

Evaluating the Impact of Data Anonymization in a Machine Learning Application

Lelio Campanile[1], Fabio Forgione[2], Michele Mastroianni[3], Gianfranco Palmiero[2], and Carlo Sanghez[2(✉)]

[1] Dip. Matematica e Fisica, Università degli Studi della Campania "L. Vanvitelli", Caserta, Italy
lelio.campanile@unicampania.it
[2] GAM Engineering srlu, Caserta, Italy
{fforgione,gpalmiero,csanghez}@gameng.it
[3] Università degli Studi di Salerno, Salerno, Italy
michele.mastroianni@unisa.it
https://www.gameng.it

Abstract. The data protection impact assessment is used to verify the necessity, proportionality and risks of data processing. Our work is based on the data processed by the technical support of a Wireless Service Provider. The team of WISP tech support uses a machine learning system to predict failures. The goal of our the experiments was to evaluate the DPIA with personal data and without personal data. In fact, in a first scenario, the experiments were conducted using a machine learning application powered by non-anonymous personal data. Instead in the second scenario, the data was anonymized before feeding the machine learning system.

In this article we evaluate how much the Data Protection Impact Assessment changes when moving from a scenario with raw data to a scenario with anonymized data.

Keywords: Risks · DPIA · Privacy · WISP · Unwanted modification of data · Illegitimate access to data · Data disappearance · GDPR

1 Introduction

The data protection impact assessment (DPIA) is a procedure aimed at describing the processing, assessing its necessity and proportionality, and facilitating the management of risks to the rights and freedoms of individuals deriving from the processing of their personal data. There are some contexts in which it is necessary to make use of personal data in order to provide a service.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13380, pp. 389–400, 2022.
https://doi.org/10.1007/978-3-031-10542-5_27

Article 35 of the GDPR (General Data Protection Regulation) prescribes that *Where a type of processing in particular using new technologies..... is likely to result in a high risk to the rights and freedoms of natural persons* the controller shall carry out a Data Protection Impact Assessment (DPIA), which aims to conduct a systematic risk assessment in order to identify privacy threats and impose technical and organizational controls to mitigate those threats. Article 35 prescribes that DPIA shall in particular be required in the case of:

- a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
- processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offenses referred to in Article 10;
- systematic monitoring of a publicly accessible area on a large scale.

Consider the case of a wireless internet service provider that wants to predict the failures of customer equipment: there is no way to avoid collecting customer personal data. In fact, the technical support of a wireless service provider normally has access to the MAC address of the customer's device, the geolocalization of the customer's device, personal and administrative data of individuals, but also, theoretically, the time slots of greatest use of the network, the favorite streaming service providers, the device most bandwidth consuming, etc. Technical support has all the data to profile customers or to know if a customer is home or not. In short, the risks for individuals are high.

The main purpose of technical support of a wireless service provider is to offer the best possible connection. The biggest challenge here is to predict device failures using both field data such as weather conditions, wind, humidity and personal data such as customer preferred streaming services, location of radio antennas, number of device connected etc.

Industry best practices indicate a need to take action before a blocking radio failure, but to take action sooner you need to somehow predict failure. Furthermore, best practices indicate two ways to reduce the risk: data anonymization and data encryption.

Failure prediction systems that rely on machine learning are already available, but the problem here is that these systems need access to personal data. Such systems are based on machine learning applications need personal data during training and they need personal data during the prediction phase.

However, there are solutions that mitigate the risk through data anonymization, data disruption or encryption, but performance in terms of failure prediction degrades significantly. Preservation of privacy in data mining and machine learning is the absolute prerequisite in the above scenarios.

This paper is the continuation of the research begun with [3], and the actual work is to assess the risk in a scenario where personal data was neither anonymized nor encrypted. Next, we measure the risk of the same scenario after applying different anonymization and encryption techniques on the data.

The research questions the paper aims to answer are:

RQ1: is there an effective reduction in risk for individuals?
RQ2: among the methods of anonymization, which is the one that guarantees the best performance in terms of failure prediction with moderate risk for individuals?

The goals of this research are to perform a risk assessment in order to reduce illegitimate access to data and to reduce unwanted data changes.

This paper is organized as follows: the Sect. 2 presents related works, in the Sect. 3 is presented the methodology we use in order to ensure pseudonimization and anonymization of data; in Sect. 4 is described the case study, while in Sect. 5 are discussed the results of the DPIA analysis. Section 6 concludes the paper.

2 Related Works

In [1] are highlighted threats and vulnerabilities of the emerging technologies (IoT - Internet of Things, Cloud Computing, and AI - Artificial Intelligence), as well as the techniques of risk mitigation as stated in the outlining standards and regulations. A comparison was then made between the three technologies and the international standards (OWASP - Open Web Application Security Project, NIST - National Institute of Standards and Technology, ISO - International Organization for Standardization, and GDPR). The study shows that, while in the case of cloud technology the standards are able to effectively support the developers in order to obtain compliance, for the case of IoT and AI there's still a lot of work to be done on standards. An interesting systematic literature review is [6], in which 159 research paper have been analyzed and a thematic analysis to all identified risks has been performed and ten noticeable DPIA assessment methodologies have been identified.

Some authors use the methodologies suggested by National Authorities to perform privacy assessment. The work of Georgiou et al. [7] deals with DPIA performing in health sector, using the CNIL-PIA methodology [5] proposed by the French privacy authority (Commission Nationale de l'Informatique et des Libertés - CNIL). This paper focuses on the first steps of the methodology, and in particular about the purpose of processing and the evaluation of the organization's GDPR compliance level and of the gaps that must be filled-in (Gap Analysis). Also in [2] is used the CNIL-PIA methodology to perform a risk assessment analysis related to the use of pseudonimization in blockchain.

In this paper, is used the tool PIA created by CNIL to evaluate the risks and their related mitigation techniques. In [8] in order to perform DPIA, is used the guideline redacted by Information Commissioner's Office (ICO) [10], the UK independent authority set up to uphold the information rights for public interest.

On the other side, other authors developed ad-hoc methodologies and tools to deal with DPIA. In [13], a methodology is proposed to perform DPIA for healthcare information systems, due to the specific constraints posed by the heterogeneous and highly sensitive nature of data and software use in hospitals. The methodology proposed uses as the starting point the Record Of Processing Activities (ROPA), stated as mandatory by Art. 30 of GDPR, and a specialized tool to perform DPIA has been developed. Also in [11] the ROPA is seen as the starting point to perform risk assessment, and the methodology proposed is based on ontologies and use of AI tools. In [12] is proposed a framework specialized in the context of Public Administration information to perform assessment.

3 Methodology

Our methodology was based on the Data Protection Impact Assessment. The DPIA is a process designed to identify risks arising out of the processing of personal data and to minimize these risks as far and as early as possible [4]. Although the DPIA is often used to demonstrate GDPR compliance, in our work we used that process to check how good the data anonymization solutions were.

The first step was to perform a preliminary Data Protection Impact Assessment on the raw data. In our case the raw data included a data set with 43,780 rows x 25 columns.

The columns included personal data:

- MAC address of the Customer Premise Equipment
- Customer ID
- name of Customer
- ZIP code of costumer
- geo-localization of radio device in customer's house

The second step was to mitigate the risks related to data breach. For this purpose, We used several good practices in terms of data obfuscation [9]:

- **Attribute suppression**: refers to the removal of the personal part from data. We simply deleted the columns with personal data from the data set.
- **Pseudo-anonymization by HASH**: refers to the replacement of personal data with anonymous data. We simply used a hash function on strings containing personal data. The new value was used instead of the old one.
- **Generalization**: refers to the converting personal data in an range of values. We simply grouped personal values in order to keep the rough information, but accepting the loss of detail.

– **Generalization and k-anonymity**: refers to the converting data in an range of values as seen before, but in this case we deliberately deleted those samples that had no other k-similar. We did not use K-anonymization due to the size of our data set.

In order to achieve an anonymous data set from the raw data set using the attribute suppression, one possible choice is to delete all columns that include personal data from the data set. As an alternative, is possible to anonymize the data set replacing all characters of the columns containing personal data with the 'x' character.

Another technique to mitigate risks, is the pseudo-anonymization of the columns containing personal data, applying a hash function directly to those data.

Finally, another useful technique is using the generalization. In order to operate in that way, it is possible to eliminate the last octet present in the 'MAC' column, and group the values present in the 'CUSTOMER CODE' column into groups of 500, eliminating the last two digits from the ZIP code present in raw data and considering only the first letter present in the values of the 'DEVICE NAME' column.

The raw data set was represented by 25 columns. Of these 25 columns, some contain personal data: the customer code and the device name of the customer, ZIP code of the customer, geo-localization of the customer's device, MAC address of the router and MAC of the customer's access point, name and surname of the customer, etc. The source of the data was represented by about 2000 radio devices connected by radio frequency.

We evaluated the performance in terms of predicting failures and in terms of the time required to process data from the devices. Obviously the machine learning model provides the best performance with the raw data set.

The best performance using the anonymous data set was obtained by applying hashing techniques, but the code is too much time consuming. The point here is that the anonymization of raw data is done on small Linux servers with low-power ARM processors.

In order to improve the performance of the fault prediction system, then, we decided to spread the computational load across multiple ARM servers. Consequently, to distribute the load between the servers running the machine learning application, we have to send the data to the servers and sending the data to the servers increases the likelihood that the data is somehow corrupted, stolen, modified, and so on.

The idea of using more powerful ARM processors was ruled out due to the higher impact in terms of power consumption of the solution.

In the case study, the focus was to achieve DPIA from raw data set and DPIA from the modificated data set.

4 Case Study

In this section we describe the case study on which is based our work.

Before continuing, it is advisable to clearly distinguish the problem to be addressed, the testbed and the expected results for each step of the methodology.

The problem to be addressed is the processing of personal data done by the fault prediction system based on artificial intelligence.

The testbed was the machine learning system and the personal data coming from Flyber's customers.

The methodology involved first evaluating the DPIA with personal data and then evaluating the DPIA with anonymous data. The expected results were to have a significant reduction in the risk of compromising user privacy.

Flyber is a Wireless Internet Service Provider (WISP) based on Teverola, Italy, that offers its customers free technical support to solve problems related to the quality of signal reception. Since the technology is wireless over open bands, the quality of the connection depends mainly on radio interference.

To check the health status of the Customer Premise Equipments (CPE), Flyber adopts a monitoring system that produces a daily report with 25 fields. The monitoring system produces information on a lot of personal data: MAC Address of the customer's device, GPS coordinates of the customer's device, customer name, etc.

Internet Service Providers technical support needs personal data to predict potential anomalies, there is no way to gather this stuff. To monitor radio quality and to prevent failures Flyber uses a machine learning application.

To avoid the disclosure of personal data, the machine learning application uses an anonymous data set. There are several aspects that determines the need of DPIA: systematic monitoring, evaluationg and predicting internet usage, data processed on a large scale, etc. Because of the above criteria, DPIA must be required.

The DPIA allowed us to assess the risk to individuals in the event of a data compromise, the impact of the data compromise, and the likelihood of adverse events. It should be noted that Flyber already used a fault monitoring system based on artificial intelligence [3].

The first version of the fault monitoring system included a single ARM Linux Appliance capable to manage data from more than 1000 radio devices. In the work [3], it is already explored the impact of adopting anonymization in a real-word application, but in that case our experiments were focused on computing performance and accuracy and computing performance was limited to the single ARM Linux server.

In order to boost performance we deployed a new architecture based on multiple computational devices. Now the updated fault monitoring system includes a cluster of four ARM Linux Appliance capable to manage data from more than 2000 radio devices for 0,5 GB of data per second. Because more and more ARM Linux server and because the need to transfer data outside the single system, the likelihood of risk got higher.

5 Results

This section shows the privacy risks, taking into account existing or planned controls. The diagrams allow us to have a global and synthetic view on the controls effects on the risks they address. This section shows risk mapping as well in order to compare the positioning of the risks hem, before and after application of the complementary controls.

In order to perform the risk assessment, we use the PIA tools developed by CNIL, the French data protection authority.

The fault prediction procedure using raw data set is divided into three steps:

- The first step is to save the raw data on the monitoring server.
- The second step is to send the raw customer data to the machine learning application.
- The third step is fault prediction through the machine learning model.

If the prediction system indicated a problem, then the system sent an e-mail to technical support with the data of the problem encountered and opened a ticket.

According to definition "a threat is a *Procedure comprising one or more individual actions on data supporting assets*", in our work we considered the following threats:

- Saving the raw data on the server.
- Sending the raw data to the AI application.
- Feeding and running the machine model.

The risk overview Fig. 1, produced by PIA, shows the relationships between impacts, threats, sources of threats and security measures implemented at this stage. We assessed the levels of risk and likelihood of the events according to the methodology of the French Data Protection Authority (CNIL).

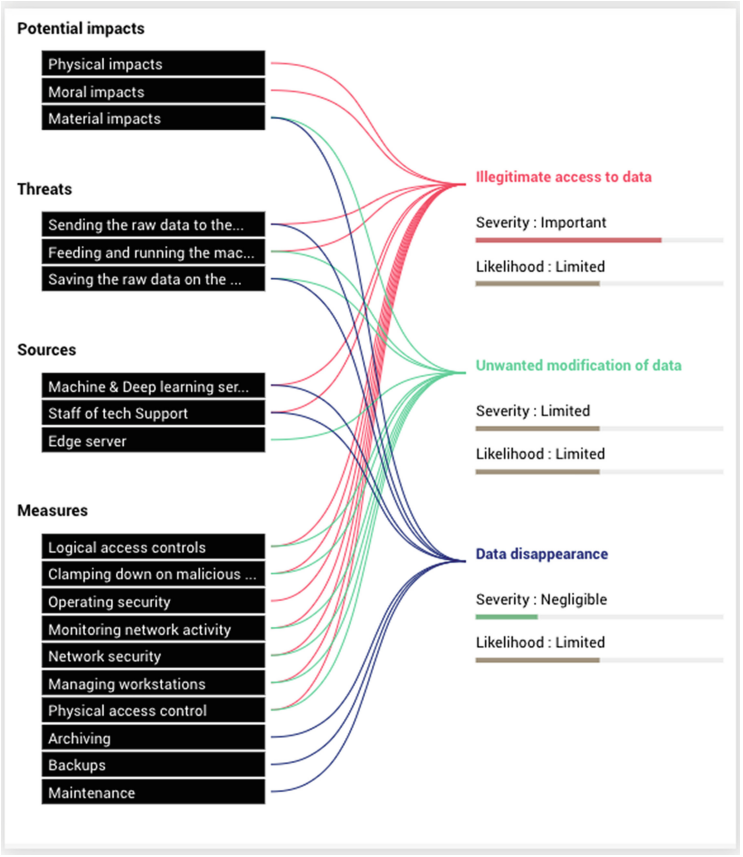

Fig. 1. Risk overview with raw data.

The framework comprises three guides [5]:

- the first guide sets out the approach,
- the second guide contains facts that could be used for formalizing the analysis,
- the third guide provides knowledge bases. This guide is a catalog of controls aimed at complying with the legal requirements and treating the risks.

We have come to the following conclusions to according to the above knowledge bases.

Unwanted Modification of Data
In the event of unwanted modification of data, we consider the risk **limited** in relationship to material impacts (outdated data or loss of access to temporary services), while on the moral side we have not assessed any particular critical issues.

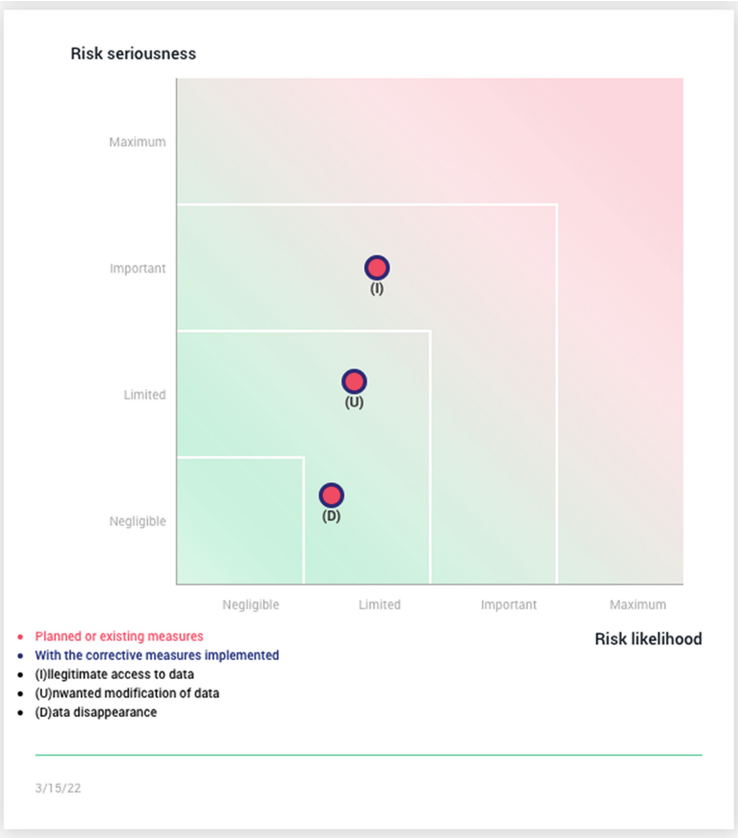

Fig. 2. Risk mapping with raw data.

Illegitimate Access to Data
In the event of illegitimate access to data, we have considered the severity level of the risk to be **important** due to potential property damage/theft.

Data Disappearance
In the event of Data disappearance, we considered the risk level to be **Negligible**.

We considered the likelihood of threats **limited**: the data are stored on servers suitably protected by both logical and physical controls.

In the Fig. 1) are shown the risk overview in case of raw data without anonymization. In the Fig. 2 is shown the risk mapping without encryption.

The fault prediction procedure using anonymous data set is divided into four steps:

- The first step is to save the raw data on the edge ARM Linux server.
- The second step is to delete any personal data from the raw data on the edge ARM Linux server.

– The third step is to send the data set to the machine learning application.
– The final step is fault prediction through the machine learning model using only anonymous data set.

In order to reduce risk likelihood in terms of illegitimate access to data and unwanted modification of data we used several methods of Data anonymization.

As a consequence of deleting personal data or hashing personal data we have significantly reduced the risk moving from a scenario where risk was important to a **limited** risky scenario.

In Fig. 3 is shown the risk assessment in case of raw data with anonymization. In Fig. 4 is shown the risk mapping with encryption.

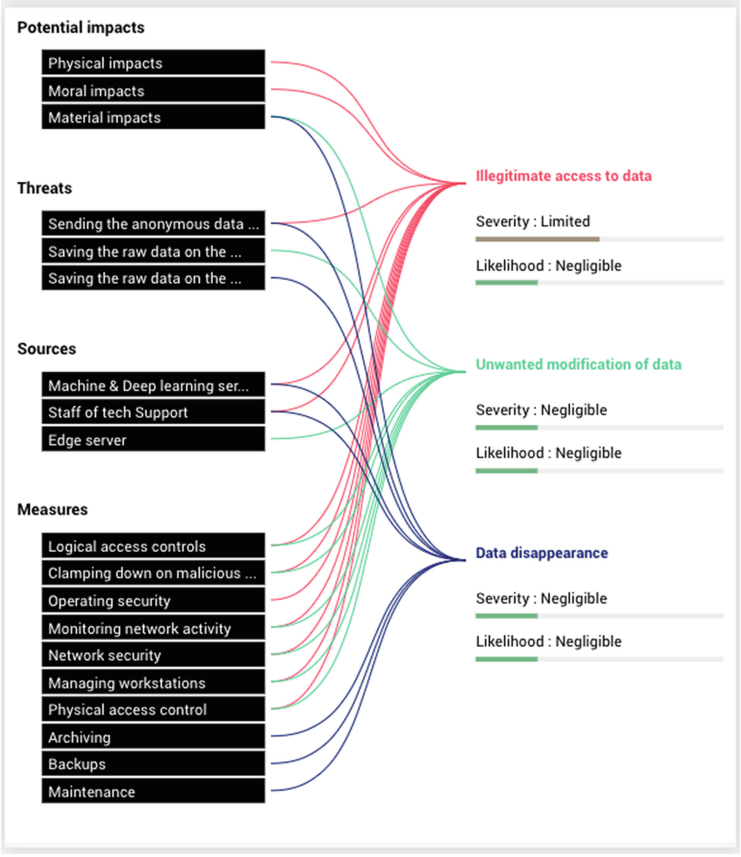

Fig. 3. Risk overview with anonymous data.

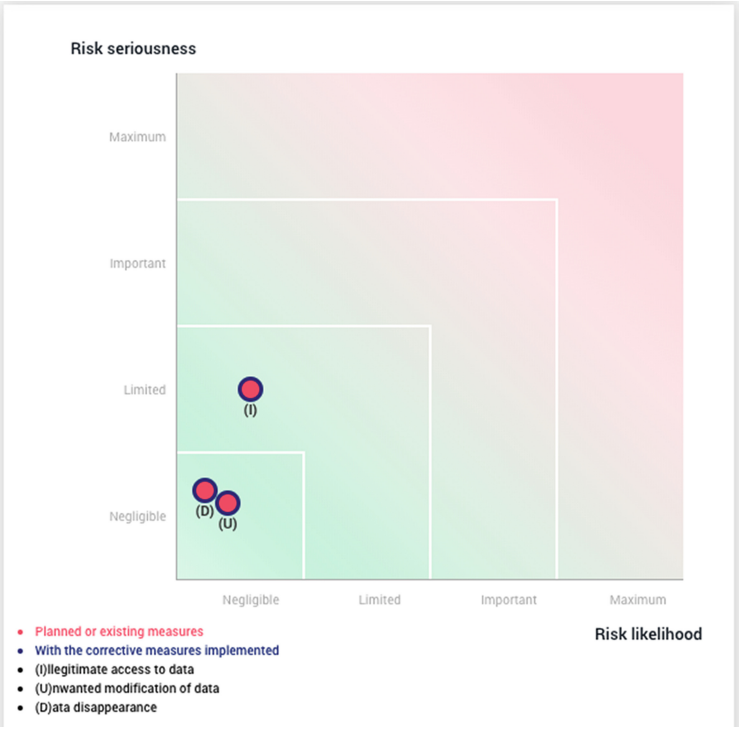

Fig. 4. Risk mapping with anonymous data.

6 Conclusion and Future Works

In this study has been studied the impact of obfuscation techniques on personal data in order to mitigate the data breaches risks on personal data. The CNIL-PIA methodology on privacy risk assessment, and the related tool PIA, has been used to show the impact of those techniques and to show which could be the better way to obtain GDPR compliance in personal data used by learning machine tools to perform technical activities. The results show that some techniques are really effective in that direction. In future work, we aim to explore with better detail the obfuscation techniques.

References

1. Attar, R.A., Al-Nemri, J., Homsi, A., Qusef, A.: Risk assessment for emerging domains (IOT, cloud computing, and AI). In: 2021 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 120–127 (2021). https://doi.org/10.1109/JEEIT53412.2021.9634156

2. Campanile, L., Cantiello, P., Iacono, M., Marulli, F., Mastroianni, M.: Risk analysis of a GDPR-compliant deletion technique for consortium blockchains based on pseudonymization. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12956, pp. 3–14. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87010-2_1
3. Campanile, L., Forgione, F., Marulli, F., Palmiero, G., Sanghez, C.: Dataset anonimyzation for machine learning: an ISP case study. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12950, pp. 589–597. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86960-1_42
4. Cantiello, P., Mastroianni, M., Rak, M.: A conceptual model for the general data protection regulation. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12956, pp. 60–77. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87010-2_5
5. French Data Protection Authority (CNIL). Privacy Impact Assessment (PIA) - Knowledge Bases (2018). https://www.cnil.fr/sites/default/files/atoms/files/cnil-pia-3-en-knowledgebases.pdf
6. Georgiadis, G., Poels, G.: Towards a privacy impact assessment methodology to support the requirements of the general data protection regulation in a big data analytics context: a systematic literature review. Comput. Law Secur. Rev. **44**, 105640 (2022) https://doi.org/10.1016/j.clsr.2021.105640
7. Georgiou, D., Lambrinoudakis, C.: Data protection impact assessment (DPIA) for cloud-based health organizations. Future Internet **13**(3) (2021). https://doi.org/10.3390/fi13030066
8. Henriksen-Bulmer, J., Faily, S., Jeary, S.: DPIA in context: applying DPIA to assess privacy risks of cyber physical systems. Future Internet **12**(5) (2020). https://doi.org/10.3390/fi12050093
9. Hosseinzadeh, S., et al.: Diversification and obfuscation techniques for software security: a systematic literature review. Inf. Softw. Technol. **104**, 72–93 (2018) https://doi.org/10.1016/j.infsof.2018.07.007
10. Information Commissioner's Office (ICO). Data protection impact assessments (2018). https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/
11. Di Martino, B., Mastroianni, M., Campaiola, M., Morelli, G., Sparaco, E.: Semantic techniques for validation of GDPR compliance of business processes. In: Barolli, L., Hussain, F.K., Ikeda, M. (eds.) CISIS 2019. AISC, vol. 993, pp. 847–855. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-22354-0_78
12. Szczepaniuk, E.K., Szczepaniuk, H., Rokicki, T., Klepacki, B.: Information security assessment in public administration. Comput. Secur. **90**, 101709 (2020) https://doi.org/10.1016/j.cose.2019.101709
13. Todde, M., Beltrame, M., Marceglia, S., Spagno, C.: Methodology and workflow to perform the data protection impact assessment in healthcare information systems. Inf. Med. Unlocked **19**, 100361 (2020) https://doi.org/10.1016/j.imu.2020.100361