of Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN), a class of deep learning models particularly apt to capture temporal dependencies underlying in the data.

4.1.2 Touchscreen. In [9], the authors performed an analysis to identify whether the user using the device was a child or an adult based on swipe and tap gestures. For this purpose, an Active User Detection (AUD) algorithm has been used, achieving 97% accuracy. In [165], a new database of children's mobile interaction was presented. The authors used touch interaction information to classify children into three groups aged 18 months to 8 years old. The authors used a SVM algorithm achieving and accuracy of 90.45%. Nguyen *et al.* [129] also conducted a study using RF on tap gestures to distinguish between an adult and a child, achieving an accuracy of 99%. Touchscreen data has also been used to extract a person's gender. Miguel-Hurtado *et al.* in [119] focused their work on the prediction of soft-biometrics from swipe gesture data. They achieved 78% accuracy rate using a decision voting scheme from four classifiers: Decision Tree (DT), Naïve Bayes (NB), SVM and Logistic Regression (LR). In [88], behavioural data from a smartphone's accelerometer, gyroscope and orientation sensors were used while the user interacted with the device. The authors used gestural attributes in which the *k*-NN classifier recognises the gender of the user, providing a classification accuracy of 93.65%.

4.1.3 Network, Location and Application. Studies have shown a strong correlation between a user's geolocation and usage patterns and their demographics. For instance, in [27], authors highlight the value of mobile device data as a means of demographic modelling and measurement, without having to deal with the logistics of traditional censuses and surveys, which limit the speed for which policies can be designed and evaluated. In [186], based on three indicators of travel behaviour (radius, eccentricity, and entropy), the authors focus on understanding how usage of mobile phones correlates with individual travel behaviour exploring indicator correlations between mobile phone call frequencies evaluating factors such as age, gender, social temporal orders and characteristics of the built environment. Similarly, in [151], an unsupervised, data-driven approach is proposed to identify different user types based on high-resolution human movement data collected from a smartphone navigation app. In [141] the authors showed how demographic information can be inferred from geo-tagged photos on social networks. Specifically, they performed an analysis of how a person's ethnicity can be extracted from their location

patterns based on spatial segregation in two metropolitan areas. They distinguished between people belonging to three different ethnicity groups with an accuracy of 72% using LR. Also, Wu *et al.* [180] studied location data to obtain information on marital status and state of residence. They extracted spatio-temporal features from human mobility patterns and used them in conjunction with semantic features based on geographical context, which provided information about the places the subjects were visiting, such as residences, parks, hospitals, schools, and shopping malls. On this ground, they were able to achieve an accuracy of 80% based on an XGBoost algorithm. In [127], starting from gender-related behavioural patterns found in application, Bluetooth, and Wi-Fi, the authors were able to estimate the user gender with an accuracy of 91.8% using RF and multinomial NB. From the network connection logs, the total frequency of every event data record is computed. After sorting the events by frequency of occurrence, an evaluation of temporal patterns is carried out on the 1,000 most frequent events. Such contextual behavioural information is employed in a variety of user services, such as in personalising ads and customising home screens.

4.2 Activity and Behaviour

It has been shown that a broad variety of users' behaviour or activities can be inferred from mobile device sensor data [52].

4.2.1 Motion Sensors. In [105] the authors were able to detect whether the person was stationary, walking, running, bicycling, climbing stairs, going downstairs or driving using only the accelerometer information. Their proposed approach, based on SVM, was able to achieve an accuracy of 93.2%. Using accelerometer and gyroscope data, Anjum and Ilyas [29] developed an application to track the user activities, while the mobile device was kept in their hand, trouser pocket, breast pocket or handbag. Using a DT classifier, they achieved an average Area Under the Receiver Operating Characteristic (AUROC) curve of over 99.0%. In [164], the movements made by a user while eating were recognised by the accelerometer on a smartwatch. In [149], the authors, based on smartphone accelerometer data, classified drinking behaviour of young adults using nightlife physical motion. Density-based Spatial Clustering of Applications (DBSCAN) algorithm was used, achieving an accuracy of 76.1%. Even the amount of alcohol taken by users can also be extracted from the accelerometer data. In [33], the authors detected if a subject is sober, tipsy or drunk based on the accelerometer data and users' self-reporting of consumption. Their system achieved an accuracy of 70% using a RF algorithm. Motion sensors have also been used to extract information related to sleep such as sleep posture and habits. In [104] accelerometer, gyroscope and orientation data from a smartwatch was used to detect the sleep posture (supine, left lateral, right lateral, prone) achieving an accuracy over 95% with the Euclidean distance of the input values, and also to detect the hand position while sleeping (placed on the abdomen, chest or head) achieving an accuracy over 88% with k-NN algorithm.

4.2.2 Network, Location and Application. From GPS data, the authors in [177] determined whether the user was standing, walking, or using other transportation with a fuzzy classifier monitoring the speed and angle of the person obtaining a matching rate of 96% at a five-second interval. Also, Wi-Fi transmitters and receivers can reveal a significant amount of information about users' activity. In [55] the Wi-Fi Received Signal Strength Indicator (RSSI) was used on a smartphone to determine what activity users were doing, among lying down, falling, walking, running, sitting down, and standing up. For that purpose, different features of each activity were studied obtaining an accuracy rate of 97.7% with a Convolutional Neural Network (CNN). In [182], the authors used three neural networks on Channel State Information (CSI) measured by the Wi-Fi module, being able to discriminate whether a person is sitting, standing, walking with an accuracy of 83%.

4.3 Health Parameters and Body Features

4.3.1 Motion Sensors. The Body Mass Index (BMI) is a mathematical ratio that associates the mass and height of an individual. The usual way to calculate it is by using the parameters of a person's height and weight. In turn, human gait is based on the interaction between hundreds of muscles and joints in the body, and motion sensors can pick them up and translate them into characteristic patterns linked to the traits of the subjects, such as BMI. Yao *et al.* [183] used a hybrid model with a CNN-LSTM architecture to estimate the continuous BMI value from the accelerometer and the gyroscope data with a maximum accuracy of 94.8%. From the BMI many health attributes can be inferred [25, 63]. Another parameter that can be measured from the accelerometer is stress. Garcia-Ceja *et al.* [75] achieved 71% accuracy using similar user models and the Naïve Bayes algorithm.

4.3.2 Touchscreen. It is possible to identify whether a person has Parkinson's disease by analysing their keystroke writing pattern independently of the written content. In [158] the authors used a SVM algorithm achieving an AUROC of 0.88 on this problem. In [47], different types of features extracted from handwriting were studied as biometrics for Parkinson disease, achieving very promising results. In [37] the authors showed how people with longer thumbs perform swipe gestures in less time.

4.3.3 Network, Location and Application. In [124], the authors aimed at identifying periods of depression using geolocation patterns acquired from mobile phones of individuals with bipolar disorder (BD). While the subjects' depressive symptomatology was monitored through a weekly questionnaire, the authors used a linear regression algorithm and a quadratic discriminant analysis algorithm achieving an 85% accuracy. GPS can also determine sleep disorders, showing a good ability to detect sleep-wake stages and sleep-disordered breathing disorders (SRBD) such as Obstructive Sleep Apnea (OSA) with an accuracy up to 92.3% using SVM algorithms [15, 85]. StayActive[3] is an application that detects stress by analysing the behaviour of the users via smartphone, using the data from the Wi-Fi, step counter, location and battery level among others. In [132], the authors used a combination of simple relaxation scores based on the information extracted from the sleeping pattern of the users (largest time interval that the user did not touch his/her screen), their social interaction and their physical activity to determine the stress level.

4.4 Mood and Emotion

The user efficiency or motivation when performing a task changes in accordance to their mood. Thus, it can be inferred from different sensors.

4.4.1 Motion Sensors. In [159], Neal and Canavan studied how mood can have a significant impact on the recognition performance of a mobile biometric system. In their study, the authors observed that the subjects with the least accurate identification (<70%) were those with the least mood changes using a RF classifier. The walk pattern data obtained from a smartwatch accelerometer and gyroscope can be used to determine a person's mood (happy, sad or neutral). The authors in [138] determined the mood with a RF algorithm achieving a mean AUROC of 81%.

4.4.2 Touchscreen. Numerous studies have shown how, from the way a user interacts with the screen of his or her mobile device, it is possible to extract their mood. In [46], the authors investigated the manifestations of psychiatric diseases unobtrusively and in the setting of patients' daily lives, exploring the possible connections between bipolar affective disorder and mobile phone usage. Based on keystroke metadata and accelerometer data, they reported a 90.31% prediction accuracy on the depression score. In [83], in order to provide people with preventive treatments before subjects reach clinical depression, the authors exploited a mobile app to capture emotional states, by the means of call logs and usage of apps, with a predictive accuracy for negative emotions

[3]StayActive App: http://www.aal-europe.eu/projects/stayactive/

of around 86%. Gao *et al.* [74] demonstrated how finger-stroke features during gameplay could automatically discriminate between four emotional states (excited, relaxed, frustrated, bored). By means of an SVM algorithm they obtained an accuracy of 69%. In [153], finger strokes were studied. These strokes were assumed to be indirect indications to the user's emotional state. The authors predicted the emotional state of a person into one of the three states: positive, negative or neutral. They achieved an accuracy of 90.47% using a linear regression.

4.4.3 Network, Location and Application. MoodExplorer [4] is an app that collects data from mobile sensors such as GPS, accelerometer and Wi-Fi among others. From them, the authors in [188] demonstrated how self-reported emotional states have high correlation with smartphone usage patterns and sensing data. The authors recognised the composite emotions (happiness, sadness, anger, surprise, fear, disgust) of users through a proposed model called Graph Factor with a performance metric called exact match of 62.9% on average.

4.5 Location Tracking

Mobile devices usually come with built-in GPS modules for the purpose of location tracking. However, even when GPS coordinates are not available explicitly, position can be inferred by other sensors.

4.5.1 Motion Sensors. Several studies have shown how the position of a person can be inferred from the accelerometer, gyroscope and magnetometer while he or she is walking, driving or using public transport. In [93], the authors compared the pre-established routes with those taken by users while using different transport modes such as walking, train, bus or taxi. They compared both routes with a Dynamic Time Warping (DTW) algorithm obtaining a Kullback-Leibler distance of 0.00057 in the case of a taxi journey. In [87], it was demonstrated how, when a person uses the subway, it is possible to track them from the accelerometer data. They achieved an accuracy of 92% when the passenger travelled through 6 stations using boosted NB and DT algorithms. In [86], the authors were able to determine the location of an individual driving in a vehicle based solely on motion sensor measurements. The approach adopted was based on deriving first an approximate motion trajectory given acceleration measurements, then on correlating such trajectory with map information to infer the location. In this way, they were able to locate a device owner to within a 200-meter radius of the true location.

4.5.2 Network, Location and Application. From the different Wi-Fi networks to which a user connects, it is also possible to determine the position of an individual. In [156] the location was determined in real time in indoor places. The authors achieved an accuracy of 85.7% using a RF algorithm.

4.6 Keystroke Logging and Text Inferring

4.6.1 Motion Sensors. Touchlogger [45] was an application created to determine the region of the phone touchscreen touched by the user, based on the device micro-movements captured by the accelerometer and the gyroscope. The screen was divided into 10 regions and with the help of a probability density function for a Gaussian distribution an accuracy of 70% was obtained. Based on this result, it could be possible to identify the text that the user is writing. In this task, Owusu *et al.* [68] obtained an accuracy of 93% using a hierarchical classification scheme. Similarly, in [11], with a controlled environment, the authors were able to identify the PIN entered 43% of the times and the pattern 73% of the time by means of LR and HMM.

5 PRIVACY METRICS FOR SENSITIVE DATA

All privacy protection methods work by modifying the original data in order to deprive it of user sensitive information. For instance, the modified data should only reveal allowed attributes (e.g., gender) in order to maintain some data utility, in terms of available information, while other attributes (e.g., ethnicity) are suppressed.

[4]MoodExplorer App: https://play.google.com/store/apps/details?id=com.examsuniverse.moodexplorer

The degree of privacy achieved is typically related to the extent of data modification; however, the utility of the resulting dataset can be significantly impacted [76].

In order to evaluate the effectiveness of privacy protection approaches, the degree of privacy protection achieved, as well as the residual data utility after data modification, should be quantified. The former task can be achieved through specific privacy metrics, whereas the latter can be expressed in terms of reduction of traditional performance metrics such as accuracy or Equal Error Rate (EER).

User sensitive data acquired through mobile interaction is very heterogeneous and can be *structured*, as in the case of high-level health data, network, location and application data, or *unstructured*, i.e. motion, position, environmental, touchscreen and low-level health data. Consequently, different metrics are required depending on the specific application scenario. In this context, we will consider data after having undergone modifications in order to suppress or alter specific sensitive attributes, while retaining utility for analysis and extraction of non-sensitive information.

In our discussion, privacy metrics will be classified based on their output, in other words, depends on the characteristics of the data that are measured with a specific metric. There is no specific metric that can be applied to every characteristic, so many studies use their own metrics. Table 3 shows the metrics considered in our discussion and input data needed for the specific metric computation, grouped by the property measured. According to this criterion, some of the most relevant privacy metrics in the context of data acquired through mobile interaction can be grouped as follows [175]:

Anonymity-based metrics. these metrics stem from the idea of k-Anonymity [106], defined as the property of a dataset ensuring that in case of release, based on an individual's disclosed information, it is not possible to distinguish than individual from at least $k-1$ individuals whose information has also been disclosed. This is achieved by grouping subject data into equivalence classes with at least k individuals, indistinguishable with respect to their sensitive attributes. k-Anonymity is independent of the information extraction technique and it quantifies the degree of privacy exclusively considering the disclosed data. It is useful to express the degree of similarity between datasets, namely the original one and the sanitised one, or it can be applied to samples within a single dataset. However, several studies have reported some limitations of k-anonymity, which have led to the development of new metrics based on the original, aiming to overcome some of its issues by imposing additional requirements. For instance, m-invariance [181] modifies k-anonymity to allow for multiple, different releases of the same dataset. (α,k)-Anonymity [179] imposes a predetermined maximum occurrence frequency for sensitive attributes within a class to protect against attribute disclosure. ℓ-diversity [13] was developed to prevent linkage attacks by specifying the minimum diversity within an equivalence class of sensitive information, namely at least ℓ well-represented different sensitive values. For a skewed distribution of sensitive attributes, t-closeness [123] and stochastic t-closeness [64] were introduced, starting from the idea that the distribution of sensitive values in any equivalence class must be close to their distribution in the entire dataset. Consequently, knowledge of the original distribution is needed to compute this metric. Similarly, starting from the original data distribution (c,t)-isolation [51] indicates the number of data samples present in the proximity of a sample predicted from the transformed data. Depending on the semantic distance between sensitive user records, such as in the case of numerical values, (k,e)-anonymity [187] requires the range of sensitive attributes in any equivalence class to be greater than a predetermined safe value. Despite the highlighted shortcomings, k-anonymity and the derived metrics are still largely employed today in a broad variety of different privacy contexts, but mainly for low-dimensional structured data [21]. It has in fact been shown that k-anonymity-based properties do not guarantee a high degree of protection in case of high-dimensional data.

Differential Privacy-based metrics. differential privacy is a definition that has become popular thanks to its strong privacy statement according to which the data subject will not be affected, adversely or otherwise, by allowing their data to be used in any study or analysis, no matter what other studies, datasets, or information

Table 3. Some of the most common privacy metrics grouped by the property measured. ADE - Adversary's Estimate: generally a posterior probability distribution. ADR - Adversary's Resources: computational power, time, etc. PAR - Parameters: for configuring privacy metrics. PK - Prior Knowledge: generally a prior probability distribution. TO - True Outcome: also known as ground truth, it can be used to evaluate the ADE.

Property	Metric	Input Data
Anonymity	k-Anonymity [106]	PAR
	m-Invariance [181]	PAR
	(α, k)-Anonymity [179]	PAR
	ℓ-Diversity [13]	PAR
	t-Closeness [123]	PAR, TO
	Stochastic t-closeness [64]	PAR, TO
	(c, t)-Isolation [51]	ADE, PAR, TO
	(k, e)-Anonymity [187]	PAR
Differential Privacy	$(d\text{-}\chi)$-Privacy [49]	PAR, TO
	Joint Differential Privacy [97]	PAR, TO
	Geo-indistinguishability [28]	PAR, TO
	Computational Differential Privacy [120]	ADE, ADR, PAR, TO
	Information Privacy [65]	ADE, PAR
Entropy	Entropy [154]	ADE
	Cross-Entropy [118]	ADE, TO
	Cumulative Entropy [92]	ADE
	Inherent Privacy [22]	ADE, TO
	Mutual Information [110]	ADE, TO
	Conditional Privacy Loss [22]	ADE, TO
Success Probability	Privacy Breach [70]	ADE, TO
	$(d\text{-}\gamma)$-Privacy [139]	ADE, TO
	(δ)-Presence [128]	ADE, TO
	Hiding Failure [8]	ADE, TO
Error	Euclidean Distance [155]	ADE, TO
Accuracy	Confidence Interval Width [24]	ADE, PAR
	$(t\text{-}\delta)$-Privacy Violation [95]	ADE, PAR, PK, TO
	Size of Uncertainty Region [56]	ADE
	Customisable Accuracy [32]	PAR
Time	Maximum Tracking Time [148]	ADE
	Mean Time to Confusion [82]	ADE, PAR

sources, are available [43]. As discussed in Sec. 6, differential privacy is generally achieved by adding noise to the original data. Therefore, in order to quantify differential privacy as a property of the data indicating the degree of privacy, it is a requirement to have knowledge of the original data. Differential privacy was defined in the context of databases to achieve indistinguishability between query outcomes, but thanks to its generality it has found application in different contexts for low-dimensional data, including biometrics and machine learning systems. It is in fact based on the requirement that independently of the presence of a particular data subject, the probability of the occurrence of any particular sequence of responses to queries is provided by a parameter, ϵ, which can be chosen after balancing the privacy-accuracy trade-off inherent to the system. For a given computational task and a given value of ϵ, there can be several differentially private algorithms, which might have different accuracy performances. As in the case of k-anonymity, many metrics were originated from the initial definition of differential privacy, including approximate differential privacy, which has less strict privacy guarantees but is

able to retain a higher utility [66]. d-χ-Privacy [49] allows different measures for the distance between datasets than the Hamming distance used in the definition of differential privacy. Joint differential privacy [97] applies to systems where a data subject can be granted access to their own private data but not to others'. In the context of location privacy, geo-indistinguishability [28] is achieved by adding differential privacy-compliant noise to a geographical location within a determined distance. In contrast to previously described metrics based on differential privacy, computational differential privacy [120] adopts a weaker adversary model, favouring accuracy. In order to adopt computational differential privacy, it is necessary to have knowledge of the posterior data distribution reconstructed from the transformed data. Similarly, information privacy [65] is met if the probability distribution of inferring sensitive data does not change due to any query output.

Entropy-based metrics. in the field of information theory, entropy describes the degree of uncertainty associated to the outcome of a random variable [154]. Metrics based on entropy are generally computed from the estimated distribution of real data obtained from the sanitised data, even though additional information can be needed for a particular metric, such as the original data or some of the data transformation parameters. When attempting to estimate sensitive information from protected user data, high uncertainty generally correlates with high privacy. Nonetheless, a correct guess based on uncertain information can still occur. In [118], the degree of privacy protection is quantified by cross-entropy (also referred to as likelihood) of the estimated and the true data distribution in the case of clustered data derived from the original data. A cumulative formulation of entropy was defined in [92] in the context of location privacy to measure how much entropy can be gathered on a route through a series of independent zones. Inherent privacy [22] represents another example of metric derived from the definition of entropy, considering the number of possible different outcomes given a number of binary guesses. Mutual information and conditional privacy loss [22, 110] are also metrics based on entropy. The former provides a measure of the quantity of information common to two random variables and it can be computed as the difference between entropy and conditional entropy, also known as equivocation, which is useful to compute the amount of information needed to describe a random variable, assuming knowledge of another variable belonging to the same dataset. The latter property is built on similar premises, but it considers the ratio between true data distribution and the amount of information provided by another variable revealed.

Success Probability-based metrics. metrics in this category do not take into account properties of the data but only the outcome of sensitive information extraction attempts, as low success probabilities indicate high privacy. However, even if this trend is observable considering the entire dataset, single users' private data could still be compromised. In [70], based on the original and estimated data, a privacy breach is defined as the event of the reconstructed probability of an attribute, given its true probability, being higher than a fixed threshold, whereas in [139], this idea was extended by (d,γ)-privacy, in which additional bounds are introduced for the ratio between the true and reconstructed probabilities. In contrast, δ-presence [128] evaluates the probability of inferring that an individual is part of some published data, assuming that an external database containing all individuals in the published data is available. Hiding Failure (HF) [8] is a data similarity metric used to detect sensitive patterns. This metric is computed as the ratio between the sensitive patterns found in the sanitised data set and those found in the original data set. If HF is equal to zero, it means all the patterns are well hidden.

Error-based metrics. these metrics measure the effectiveness of the sensitive information extraction process, for example, using the distance between the original data and the estimate. A lack of privacy generally takes place in case of small estimate errors. In location privacy, the expected estimation error measures the inference correctness by computing the expected distance between the true location and the estimated location using a distance metric, such as the Euclidean distance [155]. Furthermore, with particular regard to high-dimensional, unstructured data such as the ones acquired by mobile background sensors or images, a simple but common approach to quantify privacy consists in comparing the traditional performance metrics of sensitive attribute

extraction methods (i.e. accuracy) before and after the data modification process. A significant performance drop is a valid indicator of the effectiveness of a data modification technique.

Accuracy-based metrics. quantify the accuracy of the inference mechanism, as inaccurate estimates typically show higher privacy. The confidence interval width indicates the amount of privacy given the estimated interval in which the true outcome lies [24]. It is expressed in percentage terms for a certain confidence level. *(t, δ)* privacy violation [95] provides information whether the release of a classifier for public data is a privacy threat, depending on how many training samples are available to the adversary algorithm. Training samples link public data to sensitive data for some individuals, and privacy is violated when it is possible to infer sensitive information from public data for individuals who are not in the training samples. In location privacy, the size of the uncertainty region denotes the minimal size of the region to which it is possible to narrow down the position of a target user, while the coverage of sensitive region evaluates how a userâĂŹs sensitive regions overlap with the uncertainty region [56]. A different approach was proposed in [32]. In this work, data subjects are given the possibility to customise the accuracy of the region they are in when submitting it to an internet service. The accuracy of the obfuscated region can therefore be seen as an indicator of privacy.

Time-based metrics. time-based metrics measure the time that elapses before sensitive information can be extracted. For instance, in location tracking, to evaluate a given privacy protection method, it can be useful to measure for how long it is possible to breach privacy by successfully tracking the user, by computing the maximum tracking time [148] or the mean time to confusion [82].

6 PRIVACY PROTECTION METHODS FOR SENSITIVE DATA

Given the amount of personal and sensitive information that can be extracted from mobile device sensors, it is necessary to apply a series of techniques to protect the data, as specified in the GDPR. The data should be used for its primary purpose, consented by the user, and it should not be possible to obtain additional information from the re-purposed data. Privacy protection methods aim to decrease the effectiveness of information extraction tools by transforming data with regard to specific sensitive attributes, while preserving the utility of the data for the original application scenario. In the remainder of this section, the discussed methods are grouped according to the type of input data they work on: *(i) traditional data modification techniques* work well with structured data, as most of them were developed for the purpose of disclosing sanitised datasets and their application fulfils the requirements of some of the properties discussed above, thus guaranteeing a certain degree of privacy; *(ii) machine learning-based data modification techniques*, which are more apt in the case of complex *unstructured* data, as the relationship between privacy gains and information loss changes completely for high-dimensional, highly correlated unstructured data like images, audio signals and time sequence signals provided by background sensors in mobile devices [125, 178]. An overview of the different privacy protection methods can be found in Table 4.

6.1 Traditional Data Modification Methods

Traditional data modification techniques have proven to work well with structured data. According to [173], these methods can be divided into the following groups:

6.1.1 Data Perturbation. It is accomplished by the alteration of an attribute value by a new value. Among traditional data perturbation approaches, randomisation techniques are based on the use of noise to mask the values of the data [20]. By incorporating sufficiently large noise, individual data can in fact no longer be recovered, whilst the probability distribution of the aggregate data can be recovered and used safely from a privacy protection standpoint. Noise can be added to the original values in a number of ways:

- additive noise, which works by adding a stochastic value to confidential quantitative attributes [41, 121];

Table 4. Comparison of different state-of-the-art Privacy Protection Methods for Sensitive Data. AE- Autoencoder, SGD- Stochastic Gradient Descent, CNN- Convolutional Neural Network, GAN- Generative Adversarial Network, SAN- Semi-Adversarial Network, FAR- False Acceptance Rate, TASR- Task Assigment Sucess Rate, HF- Hiding Failure, Acc- Accuracy, SA- Sensitive Attribute, AUROC- Area Under the Receiver Operating Characteristic, AD- Attribute Disclosure, IVE- Incremental Variable Eliminator, COCR- Correct Overall Classification Rate, LFW- Labeled Faces in the Wild

Traditional Methods					
Method/ Classifier	**Field**	**Sensitive Data Protected**	**Study**	**Best Performance**	**Database**
Data Perturbation	Fingerprint Faces Images	Demographics	Sadhya *et al.* (2016) [146]	0.45% probability of success @ FAR = 10%	VC2002-DB1 Database AR Face Database
	Location Data	Location Tracking	Yang *et al.* (2018) [184]	TASR ≈ 80%	SimpleGeo Places Database Yelp Database
Data Blocking	Weather Parameters	Health Parameters	Parmar *et al.* (2011) [134]	HF = 0/3 attribute disclosure	UCI Repository: Weather Dataset
Data Aggregation or Merging	Physiologic Signals	Health Parameters	Ren *et al.* (2013) [140]	-	MIT-BIH Polysomnographic Database
Data Swapping	Personal Attributes	Health Parameters	Hasan *et al.* (2016) [79]	l-Diversity = 0 attribute disclosure	UCI Repository: Synthetic Dataset Adult Dataset
Data Sampling	Personal Attributes	Health Parameters	Liu *et al.* (2019) [111]	l-Diversity ≈ 0.15 error	UCI Repository: Adult Dataset
Machine Learning-based Methods					
Method/ Classifier	**Field**	**Sensitive Data Protected**	**Study**	**Best Performance**	**Database**
Data Level Methods					
Differential Privacy-based AE	Activity Signals, Biomarkers, Biometric Measures	Health Parameters	Phan *et al.* (2016) [136]	Acc. Privacy ≈ 85%	Own Database
SGD sanititation	Language Modeling	Text Inferring	McMahan *et al.* (2018) [115]	-0.13% in accuracy with ($4.6e10^{-9}$)-differential privacy	Reddit Dataset
Siamese CNN	Face Images	Identity	Osia *et al.* (2019) [142]	EER before ≈ 1% EER after ≈ 28%	IMDB-Wiki + LFW Datasets
	Activity Signals	Demographics		EER before ≈ 22% EER after ≈ 36%	MotionSense Dataset
Siamese CNN	Activity Signals	Demographics	Garofalo *et al.* (2019) [72]	F1-score SA before = 72.58% F1-score SA after = 52.99%	OU-ISIR Database
GAN	Activity Signals	Demographics	Ngueveu *et al.* (2020) [44]	Acc. SA before = 98.5% Acc. SA after = 61.0% Acc. SA before = 98.5% Acc. SA after = 57.0%	MotionSense Dataset MobiAct Dataset
SAN	Face Images	Demographics	Mirjalili *et al.* (2018) [171]	Error Rate SA before = 19.7% Error Rate SA after = 39.3 % Error Rate SA before = 8.0% Error Rate SA after = 39.2 % Error Rate SA before = 33.4% Error Rate SA after = 72.5 % Error Rate SA before = 16.9% Error Rate SA after = 53.8%	CelebA Dataset MORPH Dataset MUCT Dataset RaFC Dataset
	Face Images	Demographics	Mirjalili *et al.* (2020) [172]	EER SA before ≈1% EER SA after = 20% EER SA after = 20% EER SA after = 10% EER SA after = 10%	CelebA Dataset UTK-face Dataset MORPH Dataset MUCT Dataset
AE	Activity Signals	Demographics	Delgado-Santos *et al.* (2021) [62]	AUROC SA before = 99.00% AUROC SA after = 57.2%	MotionSense + MobiAct Databases
Feature Level Methods					
Decision Tree Ensemble	Face Images	Demographics	Terhorst *et al.* (2019) [163]	COCR before = 94.7% COCR after = 64.7%	FERET Database
AE	Face Images	Demographics	Bortolato *et al.* (2020) [40]	EER SA before = 1.8% EER SA after = 41.9% EER SA before = 4.9% EER SA after = 41.4% EER SA before = 14.5% EER SA after = 50.2%	CelebA Dataset LFW Dataset Adience Dataset
Sensitivity Detector + Triplet Loss	Face Images	Demographics	Morales *et al.* (2020) [14]	Acc. SA before = 95.1% Acc. SA after = 54.6%	DiveFace Database

- multiplicative noise, in which protected numerical attributes are multiplied by a stochastic value [99];
- geometric perturbation, in which a mix of additive and multiplicative perturbations are used through a rotation matrix [94];
- nonlinear transformation, applying a sigmoid distortion for mapping the data to a different space but preserving the statistical properties of the data [4, 38];
- data condensation, in which the data is transformed into a new distribution where the new data include the correlations among the different dimension [19];
- through a combination of the above techniques [48].

Differential privacy has been widely used in several applications. For instance, in [146], differential privacy was used in a privacy-preserving framework for a recognition system based on soft biometrics, such as age, gender, height, and weight extracted from fingerprints and face images. In the context of mobile devices, differential privacy has also been applied for providing rigorous protection of worker locations in a company centralised server crowdsensing application [184].

6.1.2 Data blocking. It consists in hiding a certain set of sensitive attributes by replacing the original attribute values. An existing attribute value can be replaced with a predetermined value to indicate the data suppression (it could be “?", “0" or “x" in the case of one-character values). A new sanitised dataset is generated in which the classification rules of the sensitive data can no longer be extracted [96, 134].

6.1.3 Data Aggregation or Merging. It is the combination of values in a coarser category [109] or the processing by a compression algorithm to reduce the number of embedded bits used to store the sensitive data. The compression algorithm is used to reduce sensitive embedded data to alleviate the effect of data masking on the quality of ordinary data. At the same time, data merging lowers the processing power consumption [140].

6.1.4 Data Swapping. It refers to interchanging values of individual records decreasing the risk of attribute disclosure. This technique obtains new data with no valid information making impossible for the adversary to access the real data. In addition, the data swapping method ensures that the published data satisfies *l*-diversity and guarantees that the adversary cannot violate individual privacy [79].

6.1.5 Data Sampling. It consists in the releasing data of a sample of the population. This technique is based on the conditional probability distribution of the data. However, values that appear with little recurrence in a dataset may give rise to privacy problems. Therefore, it is important to choose a sample from the set that is representative and has the same shape as the original dataset, thus achieving good results in terms of *(d-γ)*-privacy [50, 111].

Such strategies have found a large number of different implementations for structured data and are often adopted by governmental or statistical agencies. Many are available in libraries under open-source license, like ARX[5] or the R-package sdcMicro [162, 178]. However, a critical aspect of these modification techniques is often scalability, i.e. there is a significant performance drop as the number of the dimensions of the dataset increases; in addition, the computational overhead will increase exponentially with respect to the number of attributes and number of instances. These limitations of the traditional data modification methods are commonly grouped under the label of “curse of dimensionalityâĂİ [102].

6.2 Machine Learning-based Data Modification Methods

In addition to the goal of information extraction as discussed in Sec. 4, considering its potential in big data processing [137], machine learning approaches have in turn been investigated for the purpose of perturbing the data in the attempt to overcome the limitations of traditional modification techniques. Within these algorithms, a

[5]Available at https://arx.deidentifier.org.

subdivision into two groups can be made of those that operate at the *i)* data level and those that operate at the *ii)* feature level, depending on the input data. In this section we present a brief summary of the most competitive techniques of the two groups according to [40].

6.2.1 Data level methods. Algorithms that operate at the data level have raw data as input. Within the algorithm itself they are processed and the output is a transformed dataset containing the protected sensitive data.

Among privacy protection solutions adopted to protect sensitive data in the context of machine learning models, differential privacy-based mechanisms are popular in the literature. In [136], a differentially private model implementation based on perturbing the objective functions was proposed for deep Autoencoders (AE) for human behaviour prediction in a health social network. Such method can be applied to each layer of the network. Similarly, the idea of sanitising the gradient in Stochastic Gradient Descent (SGD) was introduced in [16] for CNN, and for complex sequence models for next-word prediction in [115]. Differential privacy has also been implemented in dedicated Tensorflow[6] and PyTorch[7] libraries. Generally, however, at a modest privacy budget differentially private mechanisms come with a cost in software complexity, training efficiency, and model quality [168].

Using a convolutional architecture, another possibility is offered by the Siamese architecture, which has two different input vectors while maintaining equal weights in the two halves of the network to acquire comparable output vectors. Osia *et al.* [142] used this architecture both in the field of facial images, to protect the identity of the person, and in the field of activity recognition to protect the gender of the user. The authors in [72] also used a Siamese CNN. In this case their work focused solely on activity recognition while protecting demographic information.

Also, Generative Adversarial Networks (GAN) are among the most popular techniques considered for this purpose in the literature. GAN are unsupervised methods that exploit two adversarial subnetworks (the *generator* and the *discriminator*), and are able to learn well, in a competitive manner, the statistical structure of high dimensional signals. A GAN-based approach called DySan was developed in [44] for data sanitisation, in the context of a mobile application for physical activity monitoring through the accelerometer and the gyroscope data. Before sending the data to a server hosted on the cloud, gender inferences are prevented by distorting the data while limiting the loss of accuracy on physical activity monitoring.

A similar approach for privacy protection is based on Semi-Adversarial Networks (SAN). SAN are different from typical GAN in the fact that, in addition to the generator subnetwork, they include two independent discriminator classifiers rather than one. A semi-adversarial configuration was proposed by Mirjalili *et al.* [171] for the purpose of image data perturbation. Based on the feedback of two classifiers, where one acts as an adversary of the other, this model was able to privatise gender while maintaining the same accuracy in face recognition. The authors extended their work in [172], by including, among other things, the possibility of choosing to obfuscate specific attributes (e.g., age and race), while allowing for other types of attributes to be extracted (e.g., gender).

Delgado-Santos *et al.* [62] proposed GaitPrivacyON, an autoencoder trained in an unsupervised way. The authors where able to create new transformed data that achieved significant improvements in protection in the gait verification task while sensitive information reminded private (e.g., activity and gender).

6.2.2 Feature level methods. There is a second set of methods that, instead of using raw data as input, apply on the embedding representation of the data. Therefore, a pre-trained model used as features extractor is needed. After that, this set of features will be the input of the privacy method. Finally, a transformed dataset that keeps the sensitive data privatised will be the output. Terhöst *et al.* [163] proposed an Incremental Variable Eliminations

[6] Available at https://github.com/tensorflow/privacy.
[7] Available at https://github.com/pytorch/opacus.

algorithm (IVE). The authors, by training a set of decision trees, obtain a measure of the importance of the variables that predict the sensitive attributes to be reduced.

An AE was also used by Bortolato *et al.* [40]. The authors introduced Privacy-Enhancing Face-Representation learning Network (PFRNet), a neural network-based model that works at the level of face representations (templates) from images, aiming to achieve distinct encodings for both identity and gender in the feature space. The model showed how training a loss function for gender suppression (where the distributions of male and female subjects were similar) for the identity feature space, was an effective way to preserve privacy.

Morales *et al.* [14] aimed to leave out sensitive information in the decision-making process in an image-based face recognition system without a significant drop of performance by focusing on the feature space. Developed for the purpose of ensuring fairness and transparency, their systems inherently improve the privacy of the data. It works as an independent, decoupled module on top of a pre-trained model and takes as input the embeddings generated by the model. By defining and minimising its own triplet-loss function, SensitiveNets generates new representations agnostic of gender and ethnicity information, which however still retain information useful for extraction of other attributes.

6.3 Other Perspectives

Finally, it is important to highlight that in order to protect users' privacy while handling their private data, besides data modification methods, other important perspectives to be considered to comply with secure data management practices in relation to privacy include:

6.3.1 Template protection. It is an important field of research in the area of biometrics. Templates are compact representations of users' biometric data for the purpose of storage. They are transformed into protected biometric references for security purposes. Template protection schemes should provide the following properties [126]:

- Irreversibility: it should be computationally difficult[8] to compute the original template from a subjectâĂŹs protected biometric reference.
- Revocability: it should be computationally difficult to compute the original biometric template from multiple instances of protected biometric reference derived from the same biometric trait of an individual. Biometric data is permanently associated with the data subject and it cannot be revoked and reissued if compromised, contrarily to credit cards or passwords. However, through revocable and irreversible transformations templates can be cancelable, thus mitigating the risks associated with biometric template theft [135].
- Unlinkability: it should be computationally difficult to determine whether two or more instances of protected biometric reference were obtained from the same biometric trait of a user. Unlinkability prevents cross-matching across databases.

6.3.2 Data outsourcing. Usually mobile applications exploit cloud resources for model training and inference. Therefore, users' personal data containing sensitive information may be on the internet. If stored on the cloud, data subject privacy undergoes greater risks than being stored locally in the device [157]. Performing the training and inference tasks locally is among alternative solutions investigated. However, the computational resource constraints are much stricter [53, 152].

A different approach could be federated learning, a machine-learning strategy according to which models are trained on datasets distributed across multiple devices, thus preventing data leakage [101, 112]. However, recent attacks demonstrate that simply maintaining data locality during training processes does not provide sufficient privacy guarantees as intermediate results, if exposed, could still cause some information leakage [185]. Possible solutions to this problem are given by differential privacy mechanisms and Secure Multiparty Computation (SMC) schemes, or a combination of the two [169].

[8] A problem is defined computationally difficult if it cannot be solved using a polynomial-time algorithm.

Finally, it should be pointed out that the considered techniques should be complemented by widely deployed encryption protocols that would guarantee data security, such as hash functions, secret-key and public-key cryptography, among others [57, 80].

7 CONCLUSIONS AND OPEN RESEARCH QUESTIONS

7.1 Conclusions

As demonstrated, seemingly innocuous user data can reveal personal and sensitive information about the user, which must be protected in compliance with the GDPR. We have provided a state-of-the-art review of the different kinds of sensitive data that can be extracted by the mobile device sensor data. A survey of the metrics that allow a comparison of different aspects and quantify the effectiveness of the privacy protection methods was carried out for the purpose of identifying the most suitable metric for each specific application. Some of the most popular privacy protection data modification methods were also discussed, aiming to offer useful guidelines for managing the trade-off between protecting the sensitive attributes while disclosing the allowed attributes, inherent to the privacy problem.

7.2 Open Research Questions

Many paths of development remain to be investigated. The most relevant ones are discussed below.

7.2.1 Protection of the Privacy of User Sensitive Data.

Correlation between Sensitive Attributes. It is important to observe the correlation between the different sensitive attributes, in order to identify from which sensitive attributes it is possible to extract others. For example, the user location obtained from the mobile device Wi-Fi data can also reveal information about the activity a user is involved in.

Data Modification Algorithms for Privacy Protection. As shown in Sec. 4, inferrable attributes can assume very diverse sets of values, in terms of size and number of attributes per subject. For instance, the presence of a disease or the gender of a data subject are unique to each subject and can assume a binary or limited set of values. A different scenario is given by attributes such as the age (unique attribute, but wider set of possible values), or the activity the user is involved in or their location (several possible attributes per subject, but one at a time). Depending on the formulation of the attribute output categories, at the cost of increased system complexity, it is possible to achieve a finer granularity in terms of information about the data subject, which typically relates to a higher extent of privacy invasion. Therefore, from the perspective of sensitive data protection, a possible step towards the protection of the privacy of sensitive data could be developing a system that would modify the data so that the possible sensitive attribute recognisable output categories would be fewer and coarser.

Ethical Implications. The digitalisation of data storage and communications, combined with the ever-growing capacity of computers to automatically process data, has made possible to mine structures and relationships lying in the data to extract information in unprecedented ways. Among other things, the GDPR provides a definition of personal and sensitive information to safeguard the right to privacy in the digital domain, thus laying the cornerstone of an ethical usage of user data. Nonetheless, even if the sensitive information is suppressed, it would be beneficial to assess the side effects of automated processing, with regard to sensitive attributes, paying special attention to the ethical consequences this might entail. Therefore, even if data is collected and processed for a legitimate purpose, the results yielded might be influenced by personal and sensitive information that the models are covertly recognising and exploiting. For instance, in 2018, Amazon withheld their machine learning engine in charge of selecting the most suitable job applicant profiles as it was discovered that it was biased against women, downgrading resumes that included the word âĂIJwomenâĂŹs,âĂİ as in âĂIJwomenâĂŹs chess club

captainâĂİ and graduates of all-womenâĂŹs universities. This was due to the fact that the models utilised were trained with resumes submitted to the company over a 10-year period, which came mostly from men [1]. Such risks are exacerbated by the fact that, in the case of deep learning models, it is often difficult to ascertain how such information is encoded in the intermediate layers, and that the sensitive and legitimate attributes might be entangled within their representation instances. Fairness in AI is an novel, yet very active field of investigation, deeply connected with the protection of the privacy of user sensitive data.

7.2.2 Performance of the Algorithms.

Robustness. Given the ubiquity of mobile devices, the data are captured by the built-in sensors in a variety of different scenarios. Therefore, a typical requirement of mobile device computing is robustness. For instance, with regards to the recognition of sensitive information, the property of position invariance would grant a negligible impact in the performance of an algorithm due to changes in the position of the mobile device with respect to the user who is carrying. In other words, the algorithm should be able to recognise the predetermined user attributes regardless of whether the mobile device is in the front pocket, in the hand or in the backpack or whether it is performing a specific activity such as answering a call or typing.

Reliability of Labels. It is important to identify whether the subjects themselves are in charge of the task of labelling the sensitive attribute. In this case it is important to ensure that a subject is able to do it in an objective way, in the case of mood recognition, for instance. Additionally, with particular regard to 3D motion sensor data in the time domain, labelling is often not straight-forward and it can be expensive and time-consuming. Improving or overcoming the labelling process is an interesting open problem for further investigation. A solution could be adopting self-supervised learning (SSL), a paradigm according to which the training of the feature extraction algorithms can take place in an unsupervised manner.

Impact of Hardware Differences. After performing a study of the different mobile device sensors, it would be interesting to evaluate how innate sensor characteristics affect the processes of sensitive data extraction and protection. This is due to the fact that not all smart device sensors have the same characteristics, i.e. full-scale values, resolution, sampling frequencies, etc.

Computation Time. With regard to mobile devices, time constraints are often crucial for real-time applications, and a seamless user experience is among the main user concerns. Therefore, incorporating in the processing chain additional steps aiming to protect the privacy of the sensitive data should not impact the computation time significantly.

Storage of the Algorithms. Finally, a significant aspect is related to the storage of the algorithms. The captured user raw data may then be sent to the cloud, for training the models, as more powerful hardware resources are typically available remotely. In such way, the raw data might be exposed to greater risks related to being transmitted and stored in a server. It is therefore necessary to develop systems that achieve the desired degree of sensitive data protection, without impacting the performance of the models. Among the solutions proposed for such goals is federated learning, in combination with algorithms that would guarantee differential privacy and SMC.

7.2.3 General metric framework. With regard to the protection of the privacy of sensitive data, it would be desirable to create a general metric framework that can be applied to any set of protected data and indicate with certainty the degree of protection through a score, encompassing which attributes are being protected and how many classes are being used to differentiate an attribute. Based on this, a standardised set of limit values should be established in order to indicate the point at which sensitive data is considered fully protected. In such way,

protected data could be freely processed for extraction of information without putting at stake the privacy of users' sensitive data.

ACKNOWLEDGMENTS

This project has received funding from the European UnionâĂŹs Horizon 2020 research and innovation programme under the Marie SkÅĆodowska-Curie grant agreement No 860315. R. Tolosana and R. Vera-Rodriguez are also supported by INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER).

REFERENCES

[1] 2016. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed: 2021-06-16.
[2] 2016. EU 2016/679 (General Data Protection Regulation). https://gdpr-info.eu/
[3] 2017. *Health Informatics âĂŤ Pseudonymization.* Technical Report. International Organization for Standardization.
[4] 2018. Privacy-preserving Collaborative Fuzzy Clustering. *Data & Knowledge Engineering* 116 (2018), 21–41.
[5] 2019. PriMa: Privacy Matters, H2020-MSCA-ITN-2019-860315. https://www.prima-itn.eu/
[6] 2019. TReSPAsS-ETN: TRaining in Secure and PrivAcy-preserving biometricS, H2020-MSCA-ITN-2019-860813. https://www.trespass-etn.eu/
[7] 2021. Number of Apps Available in Leading App Stores as of 3rd Quarter 2020. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/
[8] S. R. M. Oliveira and O. R. ZaÄĭane. 2002. Privacy Preserving Frequent Itemset Mining. In *Proc. IEEE international conference on Privacy, security and data mining.*
[9] A. Acien, A. Morales, J. Fierrez, R. Vera-Rodriguez, and J. Hernandez-Ortega. 2019. Active Detection of Age Groups Based on Touch Interaction. *IET Biom.* 8, 1 (2019), 101–108.
[10] A. Acien, A. Morales, R. Vera-Rodriguez, J. Fierrez, and R. Tolosana. 2019. Multilock: Mobile Active Authentication based on Multiple Biometric and Behavioral Patterns. In *Proc. International Workshop on Multimodal Understanding and Learning for Embodied Applications.*
[11] A. J. Aviv, B. Sapp, M. Blaze, and J. M. Smith. 2012. Practicality of Accelerometer Side Channels on Smartphones. In *Proc. Annual Computer Security Applications Conference.*
[12] A. Jain and V. Kanhangad. 2016. Investigating Gender Recognition in Smartphones using Accelerometer and Gyroscope Sensor Readings. In *Proc. International Conference on Computational Techniques in Information and Communication Technologies.*
[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. 2007. L -diversity: Privacy beyond K -anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1 (2007), 3.
[14] A. Morales, and J. Fierrez, and R. Vera-Rodriguez, and R. Tolosana. 2020. SensitiveNets: Learning Agnostic Representations with Application to Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
[15] A. Tal, and Z. Shinar, and D. Shaki, and S. Codish, and A. Goldbart. 2017. Validation of Contact-free Sleep Monitoring Device with Comparison to Polysomnography. *Journal of Clinical Sleep Medicine* 13, 3 (2017), 517–522.
[16] M. Abadi, A. Chu, I. Goodfellow, H. McMahan, I. Mironov, K. Talwar, and L. Zhang. 2016. Deep Learning with Differential Privacy. In *Proc. ACM SIGSAC Conference on Computer and Communications Security.* 308âĂŞ318.
[17] A. Acien, A: Morales, J. Fierrez, R. Vera-Rodriguez, and O. Delgado-Mohatar. 2020. Becaptcha: Bot Detection in Smartphone Interaction using Touchscreen Biometrics and Mobile Sensors. *arXiv* (2020).
[18] A. Acien, A. Morales, J. V. Monaco, R. Vera-Rodriguez, and J. Fierrez. 2021. TypeNet: Deep Learning Keystroke Biometrics. *arXiv* (2021).
[19] C. Aggarwal and P. Yu. 2004. A Condensation Approach to Privacy Preserving Data Mining. In *Proc. Advances in Database Technology.*
[20] C. Aggarwal and P. Yu. 2008. *A Survey of Randomization Methods for Privacy-Preserving Data Mining.* 137–156.
[21] Charu C. Aggarwal. 2005. On K-Anonymity and the Curse of Dimensionality. In *Proc. International Conference on Very Large Data Bases.* 901âĂŞ909.
[22] D. Agrawal and C. Aggarwal. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.* 247âĂŞ255.
[23] R. Agrawal, T. Imieliński, and A. Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. In *Proc. ACM SIGMOD International Conference on Management of Data.* 207âĂŞ216.
[24] R. Agrawal and R. Srikant. 2000. Privacy-preserving Data Mining. In *Proc. ACM SIGMOD international conference on Management of data.*
[25] E: Albanese, L. Launer, M. Egger, M. Prince, P. Giannakopoulos, F. Wolters, and K. Egan. 2017. Body Mass Index in Midlife and Dementia: Systematic Review and Meta-Regression Analysis of 589,649 Men and Women Followed in Longitudinal Studies. *Alzheimer's*

& Dementia: Diagnosis, Assessment & Disease Monitoring 8 (2017), 165–178.
[26] Atheer Aljeraisy, Masoud Barati, Omer Rana, and Charith Perera. 2021. Privacy Laws and Privacy by Design Schemes for the Internet of Things: A DeveloperâĂŹs Perspective. *ACM Comput. Surv.* 54, 5, Article 102 (May 2021), 38 pages. https://doi.org/10.1145/3450965
[27] A. Almaatouq, PF. rieto Castrillo, and A. Pentland. 2016. Mobile Communication Signatures of Unemployment. In *Proc. International Conference on Social Informatics*. 407–418.
[28] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. 2013. Geo-indistinguishability: Differential Privacy for Location-based Systems. In *Proc. ACM SIGSAC conference on Computer & communications security*. 901–914.
[29] A. Anjum and M. Ilyas. 2013. Activity Recognition using Smartphone Sensors. In *Proc. IEEE Consumer Communications and Networking Conference*. 914–919.
[30] A. Anjum and M. U. Ilyas. 2013. Activity Recognition using Smartphone Sensors. In *Proc. IEEE Consumer Communications and Networking Conference*. 914–919.
[31] A. D. Antar, M. Ahmed, and M. Ahad. 2019. Challenges in Sensor-based Human Activity Recognition and a Comparative Analysis of Benchmark Datasets: a Review. In *Proc. International Conference on Informatics, Electronics & Vision and International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. 134–139.
[32] C. A. Ardagna, M. Cremonini, E. Damiani, S. Di Vimercati, and P. Samarati. 2007. Location Privacy Protection through Obfuscation-based Techniques. In *Proc. IFIP Annual Conference on Data and Applications Security and Privacy*. 47–60.
[33] Z. Arnold, D. Larose, and E. Agu. 2015. Smartphone Inference of Alcohol Consumption Levels from Gait. In *2015 International Conference on Healthcare Informatics*. 417–426.
[34] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. 1999. Disclosure Limitation of Sensitive Rules. In *Proc. Workshop on Knowledge and Data Engineering Exchange*. 45–52.
[35] K. Barker, M. Askari, M. Banerjee, K. Ghazinour, B. Mackas, M. Majedi, S. Pun, and A. Williams. 2009. A Data Privacy Taxonomy. In *Proc. British National Conference on Databases*. 42–54.
[36] S. Barth, M. D.T. de Jong, M. Junger, P. H. Hartel, and J. C. Roppelt. 2019. Putting the Privacy Paradox to the Test: Online Privacy and Security Behaviors among Users with Technical Knowledge, Privacy Awareness, and Financial Resources. *Telematics and Informatics* 41 (2019), 55–69.
[37] C. Bevan and D. Fraser. 2016. Different Strokes for Different Folks? Revealing the Physical Characteristics of Smartphone Users from Their Swipe Gestures. *International Journal of Human-Computer Studies* (2016), 51–61.
[38] K. Bhaduri, M. D. Stefanski, and A. N. Srivastava. 2011. Privacy-Preserving Outlier Detection Through Random Nonlinear Data Distortion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 1 (2011), 260–272.
[39] Matthew Boakes, Richard Guest, Farzin Deravi, and Barbara Corsetti. 2019. Exploring Mobile Biometric Performance through Identification of Core Factors and Relationships. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1, 4 (2019), 278–291.
[40] B. Bortolato, M. Ivanovska, P. Rot, J. Križaj, P. Terhörst, N. Damer, P. Peer, and V. Štruc. 2020. Learning Privacy-enhancing Face Representations through Feature Disentanglement. In *Proc. International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. 45–52.
[41] R. Brand. 2002. *Microdata Protection through Noise Addition.* 97–116.
[42] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and MB. Srivastava. 2006. Participatory Sensing. (2006).
[43] C. Dwork and A. Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. TrendsÂő Theor. Comput. Sci.* 9, 3-4 (2014), 211âĂŞ407.
[44] C. R. Ngueveu, A. Boutet, C. Frindel, S. Gambs, T. Jourdan, and C. Rosin. 2020. DYSAN: Dynamically Sanitizing Motion Sensor Data Against Sensitive Inferences through Adversarial Networks. *ArXiv* (2020).
[45] L. Cai and H. Chen. 2011. TouchLogger: Inferring Keystrokes on Touch Screen from Smartphone Motion. *HotSec* 11 (2011), 9.
[46] B. Cao, L. Zheng, C. Zhang, P. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. Leow. 2017. DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[47] R. Castrillon, A. Acien, J.R. Orozco-Arroyave, A. Morales, J.F. Vargas, R.Vera-RodrÄśguez, J. Fierrez, J. Ortega-Garcia, and A. Villegas. 2019. Characterization of the Handwriting Skills as a Biomarker for Parkinson Disease. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019) - Human Health Monitoring Based on Computer Vision.*
[48] M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil. 2018. Efficient Data Perturbation for Privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing* 48 (2018), 1–19.
[49] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. 2013. Broadening the Scope of Differential Privacy using Metrics. In *Proc. International Symposium on Privacy Enhancing Technologies Symposium*. 82–102.
[50] K. Chaudhuri and N. Mishra. 2006. When Random Sampling Preserves Privacy. In *Advances in Cryptology - CRYPTO 2006*. Berlin, Heidelberg, 198–213.
[51] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. 2005. Toward Privacy in Public Databases. In *Proc. Theory of Cryptography Conference*. 363–385.

[52] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (May 2021), 40 pages. https://doi.org/10.1145/3447744

[53] Yanjiao Chen, Baolin Zheng, Zihan Zhang, Qian Wang, Chao Shen, and Qian Zhang. 2020. Deep Learning on Mobile and Embedded Devices: State-of-the-Art, Challenges, and Future Directions. *ACM Comput. Surv.* 53, 4, Article 84 (Aug. 2020), 37 pages. https://doi.org/10.1145/3398209

[54] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. 2013. Unobtrusive Sleep Monitoring using Smartphones. In *Proc. International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. 145–152.

[55] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui. 2018. WiFi CSI based Passive Human Activity Recognition using Attention based BLSTM. *IEEE Transactions on Mobile Computing* 18, 11 (2018), 2714–2724.

[56] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar. 2006. Preserving User Location Privacy in Mobile Data Management Infrastructures. In *Proc. International Workshop on Privacy Enhancing Technologies*. 393–412.

[57] Lianhua Chi and Xingquan Zhu. 2017. Hashing Techniques: A Survey and Taxonomy. *ACM Comput. Surv.* 50, 1, Article 11 (April 2017), 36 pages. https://doi.org/10.1145/3047307

[58] Tore Dalenius. 1986. Finding a Needle in a Haystack or Identifying Anonymous Census Records. *Journal of Official Statistics* 2, 3 (1986), 329.

[59] A. Dantcheva, P. Elia, and A. Ross. 2016. What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. *IEEE Transactions on Information Forensics and Security* 11, 3 (2016), 441–467.

[60] K. David and H. Berndt. 2018. 6G Vision and Requirements: Is There Any Need for Beyond 5G? *IEEE Vehicular Technology Magazine* 13, 3 (2018), 72–80.

[61] O. Delgado-Mohatar, R. Tolosana, J. Fierrez, and A. Morales. 2020. Blockchain in the Internet of Things: Architectures and Implementation. In *Proc. IEEE 44th Annual Computers, Software, and Applications Conference*. 1072–1077.

[62] Paula Delgado-Santos, Ruben Tolosana, Richard Guest, Ruben Vera, Farzin Deravi, and Aythami Morales. 2021. GaitPrivacyON: Privacy-Preserving Mobile Gait Biometrics using Unsupervised Learning. *arXiv preprint arXiv:2110.03967* (2021).

[63] J. Dobner and S. Kaser. 2018. Body Mass Index and the Risk of Infection-from Underweight to Obesity. *Clinical Microbiology and Infection* 24, 1 (2018), 24–28.

[64] J. Domingo-Ferrer and J. Soria-Comas. 2015. From t-Closeness to Differential Privacy and Vice Versa in Data Anonymization. *Knowledge-Based Systems* 74 (2015), 151–158.

[65] F. du Pin Calmon and N. Fawaz. 2012. Privacy Against Statistical Inference. In *Proc. Allerton Conference on Communication, Control, and Computing*. IEEE, 1401–1408.

[66] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. 2006. Our data, Ourselves: Privacy Via Distributed Noise Generation. In *Proc. International Conference on the Theory and Applications of Cryptographic Techniques*. 486–503.

[67] E. Davarci, B. Soysal, I. Erguler, S. O. Aydin, O. Dincer, and E. Anarim. 2017. Age Group Detection using Smartphone Motion Sensors. In *Proc. European Signal Processing Conference*.

[68] E. Owusu, J. Han, S. Das, A. Perrig, and J. Zhang. 2012. ACCessory: Password Inference using Accelerometers on Smartphones. In *Proc. Workshop on Mobile Computing Systems & Applications*.

[69] E. Ellavarason, R: Guest, F. Deravi, R. Sanchez-Riello, and B. Corsetti. 2020. Touch-dynamics based Behavioural Biometrics on Mobile Devices–A Review from a Usability and Performance Perspective. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–36.

[70] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. 2004. Privacy Preserving Mining of Association Rules. *Information Systems* 29, 4 (2004), 343–364.

[71] M. Faundez-Zanuy, J. Fierrez, M. A. Ferrer, M. Diaz, R. Tolosana, and R. Plamondon. 2020. Handwriting Biometrics: Applications and Future Trends in E-security and E-health. *Cognitive Computation* 12, 5 (2020), 940–953.

[72] G. Garofalo, D. Preuveneers, and W. Joosen. 2019. Data Privatizer for Biometric Applications and Online Identity Management. In *Proc. IFIP International Summer School on Privacy and Identity Management*.

[73] C. Gao, K. Fawaz, S. Sur, and S. Banerjee. 2019. Privacy Protection for Audio Sensing Against Multi-Microphone Adversaries. *Proc. Privacy Enhancing Technologies* 2019, 2 (2019), 146 – 165.

[74] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What Does Touch Tell Us about Emotions in Touchscreen-Based Gameplay? *ACM Trans. Comput.-Hum. Interact.* 19, 4 (2012).

[75] Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. 2018. Mental Health Monitoring with Multimodal Sensing and Machine Learning: A Survey. *Pervasive and Mobile Computing* 51 (2018), 1–26.

[76] S. L. Garfinkel. 2015. De-identification of Personal Information. *National Institute of Standards and Technology* (2015).

[77] F. Gravenhorst, A. Muaremi, J. Bardram, A. Grünerbl, O. Mayora, G. Wurzer, M. Frost, V. Osmani, B. Arnrich, P. Lukowicz, et al. 2015. Mobile Phones as Medical Devices in Mental Disorder Treatment: an Overview. *Personal and Ubiquitous Computing* 19, 2 (2015), 335–353.

[78] Muhammad Haris, Hamed Haddadi, and Pan Hui. 2014. Privacy Leakage in Mobile Computing: Tools, Methods, and Characteristics. *arXiv* (2014).
[79] A. Hasan, Q. Jiang, J. Luo, C. Li, and L. Chen. 2016. An Effective Value Swapping Method for Privacy Preserving Data Publishing. *Security and Communication Networks* 9, 16 (2016), 3219–3228.
[80] L. Hernández-Álvarez, J. M. de Fuentes, L. González-Manzano, and L. Hernández Encinas. 2021. Privacy-Preserving Sensor-Based Continuous Authentication and User Profiling: A Review. *Sensors* 21, 1 (2021), 92.
[81] N Heuveldop. 2017. Ericsson Mobility Report June 2017. *Technique Report* (2017), 7.
[82] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. 2007. Preserving Privacy in GPS Traces via Uncertainty-aware Path Cloaking. In *Proc. ACM conference on Computer and communications security*. 161–171.
[83] G. Hung, P. Yang, C. Chang, J. Chiang, and Y. Chen. 2016. Predicting Negative Emotions based on Mobile Phone Usage Patterns: an Exploratory Study. *JMIR Research Protocols* 5, 3 (2016), e160.
[84] A. Hussain, T. Ali, F. Althobiani, U. Draz, M. Irfan, S. Yasin, S. Shafiq, Z. Safdar, A. Glowacz, G. Nowakowski, et al. 2021. Security Framework for IoT Based Real-Time Health Applications. *Electronics* 10, 6 (2021), 719.
[85] J. Behar, and A. Roebuck, and M. Shahid, and J. Daly, and A. Hallack, and N. Palmius, and J. Stradling, and G. D. Clifford. 2014. SleepAp: an Automated Obstructive Sleep Apnoea Screening Application for Smartphones. *IEEE journal of biomedical and health informatics* 19, 1 (2014), 325–331.
[86] J. Han, E. Owusu, L. T. Nguyen, A. Perrig, and J. Zhang. 2012. ACComplice: Location Inference using Accelerometers on Smartphones. In *Proc. Fourth International Conference on Communication Systems and Networks*.
[87] J. Hua, Z. Shen, and S. Zhong. 2017. We Can Track You if You Take the Metro: Tracking Metro Riders Using Accelerometers on Smartphones. *IEEE Trans. Inf. Forensics Secur.* 12, 2 (2017), 286–297.
[88] A. Jain and V. Kanhangad. 2019. Gender Recognition in Smartphones using Touchscreen Gestures. *Pattern Recognition Letters* 125 (2019), 604–611.
[89] A. K. Jain, S. C. Dass, and K. Nandakumar. 2004. Soft Biometric Traits for Personal Recognition Systems. In *Proc. Biometric Authentication*, David Zhang and Anil K. Jain (Eds.). 731–738.
[90] A. K. Jain, K. Nandakumar, and A. Ross. 2016. 50 Years of Biometric Research: Accomplishments, Challenges, and Opportunities. *Pattern Recognition Letters* 79 (2016), 80–105.
[91] F. John Dian, R. Vahidnia, and A. Rahmati. 2020. Wearables and the Internet of Things (IoT), Applications, Opportunities, and Challenges: A Survey. *IEEE Access* 8 (2020), 69200–69211.
[92] F Julien, M Raya, M Felegyhazi, and P Papadimitratos. 2007. Mix-Zones for Location Privacy in Vehicular Networks. (2007).
[93] K. A. Nguyen, R. N. Akram, K. Markantonakis, Z. Luo, and C. Watkins. 2019. Location Tracking Using Smartphone Accelerometer and Magnetometer Traces. In *Proc. International Conference on Availability, Reliability and Security*.
[94] K. Chen and L. Liu. [n.d.]. Privacy Preserving Data Classification with Rotation Perturbation. In *Proc. International Conference on Data Mining*. 4 pp.
[95] M. KantarcioÇğlu, J. Jin, and C. Clifton. 2004. When Do Data Mining Results Violate Privacy?. In *Proc. CM SIGKDD international conference on Knowledge discovery and data mining*. 599–604.
[96] A. Karakasidis, G. Koloniari, and V. Verykios. 2015. Privacy Preserving Blocking and Meta-Blocking. In *Machine Learning and Knowledge Discovery in Databases*. 232–236.
[97] M. Kearns, M. Pai, A. Roth, and J. Ullman. 2014. Mechanism Design in Large Games: Incentives and Privacy. In *Proc. Conference on Innovations in theoretical computer science*. 403–410.
[98] Saad Khan, Simon Parkinson, Liam Grant, Na Liu, and Stephen Mcguire. 2020. Biometric Systems Utilising Health Data from Wearable Devices: Applications and Future Challenges in Computer Security. *ACM Comput. Surv.* 53, 4, Article 85 (July 2020), 29 pages. https://doi.org/10.1145/3400030
[99] J. Kim and W. Winkler. 2003. Multiplicative Noise for Masking Continuous Data. *Statistics* 1 (2003), 9.
[100] S. J. Kim, S. Kang, Y. Choi, M. Choi, and M. Hong. 2017. Augmented-reality Survey: from Concept to Application. *KSII Transactions on Internet and Information Systems* 11, 2 (2017), 982–1004.
[101] J. Konecny, H. Brendan McMahan, D. Ramage, and P. RichtÃąrik. 2016. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. (2016).
[102] M. Köppen. 2000. The Curse of Dimensionality. In *Proc. Online World Conference on Soft Computing in Industrial Applications*, Vol. 1. 4–8.
[103] B. Krishnamurthy and C. Wills. 2009. On the Leakage of Personally Identifiable Information via Online Social Networks. In *Proc. ACM Workshop on Online Social Networks*. 7âĂŞ12.
[104] L. Chang, and J. Lu, and J. Wang, and X. Chen, and D. Fang, and Z. Tang, and P. Nurmi, and Z. Wang. 2018. SleepGuard: Capturing Rich Sleep Information using Smartwatch Sensing Data. *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–34.

[105] L. Sun, D. Zhang, B. Li, B. Guo, and S. Li. 2010. Activity Recognition on an Accelerometer Embedded Mobile Phone with Varying Positions and Orientations. *Ubiquitous Intelligence and Computing* 6406 (2010), 548âĂŞ562.
[106] L. Sweeney. 2002. K-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
[107] R. D. Labati, V. Piuri, and F. Scotti. 2011. Biometric Privacy Protection: Guidelines and Technologies. In *Proc. International Conference on E-Business and Telecommunications*. 3–19.
[108] G. Li and P. Bours. 2018. Studying WiFi and Accelerometer Data based Authentication Method on Mobile Phones. In *Proc. International Conference on Biometric Engineering and Applications*. 18–23.
[109] Q. Li and G. Cao. 2012. Efficient and Privacy-preserving Data Aggregation in Mobile Sensing. In *Proc. IEEE International Conference on Network Protocols (ICNP)*. 1–10.
[110] Zhen Lin, Michael Hewett, and Russ B Altman. 2002. Using Binning to Maintain Confidentiality of Medical Data. In *Proc. AMIA Symposium*. 454.
[111] C. Liu, S. Chen, S. Zhou, J. Guan, and Y. Ma. 2019. A Novel Privacy Preserving Method for Data Publication. *Information Sciences* 501 (2019), 421–435.
[112] Sin Kit Lo, Qinghua Lu, Chen Wang, Hye-Young Paik, and Liming Zhu. 2021. A Systematic Literature Review on Federated Machine Learning: From a Software Engineering Perspective. *ACM Comput. Surv.* 54, 5, Article 95 (May 2021), 39 pages. https://doi.org/10.1145/3450288
[113] D. G. Luca and M. Alberto. 2016. From Proximity to Accurate Indoor Localization for Context Awareness in Mobile Museum Guides. In *Proc. ACM International Conference on Mobile Human-Computer Interaction*. 1002âĂŞ1009.
[114] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen. 2020. Sensor-based Continuous Authentication of SmartphonesâĂŹ Users Using Behavioral Biometrics: A Contemporary Survey. *ArXiv* (2020).
[115] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. 2018. Learning Differentially Private Recurrent Language Models. (2018).
[116] T. Meena and K Sarawadekar. 2020. Gender Recognition using In-built Inertial Sensors of Smartphone. In *Proc. IEEE Region 10 Conference*. 462–467.
[117] Glaucia Melo, Luiz Oliveira, Daniel Schneider, and Jano de Souza. 2017. Towards an Observatory for Mobile Participatory Sensing Applications. In *Proc. International Conference on Computer Supported Cooperative Work in Design*. 305–312.
[118] S. Merugu and Joydeep Ghosh. 2003. Privacy-preserving Distributed Clustering using Generative Models. In *Proc. IEEE International Conference on Data Mining*. 211–218.
[119] O. Miguel-Hurtado, S. Stevenage, C. Bevan, and R. Guest. 2016. Predicting Sex as a Soft-biometrics from Device Interaction Swipe Festures. *Pattern Recognition Letters* 79 (2016), 44–51.
[120] I. Mironov, O. Pandey, O. Reingold, and S. Vadhan. 2009. Computational Differential Privacy. In *Proc. International Cryptology Conference*. 126–142.
[121] K. Mivule. 2013. Utilizing Noise Addition for Data Privacy, an Overview. *arXiv* (2013).
[122] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, and S. Marcel. 2016. Keystroke Biometrics Ongoing Competition. *IEEE Access* 4 (2016), 7736–7746.
[123] N. Li, and N. Ti. 2007. T-closeness: Privacy beyond K-anonymity and L-diversity. In *Proc. Conf. on Data Engineering*.
[124] N. Palmius, and A. Tsanas, and K. Saunders, and A. C. Bilderbeck, and J. R. Geddes, and G. M. Goodwin, and M. De Vos. 2016. Detecting Bipolar Depression from Geographic Location Data. *IEEE Transactions on Biomedical Engineering* 64, 8 (2016), 1761–1771.
[125] L. Na, C. Yang, C. Lo, F. Zhao, Y Fukuoka, and A. Aswani. 2018. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. *JAMA Network Open* 1, 8 (2018), e186040–e186040.
[126] K. Nandakumar and A. K. Jain. 2015. Biometric Template Protection: Bridging the Performance Gap between Theory and Practice. *IEEE Signal Processing Magazine* 32, 5 (2015), 88–100.
[127] T .Neal and D. Woodard. 2018. A Gender-specific Behavioral Analysis of Mobile Device Usage Data. In *Proc. International Conference on Identity, Security, and Behavior Analysis*. 1–8.
[128] M. E. Nergiz, M. Atzori, and C. Clifton. 2007. Hiding the Presence of Individuals from Shared Databases. In *Proc. ACM SIGMOD international conference on Management of data*. 665–676.
[129] T. Nguyen, A. Roy, and N. Memon. 2019. Kid on the Phone! Toward Automatic Detection of Children on Mobile Devices. *Computers & Security* 84 (2019), 334–348.
[130] R. Nussbaum, C. Kelly, E. Quinby, A. Mac, B. Parmanto, and B. E. Dicianno. 2019. Systematic Review of Mobile Health Applications in Rehabilitation. *Archives of Physical Medicine and Rehabilitation* 100, 1 (2019), 115–127.
[131] L. O'Gorman. 2003. Comparing Passwords, Tokens, and Biometrics for User Authentication. *Proc. IEEE* 91, 12 (2003), 2021–2040.
[132] P. Kostopoulos, T. Nunes, K. Salvi, M. Togneri and M. Deriaz. 2015. StayActive: An Application for Detecting Stress. In *Proc. International Conference on Communications, Computation, Networks and Technologies*.
[133] R. B. Parker. 1973-1974. A Definition of Privacy. *Rutgers Law Review* 27 (1973-1974), 275.

[134] A. A. Parmar, U. P. Rao, and D. R. Patel. 2011. Blocking Based Approach for Classification Rule Hiding to Preserve the Privacy in Database. In *Proc. International Symposium on Computer Science and Society*. 323–326.
[135] V. M. Patel, N. K. Ratha, and R. Chellappa. 2015. Cancelable Biometrics: A review. *IEEE Signal Processing Magazine* 32, 5 (2015), 54–65.
[136] N. Phan, Y. Wang, X. Wu, and D. Dou. 2016. Differential Privacy Preservation for Deep Auto-Encoders: an Application of Human Behavior Prediction. *Proc. AAAI Conference on Artificial Intelligence* 30, 1 (2016).
[137] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng. 2016. A Survey of Machine Learning for Big Data Processing. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (2016), 1–16.
[138] J. C. Quiroz, E. Geangu, and M. H. Yong. 2018. Emotion Recognition using Smart Watch Sensor Data: Mixed-design Study. *JMIR Mental Health* 5, 3 (2018), e10153.
[139] V. Rastogi, D. Suciu, and S. Hong. 2007. The Boundary Between Privacy and Utility in Data Publishing. In *Proc. International Conference on Very Large Data Bases*. Citeseer, 531–542.
[140] J. Ren, G. Wu, and L. Yao. 2013. A Sensitive Data Aggregation Scheme for Body Sensor Networks based on Data Hiding. *Personal and Ubiquitous Computing* 17, 7 (2013), 1317–1329.
[141] C. Riederer, S. Zimmeck, C. Phanord, A. Chaintreau, and S. Bellovin. 2015. I don't have a photograph, but you can have my footprints. Revealing the Demographics of Location Data. In *Proc. ACM on Conference on Online Social Networks*. 185–195.
[142] S. A. Osia and A. Shahin Shamsabadi and S. Sajadmanesh and A. Taheri and K. Katevas and H. R. Rabiee and N. D. Lane and H. Haddadi. 2020. A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics. *IEEE Internet of Things Journal* 7, 5 (2020), 4505–4518.
[143] S. Majumder and M. J. Deen. 2019. Smartphone Sensors for Health Monitoring and Diagnosis. *Sensors* 19, 9 (2019).
[144] S. Singh, D. M. Shila, and G. Kaiser. 2019. Side Channel Attack on Smartphone Sensors to Infer Gender of the User: Poster Abstract. In *Proc. Conference on Embedded Networked Sensor Systems*.
[145] A. Sabir, H. Maghdid, S. Asaad, M. Ahmed, and A. Asaad. 2019. Gait-based Gender Classification using Smartphone Accelerometer Sensor. In *Proc. International Conference on Frontiers of Signal Processing*. 12–20.
[146] D. Sadhya and S. K. Singh. 2016. Privacy Preservation for Soft Biometrics based Multimodal Recognition System. *Computers & Security* 58 (2016), 160–179.
[147] D. Saha and A. Mukherjee. 2003. Pervasive Computing: a Paradigm for the 21st Century. *Computer* 36, 3 (2003), 25–31.
[148] K. Sampigethaya, L. Huang, M. Li, R. Poovendran, K. Matsuura, and K. Sezaki. 2005. *CARAVAN: Providing Location Privacy for VANET*. Technical Report. Washington Univ Seattle Dept of Electrical Engineering.
[149] D. Santani, T. Do, F. Labhart, S. Landolt, E. Kuntsche, and D. Gatica-Perez. 2018. DrinkSense: Characterizing Youth Drinking Behavior Using Smartphones. *IEEE Transactions on Mobile Computing* 17, 10 (2018), 2279–2292.
[150] M. Santopietro, R. Vera-Rodriguez, R. Guest, A. Morales, and A. Acien. 2020. Assessing the Quality of Swipe Interactions for Mobile Biometric Systems. In *Proc. IEEE International Joint Conference on Biometrics (IJCB)*. 1–8.
[151] L. Scherrer, M. Tomko, P. Ranacher, and R. Weibel. 2018. Travelers or Locals? Identifying Meaningful Sub-populations from Human Movement Data in the Absence of Ground Truth. *EPJ Data Science* 7 (2018), 1–21.
[152] S. Servia-Rodríguez, K. Rachuri, C. Mascolo, P. Rentfrow, N. Lathia, and G. Sandstrom. 2017. Mobile Sensing at the Service of Mental Well-Being: A Large-Scale Longitudinal Study. In *Proc. International Conference on World Wide Web*. 103âĂŞ112.
[153] S. Shah, J. Teja, and S. Bhattacharya. 2015. Towards Affective Touch Interaction: Predicting Mobile User Emotion from Finger Strokes. *Journal of Interaction Science* 3, 1 (2015), 1–15.
[154] C. E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.
[155] R. Shokri, G. Theodorakopoulos, J. Le Boudec, and J. Hubaux. 2011. Quantifying Location Privacy. In *Proc. IEEE symposium on security and privacy*. 247–262.
[156] V. Singh, G. Aggarwal, and BVS. Ujwal. 2018. Ensemble based Real-time Indoor Localization using Stray WiFi Signal. In *Proc. IEEE International Conference on Consumer Electronics (ICCE)*. 1–5.
[157] D. Svantesson and R. Clarke. 2010. Privacy and Consumer Risks in Cloud Computing. *Computer Law & Security Review* 26, 4 (2010), 391–397.
[158] T. Arroyo-Gallego, M. J. Ledesma-Carbayo, A. Sanchez-Ferro, I. Butterworth, C. S. Mendoza, M. Matarazzo, P. Montero, R. Lopez-Blanco, V. Puertas-Martin, R. Trincado, and L. Giancardo. 2017. Detection of Motor Impairment in ParkinsonâĂŹs Disease Via Mobile Touchscreen Typing. *IEEE Trans. Biomed. Eng.* 64, 9 (2017), 1994âĂŞ2002.
[159] T. Neal and S. Canavan. 2020. Mood Versus Identity: Studying the InïňĆuence of Affective States on Mobile Biometrics. In *Proc. IEEE International Conference on Automatic Face and Gesture*.
[160] T. T. Ngo, M. A. R. Ahad, A. D. Antar, M. Ahmed, D. Muramatsu, Y. Makihara, Y. Yagi, S. Inoue, T. Hossain, and Y. Hattori. 2019. OU-ISIR Wearable Sensor-based Gait Challenge: Age and Gender. In *Proc. International Conference on Biometrics*.
[161] I. Tayfur and M. A. Afacan. 2019. Reliability of Smartphone Measurements of Vital Parameters: A Prospective Study using a Reference Method. *The American Journal of Emergency Medicine* 37, 8 (2019), 1527–1530.

[162] M. Templ, A. Kowarik, and B. Meindl. 2015. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software, Articles* 67, 4 (2015), 1–36.
[163] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. 2019. Suppressing Gender and Age in Face Templates using Incremental Variable Elimination. In *Proc. International Conference on Biometrics*. 1–8.
[164] E. Thomaz, I. Essa, and G. D. Abowd. 2015. A Practical Approach for Recognizing Eating Moments with Wrist-mounted Inertial Sensing. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1029–1040.
[165] R. Tolosana, J. C. Ruiz-Garcia, R. Vera-Rodriguez, J. Herreros-Rodriguez, S. Romero-Tapiador, A. Morales, and J. Fierrez. 2021. Child-Computer Interaction: Recent Works, New Dataset, and Age Detection. *arXiv* (2021).
[166] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. 2020. BioTouchPass2: Touchscreen Password Biometrics Using Time-Aligned Recurrent Neural Networks. *IEEE Transactions on Information Forensics and Security* 15 (2020), 2616–2628.
[167] R. Tolosana, R. Vera-Rodriguez, R. Guest, J. Fierrez, and J. Ortega-Garcia. 2020. Exploiting Complexity in Pen-and Touch-based Signature Biometrics. *International Journal on Document Analysis and Recognition* 23, 2 (2020), 129–141.
[168] F. Tramèr and D. Boneh. 2020. Differentially Private Learning Needs Better Features (or Much More Data). *arXiv* (2020).
[169] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou. 2019. A Hybrid Approach to Privacy-Preserving Federated Learning. In *Proc. ACM Workshop on Artificial Intelligence and Security*. 1âĂŞ11.
[170] V. M. Patel, R. Chellappa, D. Chandra, and B. Barbello. 2016. Continuous User Authentication on Mobile Devices: Recent Progress and Remaining Challenges. *IEEE Signal Process. Mag.* 13, 4 (2016), 49âĂŞ61.
[171] V. Mirjalili, and S. Raschka, and A. Namboodiri, and A. Ross. 2018. Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In *Proc. International Conference on Biometrics*. 82–89.
[172] V. Mirjalili, and S. Raschka, and A. Ross. 2020. PrivacyNet: Semi-Adversarial Networks for Multi-attribute Face Privacy. *IEEE Transactions on Image Processing* 29 (2020), 9400–9412.
[173] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis. 2004. State-of-the-Art in Privacy Preserving Data Mining. *SIGMOD Rec.* 33, 1 (2004), 50âĂŞ57.
[174] S. De Capitani Di Vimercati, S. Foresti, G. Livraga, and P. amarati. 2012. Data Privacy: Definitions and Techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 20, 06 (2012), 793–817.
[175] I. Wagner and D. Eckhoff. 2018. Technical Privacy Metrics: a Systematic Survey. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 1–38.
[176] Changsheng Wan, Li Wang, and Vir V. Phoha. 2018. A Survey on Gait Recognition. *ACM Comput. Surv.* 51, 5, Article 89 (Aug. 2018), 35 pages. https://doi.org/10.1145/3230633
[177] N. Wan and G. Lin. 2016. Classifying Human Activity Patterns from Smartphone Collected GPS Data: A Fuzzy Classification and Aggregation Approach. *Transactions in GIS* 20, 6 (2016), 869–886.
[178] J. Wieringa, P.K. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera. 2021. Data Analytics in a Privacy-concerned World. *Journal of Business Research* 122 (2021), 915–925.
[179] R. C. Wong, J. Li, A. W. Fu, and K. Wang. 2006. (α, k)-Anonymity: an Enhanced k-Anonymity Model for Privacy Preserving Data Publishing. In *Proc. ACM SIGKDD international conference on Knowledge discovery and data mining*. 754–759.
[180] L. Wu, L. Yang, Z. Huang, Y. Wang, Y. Chai, X. Peng, and Y. Liu. 2019. Inferring Demographics from Human Trajectories and Geographical Context. *Computers, Environment and Urban Systems* 77 (2019), 101368.
[181] x. Xiao and y. Tao. 2007. M-invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets. In *Proc. ACM SIGMOD international conference on Management of data*. 689–700.
[182] Y. Ma, and S. Arshad, and S. Muniraju, and E. Torkildson, and E. Rantala, and K. Doppler, and G. Zhou. 2021. Location-and Person-Independent Activity Recognition with WiFi, Deep Neural Networks, and Reinforcement Learning. *ACM Transactions on Internet of Things* 2, 1 (2021), 1–25.
[183] Y. Yao, L. Song, and J. Ye. 2020. Motion-To-BMI: Using Motion Sensors to Predict the Body Mass Index of Smartphone Users. *Sensors* 20, 4 (2020), 1134.
[184] M. Yang, T. Zhu, Y. Xiang, and W. Zhou. 2018. Density-Based Location Preservation for Mobile Crowdsensing With Differential Privacy. *IEEE Access* 6 (2018), 14779–14789.
[185] Q. Yang, Y. Liu, T. Chen, and Y. Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.
[186] Y. Yuan, M. Raubal, and Y. Liu. 2012. Correlating Mobile Phone Usage and Travel Behavior–A Case Study of Harbin, China. *Computers, Environment and Urban Systems* 36, 2 (2012), 118–130.
[187] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. 2007. Aggregate Query Answering on Anonymized Tables. In *Proc. International Conference on Data Engineering*. 116–125.
[188] X. Zhang, W. Li, X. Chen, and S. Lu. 2018. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (2018).