network

MDPI

Article

Data Protection Issues in Automated Decision-Making Systems Based on Machine Learning: Research Challenges

Paraskevi Christodoulou [1] **and Konstantinos Limniotis** [1,2,*]

[1] School of Pure and Applied Sciences, Open University of Cyprus, Latsia, Nicosia 2220, Cyprus; paraskevi.christodoulou@st.ouc.ac.cy
[2] Hellenic Data Protection Authority, Kifissias 1-3, 11523 Athens, Greece
* Correspondence: konstantinos.limniotis@ouc.ac.cy or klimniotis@dpa.gr

Abstract: Data protection issues stemming from the use of machine learning algorithms that are used in automated decision-making systems are discussed in this paper. More precisely, the main challenges in this area are presented, putting emphasis on how important it is to simultaneously ensure the accuracy of the algorithms as well as privacy and personal data protection for the individuals whose data are used for training the corresponding models. In this respect, we also discuss how specific well-known data protection attacks that can be mounted in processes based on such algorithms are associated with a lack of specific legal safeguards; to this end, the General Data Protection Regulation (GDPR) is used as the basis for our evaluation. In relation to these attacks, some important privacy-enhancing techniques in this field are also surveyed. Moreover, focusing explicitly on deep learning algorithms as a type of machine learning algorithm, we further elaborate on one such privacy-enhancing technique, namely, the application of differential privacy to the training dataset. In this respect, we present, through an extensive set of experiments, the main difficulties that occur if one needs to demonstrate that such a privacy-enhancing technique is, indeed, sufficient to mitigate all the risks for the fundamental rights of individuals. More precisely, although we manage—by the proper configuration of several algorithms' parameters—to achieve accuracy at about 90% for specific privacy thresholds, it becomes evident that even these values for accuracy and privacy may be unacceptable if a deep learning algorithm is to be used for making decisions concerning individuals. The paper concludes with a discussion of the current challenges and future steps, both from a legal as well as from a technical perspective.

Keywords: deep learning algorithm; differential privacy; GDPR; impact assessment

Citation: Christodoulou, P.; Limniotis, K. Data Protection Issues in Automated Decision-Making Systems Based on Machine Learning: Research Challenges. *Network* **2024**, *4*, 91–113. https://doi.org/10.3390/network4010005

Academic Editors: Andreas Kassler and Luis Alonso

Received: 22 October 2023
Revised: 15 January 2024
Accepted: 26 February 2024
Published: 1 March 2024

1. Introduction

Machine learning (ML) and, especially, deep learning algorithms, as a type of artificial intelligence (AI) algorithms, are widely used in various fields, such as digital image processing [1], data analytics [2], autonomous systems [3], text and speech recognition [4], face recognition [5], robotics [6], traffic prediction [7], intrusion detection [8], etc. It is still a highly evolving field with a variety of innovative applications [9,10]. However, when these algorithms are used in automated decision-making systems concerning individuals, several risks of a high severity occur with regard to human rights; for example, if such a system is used to decide whether a person, who is a candidate employee for a specific job, fits well with the job in question, its decision highly affects this person, and an erroneously negative response from this system yields a significant impact for the individual, posing restrictions on one's rights.

More generally, errors in the output of an automated decision-making system may have serious consequences for individuals, such as [11] (a) legal effects, e.g., the denial of a particular social benefit, such as child or housing benefit, refusal of entry at the border, extensive supervision by authorities that should not have taken place, etc., and

(b) other significant effects, e.g., the automatic refusal of an online credit application, intrusive profiling for various purposes, etc. Although several legal frameworks, such as the prominent General Data Protection Regulation (GPDR) in Europe [12], do not allow decisions that significantly affect individuals to rely solely on automated decision-making systems (i.e., human intervention is necessary before the final decision, even if such systems are used) (see Art. 22 of the GDPR), it is essential—even in these cases—to ensure that the underlying ML algorithms provide unbiased and accurate results to the greatest possible extent.

Moreover, such systems also yield several challenging issues for the rights of privacy and personal data protection [13,14]. Indeed, making a decision about an individual necessitates the processing of one's personal data, and this entails risks with regard to the aforementioned rights if no sufficient safeguards are in place, such as transparency of the overall processing (including information about the underlying logic involved), as well as the informed consent of individuals (although this may not be applicable when decision-making is authorized by law). Apart from this aspect, one should also consider that ML algorithms are constantly trained according to a specific training dataset; if this dataset is based on personal data, further risks occur with respect to the individuals related to these data. More precisely, it may be questionable whether these individuals are aware of this processing, as well as whether these data are used only for training and not for other purposes; additionally, these data should not be more than are needed for training the model, and this, in turn, typically necessitates the requirement that these data should not allow identification of the corresponding individuals (since such an identification is not necessary for the training purposes and would yield "excessive" processing). Hence, it seems that using "anonymised" data as the training set is a prerequisite to address such data protection and privacy risks (which could involve either direct anonymization methods, such as differential privacy (see, e.g., [15]) or other advanced cryptographic techniques that suffice to perform operations on "hidden" data, such as homomorphic encryption or secure computations (see, e.g., [13,16])). However, by these means, the accuracy of the ML algorithm may be affected since "anonymisation" as a process is related to a type of modification of data, and thus, due to this modification, one could say that there is a trade-off between privacy and accuracy that needs to be efficiently accommodated.

It should be clarified that automated decision-making systems may not necessarily involve ML techniques or other types of AI; see, e.g., [17]. Such systems may also lead to data protection risks [11]. However, this paper focuses explicitly on systems based on ML (with emphasis on deep learning techniques).

1.1. Research Objectives and Methodology

This paper aims to discuss the main challenges and risks, from a data protection point of view, when automated decision-making systems based on ML systems are employed, thus providing the necessary background as a survey and paving the way for subsequent research that is needed. To understand the data protection risks, the paper will be based on the legal provisions stemming from the European legal framework, and, especially, the GDPR, which, as it has been stated (see, e.g., [18]), can serve as a useful model for other regulations to follow in terms of rights and principles. More precisely, this paper aims to address the following research objectives:

1. A classification and description of the privacy and data protection threats when using ML algorithms for automated decision-making systems, associating each of them with the nonfulfillment of specific legal provisions;
2. A review of the data protection engineering techniques that have been proposed to alleviate the aforementioned privacy and data protection threats;
3. Further investigation of the application of differential privacy (DP) techniques to the training dataset via exploring the effect that each parameter of an algorithm has when evaluating both the accuracy of the output as well as the privacy achieved for the training dataset.

On the basis of the above, the methodology and the contributions of this paper are as follows:

(a) First, important results in the field of data protection attacks on ML systems are surveyed in conjunction with relevant data protection engineering techniques that have been proposed. In this respect, for each known privacy threat in ML systems, we associate the relevant provision of the GDPR that seems to not be in place, thus establishing direct connections on how the nonfulfillment of legal requirements yields specific weaknesses (allowing effective privacy attacks) in practice;
(b) Additionally, based on the work in [15] that applies DP to the training dataset of deep learning algorithms, some new results are also given, based on a set of extensive experiments relying on the above work that we carried out, indicating that there is still much room for further research since we, indeed, manage to achieve better accuracies for the algorithms by appropriately configuring several hyperparameters. However, we also address the challenge of what should be considered "acceptable" accuracy when we refer to decision-making systems concerning individuals.

Our preliminary analysis reveals that the research outcomes need to be meticulously considered by the relevant stakeholders when ML systems are deployed, i.e., the corresponding algorithmic impact assessments should be conducted based on deep knowledge of whether and how State-of-the-Art privacy-enhancing technologies, indeed, suffice to alleviate data protection risks. The paper concludes with future steps that are suggested concerning both research as well as governance aspects.

1.2. Structure of the Paper

The paper is organized as follows: Section 2 presents the background on automated decision-making systems in general, illustrating—based on some well-known incidents—how severe the impact can be on individuals if the system's output is not correct, thus establishing the importance and the inherent difficulty of this issue (see Section 2.1). Moreover, it presents the necessary legal background on personal data protection based on the corresponding framework in Europe (see Section 2.2), whilst it concludes with a summary of the main challenges. Section 3 briefly surveys the main data protection threats for automated decision-making systems that are based on ML techniques concerning privacy attacks related to the training dataset; for each such threat, the relevant association with the legal provision is provided in terms of illustrating how the nonfulfillment of a specific legal provision gives rise to each threat. Subsequently, Section 4 briefly surveys the main techniques that can be used to alleviate these data protection threats. Subsequently, in Section 5, emphasis is placed on analyzing the application of differential privacy on the training dataset, which consists of one of the main data protection engineering techniques, by further elaborating on the results in [15] by executing a more extensive set of experiments. Finally, the overall discussion points and concluding remarks are given in Sections 6 and 7, respectively.

2. Background

2.1. Automated Decision-Making Systems and Relevant Risks for Fundamental Rights

Many automated decision-making systems operate in a nontransparent manner (see, e.g., [19]) without giving the opportunity for human intervention if needed. For example, a citizen's application to rent a house can be rejected by an automated decision-making system, which is related to the process of screening tenants who are suitable to rent a house without the candidate ever having knowledge of the reason for this rejection. Moreover, it is also known that governments regularly adopt automated decision-making systems without public knowledge [20].

In 2014, a tool implementing an algorithm that could review resumes and determine which applicants a well-known company should bring on board was abandoned since it turned out that the tool systematically discriminated against women applying for technical jobs [21]. According to [20], thousands of American citizens have lost access to social

benefits as a result of "bugs" in the relevant source code since translating complex and often ambiguous regulatory requirements into computer code is not an easy task; for example, such errors resulted in hundreds of thousands of erroneous decisions, including improper denials of health care to pregnant women and women with breast and cervical cancer, as well as improper denials of food stamps to the disabled. More recently, the Dutch tax authority had been wrongly accusing families of fraud, forcing them to repay benefits, due to the discriminatory problems arising from an algorithm in a decision-making system [22]. Another important factor that affects an algorithm's bias is the "quality" of the datasets that feed it since they may reflect structural inequities, which, in turn, may yield discrimination [23]. An overview of the ways big data can harm marginalized communities is given in [24]. For example, a commercial healthcare algorithm was identifying specific-type patients based on their ethnicity for more intensive medical care than other patients with similar health issues [25], and this occurred due to the fact that the algorithm was based on data about past healthcare expenditures in order to make predictions about patients' future need. This past information resulted in misconceptions since some individuals suffered numerous logistical, institutional, and cultural barriers to healthcare access and thus had lower cost histories [20]; a proper modification of the algorithm in order to exclude past expenditures as a factor for the decision process sufficed to eliminate this algorithm's biases on its outcomes [20]. More generally, there are several known cases of disabled people being denied desperately needed state support due to an algorithmic decision (see [20] and the references therein). As stated by Philip Alston, a UN Special Rapporteur on extreme poverty and human rights, we have entered a digital dystopia in which "systems of social protection and assistance are increasingly driven by digital data and technologies that are used to automate, predict, identify, surveil, detect, target and punish" [20].

More generally, ML techniques can be used in various critical applications involving decision-making, which include, amongst others, applications in the medical/health sector (see, e.g., [26–28]).

Therefore, it becomes essential that decisions made by automated decision-making systems should also be checked/verified by a human to avoid erroneous outcomes (e.g., approvals or rejections of applications). This conclusion is further accentuated by the fact that entities (i.e., public or private organizations) that are provided with such algorithmic software may not have a deep knowledge of how the underlying algorithms actually work, and they do not implement mechanisms for regularly monitoring their effectiveness. The need for human intervention is provisioned in the GDPR, as described next.

2.2. Personal Data Protection—Legal Provisions

The right to privacy has been recognized as a fundamental human right by the United Nations Declaration of Human Rights [29], the International Covenant on Civil and Political Rights [30], the Charter of Fundamental Rights in the European Union [31], and other international treaties. Privacy is strongly related to personal data protection; as stated in the Charter of Fundamental Rights, personal data "*must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law*" (see Article 8(2) of the Charter).

In Europe, as stated above, the main legal instrument for personal data protection is the General Data Protection Regulation (GDPR). The GDPR also applies to the processing of personal data by organizations not established in the European Union in the case of when they process the data of individuals in the Union, and the processing activities are related to either "*the offering of goods or services*" or "*the monitoring of their behavior as far as their behavior takes place within the European Union*" (see Art. 3, par. 2 of the GDPR). As explicitly stated in [32], the intentionally global reach of the GDPR, with the threat of the huge fines it sets if fundamental rights are not properly protected, has led companies around the world to amend their privacy practices, as well as countries around the world to update their privacy laws.

According to the GDPR (see Art. 4(1)), the term *personal data* refers to *any information relating to an identified or identifiable natural person*, that is a person who can be identified (referred to as a *data subject*); as explicitly stated in the same provision (i.e., Art. 4(1)) in the GDPR, an identifiable natural person *is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*. It should be pointed out that this definition goes beyond the well-known definition of personally identifiable information (PII), which is generally considered information that can be used to distinguish or trace an individual's identity; all PII is considered personal data, but the reverse does not hold (see, e.g., [33]). *Personal data processing* means *any operation that is performed on personal data, including the collection, recording, structuring, storage, adaptation or alteration, retrieval, use, disclosure by transmission, dissemination, combination and erasure* (see Art. 4(2) in the GDPR). The entity that, alone or jointly with others, determines the purposes and means of the processing of personal data, is the so-called *data controller*, whereas the entity that processes personal data on behalf of the controller is the *data processor* (see Art. 4(7) and 4(8), respectively, in the GDPR).

If personal data are modified in such a way that the person(s) is (are) no longer identifiable, then the data should be considered anonymous, i.e., nonpersonal. In such a case, as stated in Recital 26 in the GDPR, the GDPR's provisions do not apply since we do not have personal data (although the process of anonymization, by itself, constitutes personal data processing). However, to determine whether a natural person is identifiable from a dataset that is considered anonymous, "*account should be taken of all the means reasonably likely to be used to identify the natural person directly or indirectly*". (This is described in Recital 26 of the GDPR.) In other words, performing truly effective anonymization is typically a difficult task; therefore, the GDPR actually "confirms" the research on known anonymization fallacies (see, e.g., [34]).

The GDPR codifies the basic principles that need to be in place when personal data are processed, setting specific obligations to data controllers (and there are also some obligations for data processors). More precisely, according to Article 5, par. 1 of the GDPR, *the personal data shall be processed fairly and in a transparent manner, being adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (purpose limitation principle), whereas the purpose shall be specified, explicit and legitimate*. Moreover, *the personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimization principle)*. Other basic principles are also defined in the same article, such as *the accuracy of the data (i.e., the data shall be accurate and, where necessary, kept up to date), the storage limitation (i.e., data shall not kept in a form which permits identification of data subjects for longer than is necessary for the purposes for which the personal data are processed) and the security of the data*.

An important requirement for personal data processing that should always be in place regardless of the legal basis is the so-called data protection by design principle (also known as privacy by design). According to this principle (see Art. 25, par. 1), the data controller shall, both at the time of the determination of the means of processing and at the time of the processing itself, *implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the legal requirements and protect the rights of individuals.* In simple words, personal data protection should be taken into account from the very beginning of the design of data processing. It is clear that this provision implies that when an automated decision-making system is to be used, the data controller should be very cautious from the very beginning with respect to the use of this system, considering whether it is fully necessary with respect to the desired purposes and, if yes, with which properties/characteristics, as well as establishing that the rights and freedoms of individuals are respected.

The GDPR defines several rights of the individuals that can be exercised at any time, and the respective data controller is obligated to respond within specific timeframes. A notable one is the right to access (Art. 15 in the GDPR), which is also enshrined in Art. 8 of the EU Charter of Fundamental Rights; the overall aim of the right of access is to provide individuals with sufficient, transparent, and easily accessible information about the processing of their personal data so that they can be aware of and verify the lawfulness of the processing and the accuracy of the processed data, whereas, for data controllers, the main modality for providing access is to provide the data subject with a copy of their data. Another important right is the right of data erasure (deletion) pursuant to Art. 17 in the GDPR; if specific circumstances occur, data subjects may ask a data controller to delete their data, and the controller is obligated to fulfill this requirement.

Another important right for data subjects, highly relevant to automated decision-making systems, is the one defined in Art. 22, par. 1 in the GDPR: "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*". Therefore, according to the GDPR's provisions, data controllers can carry out profiling (defined as "*any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*" according to Art. 4(4) in the GDPR) and/or automated decision-making as long as (i) they meet all the main principles for legitimate processing (i.e., transparency, data minimization, etc.) and (ii) additional safeguards are in place, with human intervention being a notable one, as well as the right, for the data subject, to challenge the decision. Moreover, the transparency of this process also means that the data subject shall always receive, from the beginning, meaningful information about *the logic involved, as well as the significance and envisaged consequences* (see Art. 13, par. 2(f) in the GDPR). Human intervention in a decision process could be omitted only in very specific cases, as described in Art. 22, par. 2 in the GDPR; however, these exceptions should be interpreted strictly, and even in these cases, other safeguards still need to be in place.

The above right is actually an obligation for all data controllers using automated decision-making systems; it concerns duties which the data controller has to fulfil without any active involvement from the data subject (e.g., there is no need for an individual to ask for human intervention with respect to the decision process; the data controller needs to ensure that this is always the case).

In order to ensure fair and transparent processing, the data controller should ensure that appropriate mathematical or statistical procedures are used so that factors that result in inaccuracies in personal data are corrected, and the risk of errors is minimized, as well *as discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation* are prevented (see Recital 71 of the GDPR). Hence, the GDPR associates the principles of fair and transparent processing (which are prerequisites for legitimate processing) with the accuracy and impartiality of the algorithm. Actually, data controllers are those that are accountable for illustrating that all these requirements are fulfilled; the so-called accountability principle is also an obligation for data controllers, pursuant to Art. 5, par. 2 of the GDPR. Taking into account that software providers or developers are, typically, not the data controllers (since, in a typical scenario, the data controllers adopt available software tools), it becomes evident that there is a complex surface, consisting of different stakeholders with different roles and responsibilities, whilst software developers and providers are not explicitly regulated by the GDPR.

An important accountability tool for data controllers, as determined in the GDPR, is the so-called data protection impact assessment (DPIA), pursuant to Art. 35 of the GDPR. In simple words, a DPIA is a systematic approach to identify, at an early stage, all the data protection risks stemming from intended personal data processing in order to consider the

appropriate mitigation measures and to evaluate whether, after these measures, there are still remaining risks. (If there are remaining high risks, the data controller needs to consult with a competent data protection authority.) Clearly, conducting a DPIA facilitates, for this processing, the controller's compliance with the data protection by design principle. A DPIA is obligatory for data controllers who aim to perform processes with high risks; to this end, as stated in Art. 35, par. 3 of the GDPR, a DPIA is required, amongst other cases, when "*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*". Therefore, conducting a DPIA is another safeguard for the rights and freedoms of individuals when automated decision-making systems are used.

2.3. Summary of the Main Challenges

As implied in Section 2.2, the GDPR seems to "be aware" of the relevant risks stemming from automated decision-making systems and puts (legal) safeguards to alleviate these risks. However, in practice, it is still challenging to implement the data protection principles stemming from the GDPR in such systems; for example, it is not clear how one can ensure and demonstrate that the data protection by design principle is in place in ML systems.

In relation to the above, it should be pointed out that the advance of ML algorithms—and especially of deep neural networks—has resulted in a new ecosystem consisting of opaque decision systems involving a huge parametric space, thus yielding complex black-box models that are far from transparent [35]. This, in turn, leads to decisions that are not justifiable and do not allow explanations of the logic that the ML algorithm follows to be derived. This is inherently noncompliant with the provisions stemming from the GDPR, which promotes transparency; moreover, even if human intervention is a prerequisite for the final decision (as the GDPR determines), the individuals who make the final decisions need to have explanations with respect to the output of the model [36]. Apart from this, an understanding of a system can allow its amendment. Therefore, due to these issues, the notion of explainable artificial intelligence (XAI) has been introduced [37] in order to allow explainable ML models to be derived while maintaining a high level of output accuracy, as well as to enable individuals to understand the systems (which is prerequisite for a system's trustworthiness). XAI can be seen in terms of its desired goals as a vehicle to develop ML systems that can be considered compliant with respect to the requirement of the fulfillment of human rights. As explicitly stated in [38], *explainability implies that an AI system should provide plain and easy-to-understand information on the factors and decision processes that serve as the basis for its prediction, recommendation, or decision*. However, there is still a series of challenges for XAI that remain insufficiently addressed [35]. Apart from explainability, other similar notions also exist (i.e., explanation, justification, accountability, legitimacy, etc.); these are discussed and analysed in [39].

The above challenges are further accentuated by the fact that, in practice, there also exist other specific data protection risks mainly related to personal data that are processed in the context of training an ML algorithm through a dataset, for example, how the data minimization principle is fulfilled for the process; which is the minimal amount of possible information that the model needs to be efficiently trained, taking into account that, if this information consists of personal data, appropriate safeguards for protecting these data need to be in place; and whether it would be acceptable in terms of personal data protection to use a training dataset that could allow identification of individuals. Such data protection risks, as well as the relevant mitigation measures, will be discussed next.

3. Data Protection Risks of ML Systems and Relations with GDPR Provisions

From the previous discussion, it has become evident that automated decision-making systems pose several risks with respect to the rights of privacy and personal data protection. More precisely, the widespread use of ML algorithms that are used as "core primitives" for these systems makes these systems even more vulnerable to new privacy threats; ML

and, especially, deep learning algorithms by themselves, as also discussed in Section 2.3, raise concerns with respect to these fundamental rights, whilst our understanding of their nature and impact on individuals is still limited [13]. These algorithms aim to extract the characteristics and properties, under a specific context, of individuals, toward reaching specific conclusions that are subsequently used to derive a decision about them. To this end, an ML algorithm is based on large volumes of input data, which, in turn, also correspond to individuals, and thus, they are also personal data.

Bearing in mind the aforementioned provisions stemming from the GDPR, a main data protection challenge for these systems stems from the data minimization principle. Indeed, to ensure compliance with the GDPR, it is necessary to develop machine learning applications that receive as input the absolutely necessary personal data; for instance, the training dataset must, at any time, consist of samples that contain the absolutely necessary personal data and no more than those that are necessary to achieve the purpose of each application [40]. More precisely, the training dataset should not allow the reidentification or singling out of an individual since such a process would violate the data minimization principle. However, performing robust anonymization is not always an easy task as stated above; the simple removal of direct identifiers (e.g., identity card numbers, names, etc.), although it is a prerequisite, does not ensure an anonymous dataset (see, e.g., some famous incidents related to improper anonymization, such as [34]).

From now on, we shall refer to the notion of an adversary as any entity that can mount a so-called privacy attack, i.e., can extract more personal information than it is entitled to do so. By considering the legal provisions of the GDPR, an adversary can be either a malicious third party (i.e., an attacker) or even an entity that legitimately processes personal data and does not have a malicious purpose (e.g., the ML service provider); if this legitimate entity is practically able, e.g., to reidentify an individual in violation of the data minimization principle, then it is still considered an adversary.

3.1. Reidentification Attacks

When the computation unit is separated from the input unit, then the input data need to be transferred to the computation unit through a secure channel. However, it is most likely that the input data are already in the computation unit, e.g., in a cloud computing service, but they are not encrypted or masked (i.e., we have data in plain sight). In any case, the following risks occur: an adversary that manages to get access to these data (i.e., a "malicious" ML service provider or a malicious external user) is able to extract information from the input data itself or even from the features extracted from the said input data. This risk becomes prevalent if the data minimization principle is not fulfilled, for example, if the training dataset allows the reidentification of individuals. However, if the input data or the ML algorithm are not correct, then the results of the ML model are incorrect, leading to a wrong decision, such as making a decision that someone is not entitled to receive a social benefit when in fact she/he is entitled to receive it. In other words, if full anonymization was applied to the input data, then the results of the ML model could possibly resemble random outputs [41].

3.2. Reconstruction Attacks

It should be pointed out that even if only the features extracted from the input data are moved to and stored on the servers of the computation unit, there is still a risk of so-called reconstruction attacks. In these attacks, the malicious user tries to reconstruct the input data using the knowledge he can extract from the model's feature vectors. Reconstruction attacks occur when the feature vectors of the model are known; this is the case when the feature vectors that have been used to train the model have not been deleted after building the machine learning model [42,43]. Some examples of successful reconstruction attacks include fingerprint reconstruction, where the image of a fingerprint was reconstructed from features of the fingerprint, and screen unlock pattern reconstruction, where the screen unlock pattern was reconstructed from features of the pattern, such as direction and speed.

In both cases, the malicious user actually exploited a data protection threat (caused by not protecting the characteristics of the extracted input data) and was able to mount a security attack on mobile device user authentication systems since, in both cases, the attacker was able to bypass the user's authentication process, imitating the actual user of the mobile devices. Although the goal of the attacks was to log into a mobile device as a specific user by fooling the mobile device user authentication systems, such attacks may also reveal other personal data of the user, such as the user age or the user location. Interestingly enough, one could say that the "source" for instantiating such privacy risks is noncompliance with the GDPR's storage limitation principle.

3.3. Model Inversion Attacks

There are also cases where the feature vectors are not stored in the ML models themselves, e.g., in the case of neural networks. In such cases, an adversary can try to construct feature vectors that mimic the actual feature vectors used to construct the ML model in question using the response received from the ML model. This type of attack is known as a model inversion attack. In other words, the adversary tries to infer personal information about a data subject by exploiting the outputs of an ML model. Such attacks could use, e.g., support vector machines (SVMs) and confidence values (e.g., probabilities or the decision value of the SVM) along with a classification result to recover feature vectors that were used to build the model [44]. These attacks aim to produce an average value that represents a particular class of data, so it constitutes a privacy risk, for example, when a particular class of data is associated with a particular individual, as in facial recognition applications [45]. It is worth noting that an inversion attack can be used together with a reconstruction attack, according to the researchers M. Fredrikson et al., in the case of face recognition [44], where the input sample is identical to the feature vector. Such attacks are mainly related to the security requirements stemming from the GDPR since the feature vectors need to be protected.

3.4. Member Inference Attacks

Another important type of privacy attack is the so-called member inference attack, which aims to infer, based on the corresponding output of the ML model, whether a sample belonged to the training dataset (a good survey on these attacks can be found in [46]). For example, an adversary can mount such an attack to find out if a particular person's personal data have been used to train a machine learning model that predicts whether a person has a particular disease. Such attacks identify and exploit the differences in the ML model's predictions regarding the samples that were used to train the model in conjunction with those samples that were not used. Such attacks are investigated in [47] through training ML models that received, as input, the correct label of a sample and the machine learning model's prediction that predicted whether the sample belonged to the training dataset or not. Interestingly enough, the researchers investigated this issue in the most difficult setting, where the adversary's access to the model was limited to "black-box" queries that returned the model's output on a given input.

All the above attacks actually illustrate how difficult it is to ensure the data minimization principle in practice when all technically possible data protection risks are considered.

4. Mitigating the Data Protection Risks

To address the aforementioned data protection risks, the research community has proposed several techniques. This section reviews the main approaches.

4.1. Protection against Reidentification Attacks

In principle, toward addressing such types of data protection threats, we need to cautiously configure the whole system to strike a proper balance between the anonymization of the training dataset and the accuracy of results. Such approaches are discussed next.

4.1.1. Noncentralized Approaches

Probably the most notable approach, from a data protection engineering perspective, for multiparty settings (i.e., in distributed learning and not in centralized learning), focusing mainly on the risks related to the information that the adversary can extract from the input data, is the so-called federated learning (FL) technique [48,49]. The underlying idea of FL techniques is to build machine learning models in a distributed way; that is, each entity, with its own data, uses these data to locally train the model without providing these data to any other party and, similarly, without receiving data from other entities for this training. (For example, in a case necessitating an ML model that is based on sensitive health data from hospitals/medical centers, each such organization trains its own local model independently based on its data.) By these means, training data are not shared at all. However, there is a coordinated central aggregate server that receives the models and, based on them, creates a global model which, in turn, is distributed to all the other entities. Therefore, the data sharing that takes place does not involve personal data but only the parameters of the model that is built. However, despite the obvious advantages of FL in terms of privacy, there are still risks remaining; for example, it has been shown that data could be inferred by considering only model weights [50], whilst even local personal data can be reconstructed through inversion of the model gradients sent by the clients to the server [51,52].

More generally, privacy-preserving techniques in distributed learning settings enable the training of an ML model using aggregated data while maintaining the privacy of the training datasets. Several research works propose ML techniques consisting of multiple members and based on homomorphic encryption (i.e., advanced encryption that allows mathematical operations on ciphertexts without necessitating decrypting them), while the type of ML used is deep learning (see, e.g., [53,54]). By these means, taking as input a large amount of encrypted data, the aim is to develop practical and efficient algorithms that handle encrypted data.

In a similar principle of decentralizing the training model, researchers in [55] proposed a technique called "Privacy-Preserving Bandits (P2B)". This technique is mainly focused on systems providing personalized services, e.g., a personalized news service that would learn to recommend news articles based on past articles the user has interacted with. Such recommender systems rely on an individual's data obtained through his/her interactions. The technique provided in [55] is based on the following idea: individual agents run locally on users' devices and are able to contribute useful feedback to other agents through centralized model updates while providing privacy guarantees, namely, differential privacy guarantees.

4.1.2. Centralized Approaches

Apart from these decentralization approaches, another way to alleviate data protection risks with respect to training data is to anonymize them; however, this comes with a well-known problem that is related to every anonymization, namely, how to answer affirmatively with certainty the question "are the resulting data indeed anonymous?" Even if we can ensure that proper anonymization has taken place, there is still another important question that needs to be considered: "Are the anonymous data still useful for the training purposes?" To this end, several techniques have been proposed. The researchers in [15] propose a differential privacy (DP) technique for the training dataset that is based on the stochastic gradient descent algorithm, i.e., noise is added into each iteration of the stochastic gradient descent algorithm with the aim of preserving the privacy of the said algorithm at a minimal cost in terms of software complexity, training efficiency, and ML model accuracy. DP focuses on achieving the following: the inclusion or exclusion of a single individual from a dataset should not significantly change the results of any analysis carried out on the dataset to make it difficult to conclude anything about that individual; to achieve this, noise is added appropriately to the data. For the ML algorithm, as the noise introduced into the stochastic gradient descent algorithm increases, higher probabilities for achieving anonymization are

achieved. But, as stated before, there is a need to configure the hyperparameters of the model appropriately in order to find the proper balance between anonymization and the validity of results [15].

In [56], a new technique is proposed, which applies a "confusion" function to the training dataset before the model training process takes place. This function either introduces random noise into the samples of the training dataset or adds new samples to it, with the aim of hiding personal information regarding the characteristics of individual samples or the statistical characteristics of a group of samples. Clearly, this technique lies within anonymization techniques; however, it should be stressed that by this method, the ML model trained with the "confused" dataset can achieve accuracy to a high extent.

An alternative approach to making private computations over the training dataset relies on the use of homomorphic encryption (see, e.g., [13,16,57]). As stated above, in the context of decentralized settings, by these means, operations take place over encrypted data so that the entity performing these operations, i.e., the training of the model, does not have access to the actual training data.

4.2. Protection against Reconstruction Attacks

With respect to the risk of reconstruction attacks, an approach that could be considered is to avoid using ML models that explicitly store the feature vectors, such as SVMs (support vector machines). If such ML models are used, it is suggested that the feature vectors are not given to the output module [42,45].

Since reconstruction attacks can also be seen as a specific instance of reidentification attacks (see Section 4.1), other well-known data protection engineering techniques, like DP mentioned above (and elaborated next), are also applicable in this case. Similarly, distributed (i.e., noncentralized) settings can also help address such threats, although it may still be necessary to employ other privacy techniques in such settings. In this respect, DP may again have an important role.

4.3. Protection against Model Inversion Attacks

With respect to the risk of model inversion attacks, any access to the output module needs to be restricted to "black-box" access, and the output of the model needs to be limited to reduce the knowledge that a malicious user can gain from it. For example, the success rate of such attacks was significantly reduced when the classification algorithms presented approximate confidence values, as shown in [44], or only the predicted class label, according to [45].

Again, model inversion attacks can be seen as a specific instance of reidentification attacks (see Section 4.1), and thus, techniques like DP are also good candidates for alleviating the relevant risks.

4.4. Protection against Member Inference Attacks

With respect to the risk of member inference attacks, it has been said that regularization techniques, such as dropout [58], which are used to overcome overfitting in machine learning, can help address them in some cases. Moreover, if the training process is differentially private, then the models are, in principle, secure against membership inference attacks that are solely on the outputs of the model [47]; however, it is known that differentially private models may significantly reduce the model's prediction accuracy. Another approach has been proposed in [59] based on models that depend on the causal relationship between input features and the output.

5. Further Exploring Differential Privacy in the Training Dataset—Results

In this work, we put effort into elaborating on the work presented in [15]; more precisely, we investigate how several parameters of the ML model affect the accuracy of the model when applying DP to the training dataset. In this respect, an extensive set of experiments was carried out, focusing on two different network architectures ("dense" and

"cnn") with two different sets of input data that are well-known in the literature, namely, the MNIST dataset for handwritten digit recognition [60] and the CIFAR-10 dataset, which consists of color images classified into 10 classes, such as ships, cats, and dogs [61]. For each combination (i.e., architecture/dataset), we applied two different modes of operation of the machine learning algorithms (i.e., without the addition of noise and, for the DP, with the addition of noise) to accurately recognize the various characters with respect to the input dataset MNIST and to accurately recognize the various objects with respect to the input dataset CIFAR-10. Note that the "dense" architecture corresponds to a dense neural network, whilst the "cnn" architecture corresponds to a convolutional neural network. The ML algorithm used is the stochastic gradient descent algorithm.

As in [15], we rely on the following definition of DP, which is based on the application-specific concept of adjacent databases. More precisely, each training dataset is a set of image–label pairs, and we say that two of these sets are adjacent if they differ in a single entry, that is, if one such pair is present in one set but not in the other [15]. Therefore, we have the following definition based on the variant introduced by [62]:

Definition 1. *A randomized mechanism M: D→R, with input domain D and output range R, satisfies (ε,δ)-differential privacy if for any two d, d′ being adjacent in D and for any subset S⊆R, it holds:*

$$\Pr[M(d) \in S] \leq e^{\varepsilon} \Pr[M(d') \in S] + \delta$$

where ε is the so-called privacy-loss parameter at a differential change in data (the smaller the value is, the better privacy protection), and δ reflects the probability of a privacy leak where the added noise (controlled by ε) does not provide sufficient protection. (If δ is low, it is less probable that an individual's privacy is going to be compromised, but this may yield a high noise to be added, which, in turn, may affect the applicability of the data).

For our experiments, we provided as input to the ML algorithm (i) the images of the input dataset MNIST and (ii) the images of the input dataset CIFAR-10, and we trained the model using either the "dense" network architecture or the "cnn" network architecture, both for the classical gradient descent algorithm as well as for the private stochastic gradient descent algorithm (having added noise). The hyperparameters of the algorithms are described as follows:

(i) Batch_Size: It determines the number of training samples used to train the network in an iteration (i.e., before updating the model parameters);
(ii) Learning_Rate: This is an important hyperparameter of a neural network that controls how much to change the model in response to the estimated error each time the model weights are updated;
(iii) L2Norm_Bound: This hyperparameter specifies the bound value for the clipping of the gradient descent algorithm (i.e., to not have too large weights while ensuring that the important components in the weight vector are larger than the other components) [15];
(iv) Sigma: This hyperparameter specifies the noise scale to be added to the stochastic gradient descent algorithm [15];
(v) Use_Privacy: This hyperparameter specifies whether the so-called private stochastic gradient descent algorithm, which introduces noise in each iteration of the stochastic gradient descent algorithm, will be used or not;
(vi) N_Epochs: This hyperparameter specifies the number of epochs for which the machine learning algorithm will be trained. An epoch refers to one cycle through the full training dataset;
(vii) Eps: This hyperparameter relates to privacy that can be controlled by the data analyst to maintain the balance between privacy and accuracy. More precisely, it specifies the initial value of the parameter "ε" (for differentially private settings) to be used when training, testing, and verifying the machine learning algorithm;

(viii) Delta: This hyperparameter specifies the initial value of the parameter "δ" (for differentially private settings) to be used during the training, testing, and verification of the machine learning algorithm;
(ix) Max_Eps: This hyperparameter specifies the maximum value of the parameter "ε" to be used when training, testing and verifying the machine learning algorithm. A larger value to this hyperparameter yields less noise into the stochastic gradient descent algorithm, whilst a smaller value to this hyperparameter yields more noise into the stochastic gradient descent algorithm;
(x) Max_Delta: This hyperparameter specifies the maximum value of the parameter "δ" (for differentially private settings) to be used during the training, testing, and verification of the machine learning algorithm;
(xi) Target_Eps: This hyperparameter specifies the value of the parameter "ε" that is actually used. If this value becomes greater than the "Max_Eps" hyperparameter, the program terminates.

Figure 1 illustrates the methodology we followed. Although this figure is based on the CIFAR-10 dataset, the same approach was also followed for the MNIST database.

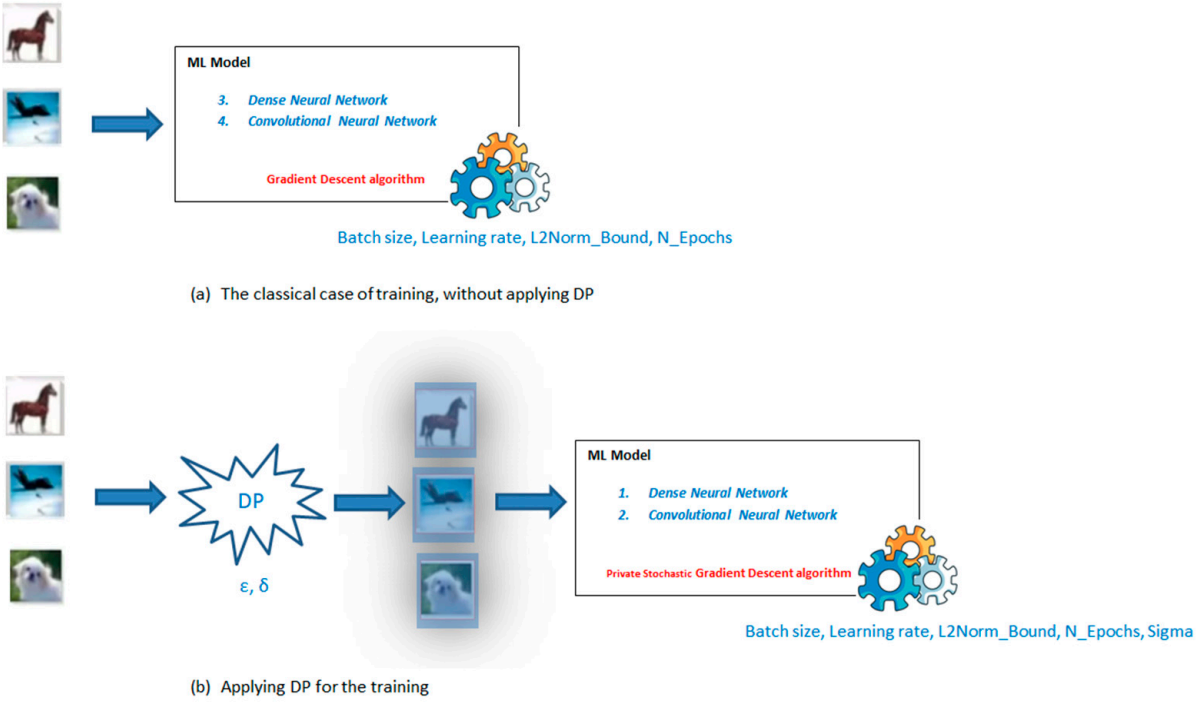

Figure 1. The methodology adopted (CIFAR-10 dataset).

All the experiments that we performed, based on the software that is publicly available in [63], are described in Table 1. Note that, for all the experiments, we set a learning rate equal to 0.01, the L2Norm_Bound equal to 4.0, and the Sigma value equal to 4.0 (i.e., low–medium noise): these are the values of the hyperparameters that were also chosen in [15]. Similarly, the value for the "δ" parameter (i.e., the Delta hyperparameter) was chosen as 1×10^{-7} (i.e., 10^{-7}) for all cases that were differentially private, similar to in [15]. Additional to the work in [15], we also examined, for the two datasets and the two models, several values for the number of epochs, the batch size, and the value of the "ε" parameter. (In [15], these parameters were chosen to be equal to 100, 64, and 1.0, respectively.) It should be pointed out that, in [15], the hyperparameters Max_Eps, Max_Delta, and Target_Eps were initialized using the values 16.0, 1×10^{-3}, and 16.0, respectively; these were also the starting values for our experiments.

Table 1. The hyperparameters chosen for the set of experiments.

#	Batch_Size	Learning_Rate	L2Norm_Bound	Sigma	Dataset	Model	Use_Privacy	N_Epochs	Eps	Delta	Max_Eps	Max_Delta	Target_Eps
1	64	0.01	4.0	4.0	MNIST	dense	False	100	____	_____	_____	_____	_____
2	64	0.01	4.0	4.0	MNIST	dense	True	100	1.0	1×10^{-7}	16.0	1×10^{-3}	16.0
3	64	0.01	4.0	4.0	MNIST	cnn	False	100	____	_____	_____	_____	_____
4	64	0.01	4.0	4.0	MNIST	cnn	True	100	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
5	64	0.01	4.0	4.0	CIFAR-10	dense	False	100	____	_____	_____	_____	_____
6	64	0.01	4.0	4.0	CIFAR-10	dense	True	100	1.0	1×10^{-7}	16.0	1×10^{-3}	16.0
7	64	0.01	4.0	4.0	CIFAR-10	cnn	False	100	____	_____	_____	_____	_____
8	64	0.01	4.0	4.0	CIFAR-10	cnn	True	100	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
9	64	0.01	4.0	4.0	MNIST	dense	True	100	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
10	64	0.01	4.0	4.0	CIFAR-10	dense	True	100	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
11	64	0.01	4.0	4.0	MNIST	dense	False	250	____	_____	_____	_____	_____
12	64	0.01	4.0	4.0	MNIST	dense	True	250	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
13	64	0.01	4.0	4.0	MNIST	cnn	False	250	____	_____	_____	_____	_____
14	64	0.01	4.0	4.0	MNIST	cnn	True	250	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
15	64	0.01	4.0	4.0	CIFAR-10	dense	False	250	____	_____	_____	_____	_____
16	64	0.01	4.0	4.0	CIFAR-10	dense	True	250	1.0	1×10^{-7}	64.0	1×10^{-2}	64.0
17	64	0.01	4.0	4.0	CIFAR-10	cnn	False	250	____	_____	_____	_____	_____
18	64	0.01	4.0	4.0	CIFAR-10	cnn	True	250	1.0	1×10^{-7}	64.0	1×10^{-3}	64.0
19	64	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
20	64	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	1.0	1×10^{-3}	1.0
21	64	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	10.0	1×10^{-2}	10.0
22	128	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	8.0	1×10^{-2}	8.0
23	32	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	16.0	1×10^{-2}	16.0
24	16	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	32.0	1×10^{-2}	32.0
25	8	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	32.0	1×10^{-2}	32.0

Table 1. *Cont.*

#	Batch_Size	Learning_Rate	L2Norm_Bound	Sigma	Dataset	Model	Use_Privacy	N_Epochs	Eps	Delta	Max_Eps	Max_Delta	Target_Eps
26	4	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	64.0	1×10^{-1}	64.0
27	2	0.01	4.0	4.0	MNIST	dense	True	250	0.5	1×10^{-7}	64.0	1×10^{-1}	64.0
28	64	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
29	64	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	1.0	1×10^{-3}	1.0
30	64	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	10.0	1×10^{-3}	10.0
31	128	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	2.0	1×10^{-3}	2.0
32	32	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	2.0	1×10^{-3}	2.0
33	16	0.01	4.0	4.0	MNIST	cnn	True	250	0.5	1×10^{-7}	4.0	1×10^{-2}	4.0
34	64	0.01	4.0	4.0	MNIST	cnn	True	500	0.5	1×10^{-7}	2.0	1×10^{-2}	2.0
35	64	0.01	4.0	4.0	MNIST	cnn	True	1000	0.5	1×10^{-7}	4.0	1×10^{-2}	4.0
36	64	0.01	4.0	4.0	MNIST	cnn	True	2000	0.5	1×10^{-7}	4.0	1×10^{-2}	4.0
37	64	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
38	64	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	1.0	1×10^{-3}	1.0
39	64	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	10.0	1×10^{-3}	10.0
40	128	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	32.0	1×10^{-3}	32.0
41	32	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	64.0	1×10^{-2}	64.0
42	16	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	64.0	1×10^{-2}	64.0
43	8	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	64.0	1×10^{-1}	64.0
44	4	0.01	4.0	4.0	CIFAR-10	dense	True	250	0.5	1×10^{-7}	128.0	1×10^{-1}	128.0
45	64	0.01	4.0	4.0	CIFAR-10	cnn	True	250	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
46	64	0.01	4.0	4.0	CIFAR-10	cnn	True	250	0.5	1×10^{-7}	1.0	1×10^{-3}	1.0
47	64	0.01	4.0	4.0	CIFAR-10	cnn	True	250	0.5	1×10^{-7}	10.0	1×10^{-3}	10.0
48	128	0.01	4.0	4.0	CIFAR-10	cnn	True	250	0.5	1×10^{-7}	2.0	1×10^{-3}	2.0
49	32	0.01	4.0	4.0	CIFAR-10	cnn	True	250	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
50	64	0.01	4.0	4.0	CIFAR-10	cnn	True	500	0.5	1×10^{-7}	4.0	1×10^{-3}	4.0
51	64	0.01	4.0	4.0	CIFAR-10	cnn	True	1000	0.5	1×10^{-7}	8.0	1×10^{-2}	8.0

The results of the experiments are shown in Table 2. For simplicity, we omitted from this table the columns corresponding to the values of the hyperparameters "Learning_Rate", "L2Norm_Bound", "Sigma", and "Delta" since they are constant for all the experiments, as shown in Table 1; however, we still keep, in Table 2, the values of the remaining hyperparameters despite the fact they are also described in Table 1 to facilitate the presentation of the results. The last column of Table 2 indicates the execution times, and the experiments took place in a Kali Linux virtual machine with two CPUs ((Intel(R) Core(TM) i7-8550U CPU @ 1.80 GHz), 8 GB RAM, 80GB HDD (SATA), NTFS).

Table 2. Results of the experiments.

	BATCH_SIZE	DATASETE	MODEL_TYPE	USE_PRIVACYE	N_EPOCHS	EPS	MAX_EPS	MAX_DELTA	TARGET_EPS	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TESTING ACCURACY (%)	EPSILON USED	DELTA USED	TRAINNING TIME
1	64	MNIST	dense	False	100	___	___	____	___	96.33	96.44	96.68	___	____	0 h 26 m
2	64	MNIST	dense	True	100	1.0	16.0	1×10^{-3}	16.0	59.64	59.96	59.10	13.993	3.6839×10^{-4}	0 h 57 m
3	64	MNIST	cnn	False	100	___	___	____	____	98.21	98.22	99.54	____	___	1 h 48 m
4	64	MNIST	cnn	True	100	1.0	64.0	1×10^{-3}	64.0	76.60	77.39	72.68	2.29	1.2746×10^{-4}	2 h 40 m
5	64	CIFAR-10	dense	False	100	____	____	____	_____	51.21	51.37	63.89	____	_____	0 h 25 m
6	64	CIFAR-10	dense	True	100	1.0	16.0	1×10^{-3}	16.0	15.64	15.71	13.73	16.0	1.4989×10^{-4}	0 h 15 m
7	64	CIFAR-10	cnn	False	100	___	_____	_____	_____	62.46	62.52	88.25	____	_____	2 h 0 m
8	64	CIFAR-10	cnn	True	100	1.0	64.0	1×10^{-3}	64.0	25.12	25.38	24.72	3.7747	1.5798×10^{-4}	2 h 12 m
9	64	MNIST	dense	True	100	1.0	64.0	1×10^{-3}	64.0	61.34	61.83	58.08	13.99	3.6839×10^{-4}	0 h 55 m
10	64	CIFAR-10	dense	True	100	1.0	64.0	1×10^{-3}	64.0	15.32	15.46	14.97	36.5621	7.9777×10^{-4}	1 h 57 m
11	64	MNIST	dense	False	250	_____	_____	_____	_____	97.27	97.29	98.49	_____	_____	0 h 34 m
12	64	MNIST	dense	True	250	1.0	64.0	1×10^{-3}	64.0	66.39	66.68	64.74	22.1849	9.1887×10^{-4}	1 h 50 m
13	64	MNIST	cnn	False	250	_____	_____	_____	_____	98.28	98.28	99.83	_____	_____	5 h 13 m
14	64	MNIST	cnn	True	250	1.0	64.0	1×10^{-3}	64.0	81.44	81.68	78.95	3.6135	3.1085×10^{-4}	6 h 08 m
15	64	CIFAR-10	dense	False	250	_____	_____	_____	_____	52.59	52.61	80.59	_____	_____	6 h 0 m
16	64	CIFAR-10	dense	True	250	1.0	64.0	1×10^{-2}	64.0	16.90	16.99	17.31	5.79335	1.99118×10^{-3}	4 h 39 m
17	64	CIFAR-10	cnn	False	250	___	___	___	___	62.70	62.70	94.68	____	___	6 h 20 m
18	64	CIFAR-10	cnn	True	250	1.0	64.0	1×10^{-2}	64.0	30.06	30.26	19.49	5.9733	8.3994×10^{-4}	6 h 24 m
19	64	MNIST	dense	True	250	0.5	4.0	1×10^{-3}	4.0	54.52	54.60	57.98	4.0	2.0691×10^{-4}	0 h 25 m
20	64	MNIST	dense	True	250	0.5	1.0	1×10^{-3}	1.0	23.62	22.66	37.71	1.0	1.2940×10^{-4}	0 h 2 m
21	64	MNIST	dense	True	250	0.5	10.0	1×10^{-2}	10.0	67.72	67.49	66.10	8.4436	1.9570×10^{-4}	1 h 59 m
22	128	MNIST	dense	True	250	0.5	8.0	1×10^{-2}	8.0	66.83	66.97	65.23	5.9668	4.6011×10^{-4}	0 h 59 m
23	32	MNIST	dense	True	250	0.5	16.0	1×10^{-2}	16.0	69.59	72.11	72.33	11.9089	1.83497×10^{-3}	3 h 50 m
24	16	MNIST	dense	True	250	0.5	32.0	1×10^{-2}	32.0	75.43	75.65	73.84	16.7136	3.6697×10^{-2}	7 h 07 m
25	8	MNIST	dense	True	250	0.5	32.0	1×10^{-2}	32.0	78.80	78.94	77.17	23.7963	7.3353×10^{-2}	8 h 09 m
26	4	MNIST	dense	True	250	0.5	64.0	1×10^{-1}	64.0	81.49	81.26	80.52	32.355	1.467053×10^{-2}	28 h 3 m
27	2	MNIST	dense	True	250	0.5	64.0	1×10^{-1}	64.0	83.81	83.90	83.23	44.8134	2.933942×10^{-2}	55 h 25 m
28	64	MNIST	cnn	True	250	0.5	4.0	1×10^{-3}	4.0	80.25	80.06	80.24	1.3608	3.1085×10^{-4}	6 h 21 m
29	64	MNIST	cnn	True	250	0.5	1.0	1×10^{-3}	1.0	75.68	76.21	79.84	0.9996	1.6747×10^{-4}	3 h 04 m
30	64	MNIST	cnn	True	250	0.5	10.0	1×10^{-3}	10.0	83.46	83.65	80.78	1.3608	3.1085×10^{-4}	6 h 16 m
31	128	MNIST	cnn	True	250	0.5	2.0	1×10^{-3}	2.0	79.64	79.93	77.05	0.9655	1.5764×10^{-4}	3 h 23 m
32	32	MNIST	cnn	True	250	0.5	2.0	1×10^{-3}	2.0	81.24	81.42	78.31	1.9207	6.3528×10^{-4}	11 h 09 m
33	16	MNIST	cnn	True	250	0.5	4.0	1×10^{-2}	4.0	75.38	75.71	73.57	2.7538	1.24625×10^{-3}	23 h 03 m
34	64	MNIST	cnn	True	500	0.5	2.0	1×10^{-2}	2.0	85.84	85.90	83.90	1.9208	6.3540×10^{-4}	10 h 54 m
35	64	MNIST	cnn	True	1000	0.5	4.0	1×10^{-2}	4.0	85.84	85.90	83.80	2.7532	1.24579×10^{-3}	22 h 47 m

Table 2. *Cont.*

	BATCH_SIZE	DATASETE	MODEL_TYPE	USE_PRIVACYE	N_EPOCHS	EPS	MAX_EPS	MAX_DELTA	TARGET_EPS	TRAIN ACCURACY (%)	VALIDATION ACCURACY (%)	TESTING ACCURACY (%)	EPSILON USED	DELTA USED	TRAINNING TIME
36	64	MNIST	cnn	True	2000	0.5	4.0	1×10^{-2}	4.0	89.62	89.63	87.71	3.8095	2.52171×10^{-3}	42 h 01 m
37	64	CIFAR-10	dense	True	250	0.5	4.0	1×10^{-3}	4.0	12.53	11.82	10.45	4.0	6.215×10^{-5}	0 h 9 m
38	64	CIFAR-10	dense	True	250	0.5	1.0	1×10^{-3}	1.0	9.77	9.41	7.81	1.0	3.88×10^{-5}	0 h 1 m
39	64	CIFAR-10	dense	True	250	0.5	10.0	1×10^{-3}	10.0	13.67	13.66	13.99	10.0	3.8742×10^{-3}	50 h 01 m
40	128	CIFAR-10	dense	True	250	0.5	32.0	1×10^{-3}	32.0	16.12	16.14	16.20	16.0331	9.9818×10^{-4}	2 h 15 m
41	32	CIFAR-10	dense	True	250	0.5	32.0	1×10^{-3}	32.0	18.50	18.61	18.82	32.1464	4.02584×10^{-3}	9 h 10 m
42	16	CIFAR-10	dense	True	250	0.5	64.0	1×10^{-2}	64.0	20.33	20.39	20.62	45.4614	7.96743×10^{-3}	15 h 50 m
43	8	CIFAR-10	dense	True	250	0.5	64.0	1×10^{-1}	64.0	21.12	21.15	21.06	63.7962	1.560910×10^{-2}	44 h 25 m
44	4	CIFAR-10	dense	True	250	0.5	128.0	1×10^{-1}	128.0	22.31	22.36	22.24	91.1240	3.364267×10^{-2}	68 h 13 m
45	64	CIFAR-10	cnn	True	250	0.5	4.0	1×10^{-3}	4.0	29.59	29.69	30.16	2.2675	3.8994×10^{-3}	6 h 09 m
46	64	CIFAR-10	cnn	True	250	0.5	1.0	1×10^{-3}	1.0	24.51	22.87	23.03	1.0	7.617×10^{-5}	1 h 21 m
47	64	CIFAR-10	cnn	True	250	0.5	10.0	1×10^{-3}	10.0	29.14	29.25	29.04	2.2675	3.8994×10^{-3}	1 h 0 m
48	128	CIFAR-10	cnn	True	250	0.5	2.0	1×10^{-3}	2.0	29.64	29.87	29.39	1.6026	1.9852×10^{-4}	2 h 59 m
49	32	CIFAR-10	cnn	True	250	0.5	4.0	1×10^{-3}	4.0	19.54	19.54	19.89	3.1539	7.7702×10^{-4}	11 h 26 m
50	64	CIFAR-10	cnn	True	500	0.5	4.0	1×10^{-3}	4.0	31.23	31.26	31.29	3.1524	7.7616×10^{-4}	12 h 03 m
51	64	CIFAR-10	cnn	True	1000	0.5	8.0	1×10^{-2}	8.0	32.30	32.30	32.06	4.4014	1.66101×10^{-3}	23 h 31 m

The results obtained with respect to the dense neural networks when DP is applied to the training dataset are promising in the case of identifying characters (i.e., for the MNIST input set). More precisely, for dense neural networks, we managed to achieve training accuracy, validation accuracy, and testing accuracy of 83.81%, 83.90%, and και 83.23%, respectively, for (44.8, 2.933942×10^{-2})-differential privacy (see the 27th experiment). This was achieved for an increased number of epochs, namely, 250 (instead of 100 that were used in [15]). Similarly, even when we used convolutional neural networks, we managed to achieve training accuracy, validation accuracy, and testing accuracy of 89.62%, 89.63%, and 87.71%, respectively, for (3.8095, 2.52171×10^{-3})-differential privacy (see the 36th experiment). Again, this was achieved for an increased number of epochs, namely, 2000. As can also be seen, a smaller value for the "ε" parameter was used in the latter case compared to the former case, thus achieving better privacy in the convolutional neural networks than in the deep neural networks.

The results significantly changed when we focused on the case of identifying objects within images (i.e., for the CIFAR-10 input set). More precisely, for dense neural networks, the best case is the one achieving training accuracy, validation accuracy, and testing accuracy of 22.31%, 22.36%, and 22.24%, respectively, for (91.1240, 3.364267×10^{-2})-differential privacy (see the 44th experiment), and, as in the case of the MNIST dataset, this was achieved for an increased number of epochs. The results slightly improved for the convolutional neural networks, where we managed to achieve training accuracy, validation accuracy, and testing accuracy of 32.30%, 32.30%, and 32.06%, respectively, for (4.4014, 1.66101×10^{-3})-differential privacy (see the 51st experiment). Hence, again, we see that through convolutional neural networks, we achieve better privacy than in deep neural networks, but the input dataset seems to have a very crucial role with respect to the accuracy of the model.

6. Discussion

The aforementioned extended set of experiments (see Tables 1 and 2) indicates how difficult it is to find a proper balance between privacy and an algorithm's accuracy due

to the variety of hyperparameters that highly affect the overall results. Although there are, indeed, promising techniques (and differential privacy is one of them), it seems that currently, there is no one-size-fits-all solution, whilst the type of the input data seems to also have a dominant role. In any case, it has become evident that it deserves attention to further explore how the hyperparameters affect both privacy and accuracy, even on an ad hoc basis.

However, it should be stressed that the context of the use of an ML algorithm is also of high importance. For example, for automated decision-making systems, it is questionable even if accuracy at 83%—as achieved by our experiments—can be considered acceptable despite the fact that, in cases that do not involve personal data processing, it may, indeed, be sufficient. Therefore, a systematic risk analysis is needed for each case, and special emphasis should always be put on the transparency of the processing. This is a very challenging task for all relevant stakeholders (AI product/software developers, organizations using such tools, competent authorities, legislators, etc.).

The above challenges also include issues stemming from the obligations that the data controllers have with respect to responding to specific individuals' rights. For example, what if a data subject asks to either access their data that are included in a training dataset or to delete these data? Interestingly enough, according to the GDPR (see Art. 11), if the data controller is able to demonstrate that they are not in a position to identify the data subject, then the GDPR's Articles 15 to 20, which correspond to data subject rights, such as the right of access and the right of data erasure, shall not apply except where the data subject, for the purpose of exercising the rights under those articles, provides additional information enabling the data subject's identification. Therefore, if the training datasets are properly anonymized, and the data controller is able to prove this, there is no obligation for the controller to "find ways" to respond to data subject requests, and actually, the data controller is not obliged to maintain, acquire, or process additional information in order to identify the data subject for the sole purpose of fulfilling data subject rights. Of course, the provisions regarding the right that an individual will not be subject to a decision based solely on automated processing still remain.

The above illustrates how important it is to establish proper systematic procedures to evaluate a decision-making system based on ML algorithms in terms of its impact on fundamental human rights. In our view, the GDPR's provisions pave the way for some important prerequisites that need to be fulfilled, including the prohibition of fully automated decisions. Moreover, conducting a DPIA is necessary to identify, in time, all the main risks with respect to personal data protection and privacy; however, such a DPIA—which is a generic accountability tool for any personal data process with high risk—should also meticulously consider all the main properties and characteristics of an ML algorithm, striving to provide convincing answers to questions such as:

(a) Is the usage of an ML algorithm fully necessary? If yes, which type of ML algorithm fits better with our needs and why?;
(b) Are the identities of the individuals whose data form the training dataset protected?;
(c) Is it possible to extract personal information concerning the training dataset from the features/parameters of the ML model?;
(d) Is the model resistant to reconstruction attacks, model inversion attacks, and member inference attacks?;
(e) Under the assumption that the above is ensured (which needs to be demonstrated), is the algorithm accurate? How this can be proved?

Most importantly, many of the above questions cannot be directly addressed by data controllers (i.e., the organizations using such tools, which have been developed by other stakeholders). However, it is essential that before the use of such systems, systematic procedures should be adopted to address all these risks within the context of a well-established impact assessment. In Europe, the proposal for the EU Artificial Intelligence Act (which had not finalized yet during the preparation and writing of this paper) introduces the concept of the fundamental rights impact assessment (FRIA) (see, e.g., [64]). Due to

this provision, the deployers of high-risk systems (such systems are also defined within the AI Act) should conduct an FRIA and develop a risk mitigation plan in coordination with the competent data protection supervisory authority and relevant stakeholders before market entry (see Recital 58a of this proposal). This is a significant safeguard for personal data protection and privacy but, in turn, sets several challenges:

(a) What should be the interplay between an FRIA and DPIA? For example, the questions stated above concerning how privacy risks are mitigated by an ML algorithm seem to be better addressed first in the context of an FRIA;
(b) How easy is it to conduct a robust FRIA/DPIA that demonstrates that all the State-of-the-Art research outcomes on privacy-enhancing technologies for ML systems have been meticulously taken into account? Note that, recalling our previous analysis based on experimental results, this is a very challenging task.

7. Conclusions

The analysis in this paper indicates how challenging it is to ensure that personal data protection principles, as provisioned in the GDPR, are in place when ML techniques are used. Indeed, we focused on DP as a well-known privacy-enhancing technique to alleviate relevant threats, and we confirmed, through an extensive set of experiments, that there is no guarantee that all the desired goals (i.e., the accuracy of the algorithm's output while protecting personal data in the training dataset) can be achieved. Therefore, there exist challenges related to the facts that (i) no privacy-enhancing technique should be considered a panacea with respect to mitigating the relative risks stemming from the process of model training, and (ii) the application of a privacy-enhancing technique may also affect the overall accuracy of the model. This, in turn, in decision-making systems, also has an impact on fundamental rights, including the right to personal data protection. Hence, the experimental results in Section 6 further reveal the inherent difficulties in striking a proper balance between privacy and accuracy. Therefore, a very careful assessment of the actual data protection risks is needed for each specific case.

As a generic conclusion, it can be considered that the European legal provisions, which (try to) embed safeguards, such as impact assessments before the start of any data processing, could be considered a "nice model" globally for establishing appropriate legal frameworks, whilst, especially the requirement for human intervention in automated decision-making systems should be considered a prerequisite for respecting fundamental human rights. Indeed, with respect, e.g., to personal data protection, adopting the data protection by design principle—although it has its own challenges in terms of its proper implementation—seems to be the proper way to alleviate data protection issues from the very beginning of the process since posteriori privacy-enhancing solutions may not be sufficient. On the other side, it is also essential that the different communities, i.e., researchers, AI systems developers, AI systems providers, AI systems users (i.e., data controllers in terms of the GDPR's provisions), and data protection authorities, should be in a close relationship to eliminating any possible lack of understanding between them. Especially the outcomes and methodologies from the research community need to be clearly identified in the context of an impact assessment. More specifically, stakeholders that are to conduct (either jointly or alone) an impact assessment should be fully aware of the relevant State-of-the-Art research outcomes on data protection engineering in order to justify any decision that is to be made with respect to the evaluation of the remaining risks. All these constitute, in turn, new challenges that need to be addressed.

In a recent report [38], governance aspects in terms of the risk management process for AI systems are elaborated and described as the key to achieving trustworthy AI. Governance is described, therein, as an activity with two elements: the first concerns the governance of the risk management process itself and includes monitoring, reviewing, documenting, communicating, and consulting on the process and its outcomes, whilst the second ensures the effectiveness of the risk management process by embedding it in the culture and broader governance processes of organizations. Based on this, as well as on the discussion

and findings in this paper, it is evident that the aforementioned governance procedures should incorporate considerations for employing efficient data protection engineering in AI systems, which—as stated above—entails the active involvement and collaboration of several stakeholders, including the research community and academia. Clearly, data protection engineering should not be considered an obstacle to ensuring XAI, which is also important for accountability.

As a concluding statement, trust is a prerequisite for the secure adoption of ML technologies, and personal data protection engineering provides the means to achieve trust. However, well-known privacy-enhancing technologies may not necessarily alleviate all the risks in such complex environments, and thus, further research should be definitely deemed necessary. However, despite the need for further research, the requirement for always taking appropriately into account existing privacy-enhancing technologies within the framework of a systematic risk assessment is still (and will, of course, remain) essential.

Author Contributions: Conceptualization, K.L.; methodology, P.C. and K.L.; software, P.C.; writing—original draft preparation, K.L. and P.C.; writing—review and editing, K.L. and P.C.; supervision, K.L.; project administration, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The experimental results were obtained based on publicly available software for research purposes that is cited in this paper.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very constructive comments and suggestions, which helped to greatly improve the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
DP	Differential privacy
DPIA	Data protection impact assessment
FL	Federated learning
FRIA	Fundamental rights impact assessment
GDPR	General Data Protection Regulation
ML	Machine learning
PII	Personally identifiable information
SVM	Support vector machine
UN	United Nations
XAI	Explainable artificial intelligence

References

1. Bergs, T.; Holst, C.; Gupta, P.; Augspurger, T. Digital image processing with deep learning for automated cutting tool wear detection. *Procedia Manuf.* **2022**, *48*, 947–958. [CrossRef]
2. Qin, S.J.; Chiang, L.H. Advances and opportunities in machine learning for process data analytics. *Comput. Chem. Eng.* **2019**, *126*, 465–473. [CrossRef]
3. Mallozi, P. Combining Machine-Learning with Invariants Assurance Techniques for Autonomous Systems. In Proceedings of the IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C), Buenos Aires, Argentina, 20–28 May 2017; pp. 485–486. [CrossRef]
4. Ivanko, D.; Ryumin, D.; Karpov, A. A Review of Recent Advances on Deep Learning Methods for Audio-Visual Speech Recognition. *Mathematics* **2023**, *11*, 2665. [CrossRef]
5. Sharma, S.; Bhatt, M.; Sharma, P. Face Recognition System Using Machine Learning Algorithm. In Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1162–1168. [CrossRef]
6. Mosavi, A.; Varkonyi-Koczy, A. Integration of Machine Learning and Optimization for Robot Learning. In Proceedings of the 15th International Conference on Recent Global Research and Education: Technological Challenges, Warsaw, Poland, 26–28 September 2016; pp. 349–356. [CrossRef]

7. Boukerche, A.; Wang, J. Machine Learning-based traffic prediction models for Intelligent Transportation Systems. *Comput. Netw.* **2020**, *181*, 107530. [CrossRef]
8. Liu, H.; Lang, B. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Appl. Sci.* **2019**, *9*, 4396. [CrossRef]
9. Shinde, P.P.; Shah, S. A Review of Machine Learning and Deep Learning applications. In Proceedings of the 4th International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 14–16 August 2018; pp. 1–6. [CrossRef]
10. Braker, C.; Shiaeles, S.; Bendiab, G.; Savage, N.; Limniotis, K. BotSpot: Deep Learning Classification of Bot Accounts Within Twitter. In Proceedings of the Internet of Things, Smart Spaces, and Next Generation Networks and Systems (NEW2AN ruSMART), St. Petersburg, Russia, 26–28 August 2020; pp. 165–175. [CrossRef]
11. Data Protection WP Art. 29. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. 2018. Available online: https://ec.europa.eu/newsroom/article29/items/612053 (accessed on 24 November 2023).
12. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *119*, 1–88.
13. Mohassel, P.; Zhang, Y. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 2017; pp. 19–38. [CrossRef]
14. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]
15. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In Proceedings of the ACM SIGSAC Conf. on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318. [CrossRef]
16. Hesamifard, E.; Takabi, H.; Ghasemi, M.; Wright, R.N. Privacy-preserving Machine Learning as a Service. In Proceedings of the Privacy Enhancing Technologies Symposium, Barcelona, Spain, 24–27 July 2018; pp. 123–142. [CrossRef]
17. Chhetri, T.; Kurteva, A.; DeLong, R.; Hilscher, R.; Korte, K.; Fensel, A. Data Protection by Design Tool for Automated GDPR Compliance Verification Based on Semantically Modeled Informed Consent. *Sensors* **2022**, *22*, 2763. [CrossRef] [PubMed]
18. Michael, J.B.; Kuhn, R.; Voas, J. Security or Privacy: Can You Have Both? *Computer* **2020**, *53*, 20–30. [CrossRef]
19. BEUC. Automated Decision Making and Artificial Intelligence—A Consumer Perspective. 2018. Available online: https://www.beuc.eu/sites/default/files/publications/beuc-x-2018-058_automated_decision_making_and_artificial_intelligence.pdf (accessed on 24 November 2023).
20. Gilman, M. Poverty Algorithms. Data and Society 2020. Available online: https://datasociety.net/wp-content/uploads/2020/09/Poverty-Lawgorithms-20200915.pdf (accessed on 24 November 2023).
21. Goodman, R. Why Amazon's Automated Hiring Tool Discriminated Against Women. ACLU 2018. Available online: https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against (accessed on 24 November 2023).
22. Hirvonen, H. Just accountability structures—A way to promote the safe use of automated decision-making in the public sector. *AI Soc.* **2024**, *39*, 155–167. [CrossRef]
23. Center for Democracy and Technology. AI & Machine Learning. 2019. Available online: https://cdt.org/ai-machine-learning/ (accessed on 22 October 2023).
24. Madden, M.; Gilman, M.; Levy, K.; Marwick, A. Privacy, Poverty, and Big Data: A Matrix of vulnerabilities for poor Americans. *Wash. Univ. Law Rev.* **2017**, *95*, 53. Available online: https://openscholarship.wustl.edu/law_lawreview/vol95/iss1/6 (accessed on 24 November 2023).
25. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef] [PubMed]
26. Sharma, C.; Damani, D.; Chariar, V. Review and content analysis of textual expressions as a marker for depressive and anxiety disorders (DAD) detection using machine learning. *Discov. Artif. Intell.* **2023**, *3*, 38. [CrossRef]
27. Parmar, S.; Paunwala, C. Early detection of dyslexia based on EEG with novel predictor extraction and selection. *Discov. Artif. Intell.* **2023**, *3*, 33. [CrossRef]
28. Ghaffar Nia, N.; Kaplanoglu, E.; Nasab, A. Evaluation of Artificial Intelligence techniques in disease diagnosis and prediction. *Discov. Artif. Intell.* **2023**, *3*, 5. [CrossRef]
29. United Nations. Universal Declaration of Human Rights. 1948. Available online: https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed on 24 November 2023).
30. United Nations. International Covenant on Civil and Political Rights. 1966. Available online: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights (accessed on 24 November 2023).
31. European Union. Charter Of Fundamental Rights of the European Union. *Off. J. Eur. Communities* **2000**, 1–64. Available online: https://www.europarl.europa.eu/charter/pdf/text_en.pdf (accessed on 24 November 2023).
32. Kaminski, M. A Recent Renaissance in Privacy Law. *Commun. ACM* **2020**, 24–27. Available online: https://scholar.law.colorado.edu/articles/1292/ (accessed on 24 November 2023).
33. Gellert, R. Personal data's ever-expanding scope in smart environments and possible path(s) for regulating emerging digital technologies. *Int. Priv. Law* **2021**, *11*, 196–208. [CrossRef]

34. Narayanan, A.; Shmatikov, V. Robust de-anonymization of large sparse datasets. In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, USA, 18–21 May 2018; pp. 115–125. [CrossRef]
35. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
36. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]
37. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
38. OECD. Advancing Accountability in AI—Governing and Managing Risks throughout the Lifecycle for Trustworthy AI. 2023. Available online: https://www.oecd-ilibrary.org/science-and-technology/advancing-accountability-in-ai_2448f04b-en (accessed on 24 November 2023).
39. Henin, C.; Le Métayer, D. Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI Soc.* **2022**, *2*, 1397–1410. [CrossRef]
40. Goldsteen, A.; Ezov, G.; Shmelkin, R.; Moffie, M.; Farkash, A. Data minimization for GDPR compliance in machine learning models. *AI Ethics* **2022**, *2*, 477–491. [CrossRef]
41. Slijepčević, D.; Henzl, M.; Klausner, L.D.; Dam, T.; Kieseberg, P.; Zeppelzauer, M. k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Comput. Secur.* **2021**, *111*, 102488. [CrossRef]
42. Feng, J.; Jain, A.K. Fingerprint Reconstruction: From Minutiae to Phase. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 209–223. [CrossRef]
43. Papernot, N.; McDaniel, P.; Sinha, A.; Wellman, M.P. SoK: Security and Privacy in Machine Learning. In Proceedings of the IEEE European Symposium on Security and Privacy, London, UK, 23–27 April 2018. [CrossRef]
44. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1322–1333. [CrossRef]
45. Al-Rubaie, M.; Chang, J.M. Reconstruction Attacks Against Mobile-Based Continuous Authentication Systems in the Cloud. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 2648–2663. [CrossRef]
46. Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P.S.; Zhang, X. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–37. [CrossRef]
47. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 2017; pp. 3–18. [CrossRef]
48. Hu, K.; Yue, H.; Guo, L.; Yuanxiong, G.; Yuguang, F. Privacy-Preserving Machine Learning Algorithms for Big Data Systems. In Proceedings of the IEEE 35th International Conference on Distributed Computing Systems, Columbus, OH, USA, 29 June–2 July 2015; pp. 318–327. [CrossRef]
49. McMahan, B.; Ramag, D. Federated Learning: Collaborative Machine Learning without Centralized Training Data. 2017. Available online: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (accessed on 22 October 2023).
50. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–22 May 2019; pp. 739–753. [CrossRef]
51. Melis, L.; Song, C.; De Cristofaro, E.; Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In Proceedings of the IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–22 May 2019; pp. 691–706. [CrossRef]
52. Geiping, J.; Bauermeister, H.; Dröge, H.; Moeller, M. Inverting Gradients-How Easy Is It to Break Privacy in Federated Learning? In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver BC, Canada, 6–12 December; 2020; pp. 16937–16947.
53. Takabi, H.; Hesamifard, E.; Ghasemi, M. Privacy preserving multi-party machine learning with Homomorphic encryption. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; Available online: https://pmpml.github.io/PMPML16/papers/PMPML16_paper_14.pdf (accessed on 24 November 2023).
54. Xu, G.; Li, G.; Guo, S.; Zhang, T.; Li, H. Secure Decentralized Image Classification with Multiparty Homomorphic Encryption. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3185–3198. [CrossRef]
55. Malekzadeh, M.; Athanasakis, D.; Haddadi, H.; Livshits, B. Privacy-Preserving Bandits. *arXiv* **2020**, arXiv:1909.04421.
56. Zhang, T.; He, Z.; Lee, R.B. Privacy-preserving Machine Learning through Data Obfuscation. *arXiv* **2018**, arXiv:1807.01860.
57. Crockett, E. A Low-Depth Homomorphic Circuit for Logistic Regression Model Training. Cryptology ePrint Archive, Paper 2020/1483, 2020. Available online: https://eprint.iacr.org/2020/1483 (accessed on 22 October 2023).
58. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. Available online: https://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf (accessed on 24 November 2023).
59. Tople, S.; Sharma, A.; Nori, A. Alleviating Privacy Attacks via Causal Learning. In Proceedings of the 37th International Conference on Machine Learning (ICML), Vienna, Austria, 13–18 July 2020; pp. 9537–9547.
60. LeCun, Y.; Cortes, C.; Burges, C.J. THE MNIST Database. 2023. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 22 October 2023).

61. Krizhevsky, A.; Nair, V.; Hinton, G. The CIFAR-10 Dataset. 2023. Available online: https://www.cs.toronto.edu/~kriz/cifar.html (accessed on 22 October 2023).
62. Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; Naor, M. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In Proceedings of the Advances in Cryptology—EUROCRYPT, Saint Petersburg, Russia, 28 May–1 June 2006; pp. 486–503. [CrossRef]
63. Lillelund, C.; Hopkins, T. dpsgd-Optimizer. Available online: https://github.com/thecml/dpsgd-optimizer (accessed on 22 October 2023).
64. Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; Floridi, L. Taking AI risks seriously: A new assessment model for the AI Act. *AI Soc.* **2023**. [CrossRef]