



Article

Toward Learning Trustworthily from Data Combining Privacy, Fairness, and Explainability: An Application to Face Recognition

Danilo Franco ¹, Luca Oneto ^{1,*}, Nicolò Navarin ² and Davide Anguita ¹

¹ Department of Computer Science, Bioengineering, Robotics and Systems Engineering, University of Genoa, Via Opera Pia 11a, 16145 Genova, Italy; danilo.franco@edu.unige.it (D.F.); davide.anguita@unige.it (D.A.)

² Dipartimento di Matematica “Tullio Levi-Civita”, University of Padua, Via Trieste 63, 35121 Padova, Italy; nnavarin@math.unipd.it

* Correspondence: luca.oneto@unige.it

Abstract: In many decision-making scenarios, ranging from recreational activities to healthcare and policing, the use of artificial intelligence coupled with the ability to learn from historical data is becoming ubiquitous. This widespread adoption of automated systems is accompanied by the increasing concerns regarding their ethical implications. Fundamental rights, such as the ones that require the preservation of privacy, do not discriminate based on sensible attributes (e.g., gender, ethnicity, political/sexual orientation), or require one to provide an explanation for a decision, are daily undermined by the use of increasingly complex and less understandable yet more accurate learning algorithms. For this purpose, in this work, we work toward the development of systems able to ensure trustworthiness by delivering privacy, fairness, and explainability by design. In particular, we show that it is possible to simultaneously learn from data while preserving the privacy of the individuals thanks to the use of Homomorphic Encryption, ensuring fairness by learning a fair representation from the data, and ensuring explainable decisions with local and global explanations without compromising the accuracy of the final models. We test our approach on a widespread but still controversial application, namely face recognition, using the recent FairFace dataset to prove the validity of our approach.

Keywords: trustworthy artificial intelligence; deep neural networks; Algorithmic Fairness; learning fair representation; privacy-preserving machine learning; Homomorphic Encryption; explainable artificial intelligence; attention maps; dimensionality reduction



Citation: Franco, D.; Oneto, L.; Navarin, N.; Anguita, D. Toward Learning Trustworthily from Data Combining Privacy, Fairness, and Explainability: An Application to Face Recognition. *Entropy* **2021**, *23*, 1047. <https://doi.org/10.3390/e23081047>

Academic Editor: Friedhelm Schwenker

Received: 30 June 2021

Accepted: 11 August 2021

Published: 14 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Trustworthiness in artificial intelligence (AI) stands out as one of the main problems to be addressed in developing the future of the modern technological societies [1]. One of the first stances that depicted the necessity of deploying trustworthy information and communications technologies goes back to 1999 with “Trust in Cyberspace” [2], where researchers showed that governments started to become dependent on possibly unreliable algorithms for operating their critical infrastructures, such as communication, transportation, and power distribution [3]. Analogously to what has been done in the past for these infrastructures, there is nowadays a need to address trustworthiness in AI systems as a holistic property able to guarantee fundamental rights, encapsulate the ethical principles of the society, enforce resilience to disruption, and cope with human errors or hostile attacks [4]. The resulting benefits are numerous and multifaceted. For example, it can contribute to increasing well-being both on a collective and an individual level, for example by generating wealth [5] or taking care of tedious or dangerous tasks [6]. Moreover, it can promote fairer behaviors toward social and political equality [4].

In general, this ambitious objective cannot be reached in a single step, but there is a need to first face specific sub-problems and then combine the results toward a more holistic

approach [7]. In the context of AI, a fundamental building block is the ability to learn from data by means of machine learning (ML)-based technologies [8]. This ability allows us to make predictions based on historical data supporting decision makers (human or autonomous) [9]. Models learned from data have been shown to deliver very accurate results in recent years, outperforming human abilities in some specific applications [10–12] with the use of increasingly complex ML algorithms on the increasing number of available data [13]. Simultaneously, researchers have begun to show the drawbacks of rushing towards these accuracy levels: models have started to also learn the human biases and misbehavior [14–16], to break the privacy of the single individuals [17,18], to show limited robustness to (malicious) data perturbations [19–23], and to be less and less understandable, undermining the fundamental right of explanation principle [24]. For these reasons, researchers have started to study these problems separately, developing the fields of Algorithmic Fairness [25–27], Privacy-Preserving Data Analysis [28], Adversarial Machine Learning [29], and Explainable Machine Learning [30–32], respectively. Unfortunately few works in the literature have tried to address more than one of these problems simultaneously. Some have tried to face two of them: for example some works combine fairness with privacy [33–41], others [42–44] combine adversarial learning with fairness, fairness with explainability [45–47], adversarial learning with explainability [48,49], and adversarial learning with privacy [50–53]. For this reason, in this work, we drive toward the development of systems able to ensure trustworthiness by delivering privacy, fairness, and explainability by design.

Privacy requires protecting the data of the single individuals along with all the information generated during the entire data lifecycle [4]. It is easy to explain the practical necessity of privacy guarantees, especially in those applications where digital records directly contain (or can be exploited to infer) highly sensitive information, such as gender, ethnicity, or sexual or political orientations [28,54]. For this reason, it is required to cope with the problem of developing ML models able to simultaneously extract useful and actionable information from data and not violate the privacy of single individuals. Algorithmic Fairness requires the outputs of ML-based models to not depend on sensitive attributes (e.g., gender, race, and political/sexual orientation) [26,55]. In fact, datasets may contain historical biases (e.g., discrimination against historically mistreated subgroup in the population) or may suffer from a coarse view of the modern societies (e.g., underrepresented subgroups). ML models trained on those biased data may exacerbate unfairness, generating a cascade effect [4,7,56]. For this reason, it is required to cope with the problem of developing methods to mitigate such biases. Explainability in ML is the ability to provide an explanation for the output of an ML-based model [30,32,57]. Explanations can be local (i.e., why the model gave a particular output provided a particular input) or global (i.e., what the model actually learned from data) [32,57]. An increasing level of explainability in the decision-making process also facilitates model traceability, which, in turn, could help reveal the possible points of failure and prevent future mistakes [4,57]. Making state-of-the-art ML-based models (i.e., deep neural networks) explainable is a quite challenging task, which needs to be directly addressed to cope with the right of explanation [24,58] but also for understanding other problems (i.e., unfair behavior, limited robustness, or leaks in privacy of the model itself [30,57]).

In this work we show how to adapt and combine state-of-the-art approaches with the purpose of learning from data under privacy, fairness, and explainability requirements. In particular, we show that it is possible to learn from data, leveraging on deep pretrained models [59,60], simultaneously preserving the privacy of the individuals thanks to the use of Homomorphic Encryption (HE) [61], ensuring fairness by learning a fair representation from the data [62–64], and delivering explainable decisions with local and global explanations [57] without compromising the accuracy of the final models. Then we will test our approach on a widespread and controversial problem, namely facial recognition, using the recent FairFace [65] dataset to prove the validity of our approach. In fact, deep pretrained networks allows one to easily and inexpensively extract a representation vector

that can be then used and fine-tuned for a specific application at hand [59,60,66]. This avoids the need to design and train from scratch a new network, which would require a huge number of data and computational resources, which is seldomly available in practical applications [67]. Nevertheless, even if an architecture is already available (with its tuned weights), we need to find smart ways to fine tune the network with limited data and a number of increasing constraints [68], especially when new data become available (e.g., the phenomena is changing) or new requirements arise (e.g., privacy and/or fairness requirements). HE has gained a lot of attention in the field of privacy-preserving machine learning since it allows working on encrypted data seamlessly, as the computations are performed on their original non-encrypted version [69]. Three main approaches exist (i.e., Partially, Somewhat, and Fully HE [70]), for which there is a tradeoff between the number of recoverable computations and the operations type allowed. For our purpose, namely ML-related applications, Somewhat HE is the most exploited approach, since it delivers the best trade-off [71] for our application. HE, on one hand, allows ensuring the privacy of the single individual, especially in the commonly adopted case where the computing and memory resources are outsourced to a third-party service provider, but on the other hand dramatically increases the computational requirements and reduces the possible network architectural choices [72,73]. Algorithmic Fairness deals with the problem of ensuring that the learned ML does not discriminate subgroups in the population using pre- in- and post-processing methods [26,74]. When deep models are exploited, such as in our use-case, learning a fair representation from the data (instead of simply trying to make the models fair) has been shown to be the best approach [62–64,75]. Still, these approaches can hardly be combined with HE since not all the operation and architectural choices are allowed due to the intrinsic limitations of HE [72,76]. For this purpose, in this work, we show that a particularly simple yet effective constraint for learning fair representation [64,77] can be combined with deep models and HE to deliver deep, fair and private models. In our work, fairness is measured according to the Demographic Parity (DP) [78], which requires the probability of the possible model decisions to be independent of the sensitive information. Finally, to deliver both local and global explainability, we rely on two state-of-the-art approaches. For local explainability, we exploit the attention maps of Deep Neural Networks through the Grad-CAM [79] algorithm, which highlights the most significant input features for a particular prediction. For global explainability, we will rely on both average attention maps and a dimensionality reduction algorithm, namely t-SNE [80,81]. Since these methods are straightforwardly applicable in conjunction with HE, they can be used to check the effect of the fairness constraints on what the deep models actually learned from data. In this sense, we are using explainability as a provision for the user right of explanation and as an inspection mechanism for the model creator as well.

The rest of the paper is organized as follows. Section 2 summarizes the works in the literature related to our research. Section 3 reports some preliminary notions instrumental for understanding our work. Section 4 presents the proposed method. The results of applying the method proposed in Section 4 on the face recognition task by means of the FairFace dataset are presented in Section 5. Section 6 concludes the paper.

2. Related Works

This section is devoted to a brief review of the works related to the context of our paper. For what concerns the fairness mitigating methods, they are usually categorized depending on the way they actually work [25,26,74,82]. For classical (i.e., shallow) ML models [83] trained on manually engineered features based on domain knowledge, we have three main families of mitigation methods: pre-, in-, and post-processing. Pre-processing methods try to remove the biases in the data so that any learning algorithm trained on those cleaned data should generate a fair model. In-processing methods impose the fairness constraints directly into the learning phase, enforcing fairness in the model's inner structures. Finally, post-processing tracks the output of an already trained model to make it more fair. When it comes to dealing with deep learning [66], where the ML models try to simultaneously

extract a synthetic yet expressive representation of the raw data (e.g., images or natural language) without any prior knowledge or human intervention, it has been recently shown [62,64,84–86] that the best approach is to learn a so-called fair representation. This fair representation can be reused to train other models which will be, again, fair by-design.

For what concerns the methods for making ML models privacy-aware/compliant, they can be divided into anonymization techniques, perturbation techniques, and distributed protocols [28,87,88]. The anonymization techniques try to maintain the privacy of the data subjects by obscuring personally identifying information within a dataset while preserving data utility. *k*-Anonymization [89], *l*-Diversity [90], and *t*-Closeness [91] are the most known approaches for anonymization. Perturbation techniques exploit noise to corrupt the data, the ML algorithm, or the learned model quantifying the disclosed information in terms of the power of the noise. Differential Privacy [92] is the most prominent theoretical framework for the perturbation techniques. Anonymization and perturbation techniques assume the existence of a trusted curator of the data. When this is not available, we need to use distributed protocol techniques [93].

For example, Federated Learning [94], one of the prominent approaches in distributed protocol techniques, requires participants to train their models privately and then to share the results. However, privacy may still be compromised once the local parameters are shared, such as the updating gradients, which that may disclose information on the user's private data. Recent works mix the use of Federated Learning with different HE schemes to address these issues [95,96]. In fact, HE recently attracted a lot of attention since it allows one to work on encrypted data as the computations are performed on their original non-encrypted version [97,98]. Consequently HE, contrarily to anonymization and perturbation techniques, entirely preserves utility and, contrarily to other simple distributed protocols techniques, automatically guarantees preserving the privacy of the single individuals. The major drawback of HE is its high computational overhead and the limitations for some operations [72,73]. In particular, three possible approaches are defined: Partially, Somewhat, and Fully HE. Partially HE benefits from an unlimited number of computations, but only one operation is allowed (multiplication: RSA [99], addition: Pailler [100]). Somewhat HE allows for multiple operations but suffers from a limited number of computations due to an increasing amount of computations-derived noise (BFV [101], CKKS [102]). Fully HE allows both a multiple number of operations and an unlimited number of computations, but generally suffers from huge computational costs (Gentry's [103]).

Concerning explainability, as a general rule, the complexity of a ML model is inversely proportional to its level of clarity and interpretability [31,57,104]. One solution to this issue is to design and implement intrinsically explainable algorithms. Alternatively, another widely used possibility is to build a highly accurate black-box model and then design a post-hoc explanation. Post-hoc explanation methods can be categorized into two families: global and local explanations. The former aims at understanding the entire logic of a system and retracing back the predictions' reasoning, while the latter is specific to a single instance and tries to justify single decisions. Post-hoc explanations are often model-agnostic, meaning that they are not tied to a particular type of ML systems [31,57]. Since explanations are mostly meant to be exploited by humans, they are usually meant to be visualized. For example, LIME [105] exploits a local surrogate model to explain the reason for a particular output. Relative to the context of computer vision, attention maps (e.g., using the Grad-CAM [79] algorithms) allow one to identify the influence regions of an image that most contribute to a particular decision. Both approaches can be used either as local methods of explanation if applied to only one image or as global methods of explanation if applied to a subset of the data [57]. In addition, dimensionality reduction methods, such as *t*-SNE [80], can be exploited as usually global explanations since they allow understanding how data, representation, and decisions are distributed and how changes in models or in the constraints (e.g., fairness) influence this distribution.

Many works [33–41] have tried to address fairness and privacy guarantees together. Kilbertus et al. [34] is one of the first proposals that addressed the need for combining

fairness requirements with privacy guarantees. Their approach is to mitigate Disparate Impact [106] (i.e., discrimination due to the correlation between sensitive and non-sensitive attributes) without disclosing sensitive information through secure multi-party computation. Jagielski et al. [35] builds on top of the previous work, stating that secure multi-party computation offers insufficient privacy guarantees due to the possible leakage of sensitive attributes. They provide a different approach based on Differential Privacy, where privacy is guaranteed through an injectable amount of noise able to mask the presence of a protected individual in a particular dataset. In this direction, other works [36–39] aimed to learn fair and differentially private ML models. Cummings et al. [36], while showing that it is impossible to achieve both differential privacy and exact fairness without non-trivial accuracy, provides a Probably Approximately Correct [107] learner that is differentially private and approximately (with high probability) fair. Xu et al. [38] presents two methods for achieving Differential Privacy and Algorithmic Fairness within a logistic regression framework through Functional Mechanism [108] that achieves privacy and fairness by injecting Laplacian noise into the model objective function. Mozannar et al. [37] proposes a two step algorithm where the first phase finds an approximately non-discriminatory predictor, while the second produces a final predictor with Local Differential Privacy guarantees [109]. Besides all the characteristics of standard Differential Privacy, Local Differential Privacy excludes the possibility in which an adversary is able to learn any sensitive information about a particular data point. On a related note, Bagdasaryan et al. [39] observed that standard Differential Privacy methodologies, such as gradient clipping and noise addition [110], yield disparate impact for underrepresented subgroups: the accuracy for those classes in a privacy-enhanced model tends to deteriorate more when compared to the non-private case. In this sense, this work empirically demonstrated that carelessly managing Differential Privacy will end up in exacerbating unfairness, hence supporting the need for alternative choices for pursuing privacy-preserving ML. Oneto et al. [33], instead, studies the privacy and fairness properties of randomized algorithms, proving that in this framework, it is possible to naturally impose fairness (measured with a generalized notion of fairness contemplating Equal Opportunity, Equal Odds, and Demographic Parity) and quantify the amount of disclosed information (via differential privacy) with theoretical guarantees. Unfortunately, the approach is still quite theoretical and practical evidence is still missing.

Some other works have tried to discuss the need for theoretical and practical ethical ML-enforcing privacy, fairness, and explainability properties [4,111–113]. Nevertheless, to the best of the authors' knowledge, in the literature, there are no works that simultaneously focus on enforcing privacy (especially HE), fairness, and (local and global) explainability in a theoretically grounded way and with actual empirical evidence in a realistic application.

In our work, we focus on a common face recognition problem using the recently released FairFace dataset [65]. Facial recognition is becoming a widespread and controversial tool used in many different contexts (e.g., from recreational activities to policing). Its popularity has increased so rapidly over the last few years that facial recognition software is commonly also used by government agencies [114]. Nevertheless, much recent evidence [115–117] shows how these algorithms can be biased against black people and women. In reaction to these issues, according to CNN [118], some governments banned the usage of facial recognition systems in law enforcement agencies and public-facing businesses. Making face recognition algorithms more trustworthy (fair, private, and explainable) would greatly improve the public opinion of them and their general acceptance.

Historically, traditional methods for facial recognition attempted to extract hand-crafted shallow features (e.g., Viola-Jones [119], Gabor [120], LBP [121]), and, before the advent of deep ML models, they represented the state of the art for classical benchmark datasets [122]. Deep learning models have recently been shown to outperform these classical methods, being more robust to changes in illumination, face pose, aging, expressions, and occlusions [123]. In particular, Convolutional Neural Networks (CNN) are

designed to be particularly proficient in facial recognition tasks and image recognition in general [124,125], employing a series of convolutional, pooling, and activation layers for extracting expressive representation from the input images. Moreover, the possibility of exploiting pretrained networks (e.g., LeNet [126], AlexNet [127], GoogleNet [128], VGGNet [129], and ResNet [130]) as-is or fine-tuned represents the state-of-the-art approach for different computer vision tasks [124]. In this work, we rely on the VGGNet architecture since it offers a good trade-off between accuracy, computational resources, and ease of use. Moreover, VGGNet differs by a few percentage points in accuracy from other state-of-the-art deep neural networks [122,131–133].

3. Preliminaries

Let us consider the probability distribution μ on $\mathcal{I} \times \mathcal{S} \times \mathcal{Y}$, where \mathcal{I} is the input space, $\mathcal{S} = \{1, 2\}$ identifies a binary sensitive variable (in our case the binary gender, i.e., male and female) and $\mathcal{Y} = \{0, 1\}$ is a binary label (in our case $<$ and \geq of 30 years old). For \mathcal{S} , our method easily extends to multiple sensitive variables and continuous variables, but to ease the presentation, we consider only the binary case in the paper. In our work, $\mathcal{I} \subseteq \mathbb{R}^{h \times w \times 3}$ is the space of all RGB images of human faces, where h and w are the height and width of the image, while the third dimension defines the three standard color channels (Red, Green, and Blue). Let $\mathcal{D} = (I_i, s_i, y_i)_{i=1}^n \in (\mathcal{I} \times \mathcal{S} \times \mathcal{Y})^n$ be a set of n samples from μ . For each $s \in \{1, 2\}$, let $\mathcal{D}_1 = \{(I, s, y) \in \mathcal{D} \mid s = 1\}$ and $\mathcal{D}_2 = \{(I, s, y) \in \mathcal{D} \mid s = 2\}$ be the set of samples in the first and second group, respectively. The goal is to learn a model $h : \mathcal{Z} \rightarrow \mathcal{Y}$ able to approximate $\mathbb{P}\{y \mid Z\}$ where $Z \in \mathcal{Z}$ may contain ($\mathcal{Z} = \mathcal{I} \times \mathcal{S}$) or not ($\mathcal{Z} = \mathcal{I}$) the sensitive attribute, depending on the specific regulation [134,135]. The ability of h of approximating $\mathbb{P}\{y \mid Z\}$ is measured with different indices of performance $P(h)$ based on the required properties and the different tasks under examination [66]. For example, in binary classification $P(h)$ can be the Accuracy or the Mean Square Error.

Within the context of the increasingly popular deep ML models, h can be described as a composition of simpler models $m(r(Z))$, where $m : \mathbb{R}^d \rightarrow \mathcal{Y}$ is a (non-)linear function and $r(Z) \in \mathbb{R}^d$ is a function mapping the input data into a vector, usually referred to as the representation vector. Note that r can be a composition of functions as well $r : r_1 \circ \dots \circ r_2 \circ r_1$, for example, in a deep neural network of l layers [66]. In other words, the function r creates a compact and expressive description of the input space that can deliver high accuracy when used by m to solve a specific task. r , learned in a particular context, can be reused by many models m as it is or fine tuned for the specific task at hand.

According to Algorithmic Fairness, we expect the model h to be fair with respect to one or more notions of fairness [26]. As recently theoretically studied in [64] and empirically demonstrated in many works [62,63,136–139], when deep learning models are developed, learning a fair representation actually allows one to make the entire network fairness-aware. Intuitively, this fair representation could be subsequently exploited by other ML models, for example, within the context of Transfer Learning [140], enforcing fairness by-design. In our work, we require the representation vector to satisfy the DP constraint [78]. Other notions of fairness could be exploited in this paper such Equal Opportunity and Equal Odds [141], but this extension is straightforward and out of the scope of this paper.

$$\mathbb{P}_Z\{r(Z) \in \mathcal{C} \mid s=1\} = \mathbb{P}_Z\{r(Z) \in \mathcal{C} \mid s=2\}, \forall \mathcal{C} \subseteq \mathbb{R}^d, \quad (1)$$

namely, the two conditional distributions of the representation vector should be equal with respect to the sensitive attribute. The constraint of Equation (1) directly implies that any model m learned on top of a fair representation will be again fair

$$\mathbb{P}_Z\{m(r(Z))=1 \mid s=1\} = \mathbb{P}_Z\{m(r(Z))=1 \mid s=2\}. \quad (2)$$

The performance $P(h)$ of the final models h will be evaluated with the accuracy metric ($\text{ACC}_y(h)$), namely percentage of correctly classified samples, computed on the test set (i.e.,

data not exploited to train h) [142]. Exploiting Equation (1), the fairness of the final models h will be measured by means the Difference of Demographic Parity (DDP) [64]

$$\left| \frac{1}{|\mathcal{D}_1|} \sum_{(Z,y) \in \mathcal{D}_1} [h(Z)=1] - \frac{1}{|\mathcal{D}_2|} \sum_{(Z,y) \in \mathcal{D}_2} [h(Z)=1] \right|, \quad (3)$$

where the Iverson bracket notation is used.

We will rely on HE for enforcing privacy guarantees. In linear algebra, a homomorphism is a transformation between two algebraic structures that preserves the defined operations. For example, let $\phi : \mathcal{A} \rightarrow \mathcal{B}$ be a homomorphic map between two sets \mathcal{A} and \mathcal{B} with the same algebraic structure, if \oplus is a binary operation on that structure, then $\phi(A_1 \oplus A_2) = \phi(A_1) \oplus \phi(A_2)$, $\forall A_1, A_2 \in \mathcal{A}$. Hence, HE is an encryption protocol that relies on homomorphic transformations obtained through the definition of public (i.e., encryption) and private (i.e., decryption) keys. Thanks to the property of homomorphism, some operations can be performed on the encrypted data as they were carried on the non-transformed version preserving the privacy of the original data. Specifically, we will rely on Somewhat HE using the CKKS scheme, which allows a bounded number of computations limited to addition, multiplications, and rotations. The CKKS algorithm defines four phases: encoding, encrypting, decrypting, and decoding [102]. First, the input data, which consist of a vector of real values, are encoded into a polynomial (i.e., the plaintext) of degree p , where p is a power of 2. CKKS works with cyclotomic polynomials from the Ring theory because they offer a good trade-off between security and efficiency [102]. The plaintext is then encrypted into a pair of different polynomials (i.e., the ciphertext) through the use of a public encryption key. The homomorphism of this encryption is achieved thanks to the theory of Ring Learning With Error [143], where, of particular interest for this work, addition and multiplication are preserved. While additions cause no obstacles, multiplications increase the noise kept in the pair of ciphertexts; therefore, only a limited number of products is allowed. However, higher polynomial degrees allow for wider computational bounds, yet they are more expensive in terms of processing and memory requirements. Once the required computations are performed, the pair of ciphertexts can be reverted back first to the plaintext polynomial through the use of the secret decryption key, and then to the vectors of values through the final decoding phase. The output vectors will yield approximate results, close to the real solution thanks to the property of homomorphism. The polynomial degree p must be chosen as small as possible to guarantee the correctness of the results without increasing too much the computational requirements [102].

In order to improve the readability of the technical parts, we added the list of notations in Table 1 that are exploited in the paper.

Table 1. Notations.

Symbol	Description
\mathcal{I}	General input space of RGB images
h	RGB images height
w	RGB images width
\mathcal{S}	Binary sensitive attribute space
s	s^{th} sensitive group
\mathcal{Y}	Binary label space
\mathcal{D}	Full dataset
\mathcal{D}_s	Samples from \mathcal{D} in the s^{th} sensitive group
\mathcal{Z}	Model input space that may (or not) contain the sensitive attribute
Z	Model input

Table 1. Cont.

Symbol	Description
h	General end-to-end model
h^*	Learned end-to-end model
r	Sub-model that learn the data representation (embedding layers)
m	Sub-model that learn the task from the representation (task specific layers)
P	General utility measure
ACC	Accuracy Measure
ϕ	Homomophic map
\mathcal{A}, \mathcal{B}	Sets with same algebraic structure
p	Degree of the encoded cyclotomic polynomial
F	General fairness measure
λ	Fairness regularization hyper-parameter (Tikhonov formulation)
η	Fairness regularization hyper-parameter (Ivanov formulation)
DP	Demographic Parity
DDP	Difference of Demographic Parity
AVG	First order (convex and differentiable) approximation of the DDP
y	Classification prediction target
y_n	Non-normalised classification model prediction for y
A	Convolutional layer output
A_k	Matrix relative to channel k in layer output A
G_{y_n, A_k}	Gradients matrix of y_n w.r.t. A_k
α_{y, A_k}	Importance weight of A_k w.r.t. the target y
L_y	Grad-CAM map w.r.t. the target y
M_s	Dataset average Grad-CAM for the s^{th} sensitive group
FRO	Frobenius distance
KS	Kolmogorov–Smirnov distance

4. Proposed Method

In this section we will present our approach to learn private, fair, and explainable deep ML models. In particular, we will start presenting our approach to private deep ML models based on HE showing the limitations that are implied in terms of computations and operations (Section 4.1). Then we will present the chosen architecture, with particular reference to the exploited facial recognition application. The proposed architecture slightly differs from the classical one due to the handling of the HE limitations (Section 4.2). Following this analysis, we will show how to impose the fairness constraint, again taking into account the limitation imposed by HE, using the fair representation framework (Section 4.3). Finally, we will empower the proposal with explainability properties that will be used also to understand what is learned from the deep model and whether the fairness constraint actually changes how and what the architecture perceives from the images (Section 4.4).

4.1. Making the Model Private

As previously mentioned, to enforce privacy, we relied on HE during both the training and forward phases of the deep ML models. During training time, each sample is encrypted following the CKKS [102] scheme to a high order polynomial that masks the real data attributes and labels. Then these encrypted values are fed to the DNNs, which

output encrypted predictions. Thanks to the homomorphism property, the masked labels and predictions can be compared through a loss function. The loss function needs to be expressed in terms of additions and multiplications (the only operations allowed by CKKS) so a polynomial loss function is the most natural choice (e.g., the Mean Square Error [144]). During the training phase, we rely on Gradient Descent algorithms [66,145], which natively require us to compute just additions and multiplications. This is true only if the architecture of the deep ML model does not contain special (non-polynomial activations) functions whose derivative cannot be expressed easily with additions and multiplications (this limits our architectural choices; e.g., the widely-used RELU activation function cannot be deployed). Belonging to Somewhat HE, CKKS adds a certain amount of noise to the encrypted data, which increases with the number of stacked layers [72,76]. This fact also limits the depth of the network. Finally, the CKKS scheme heavily increases the memory and computational requirements for storing and processing the data, further limiting the architectural choices and the number of data that we can use to train the network. Note that the privacy of the deep network can be enhanced by also encrypting the weights of the network [146] (e.g., to avoid or at least mitigate adversarial attacks (<https://blog.f-secure.com/mitigations-against-adversarial-attacks> (accessed on 11 August 2021))). The process of encryption/decryption is performed through the python TenSEAL [147] library for the CKKS scheme which easily allows the integration with common deep ML software frameworks like PyTorch [148]. Other libraries, even more efficient, exist [149,150], but they can be hardly combined with deep ML software frameworks.

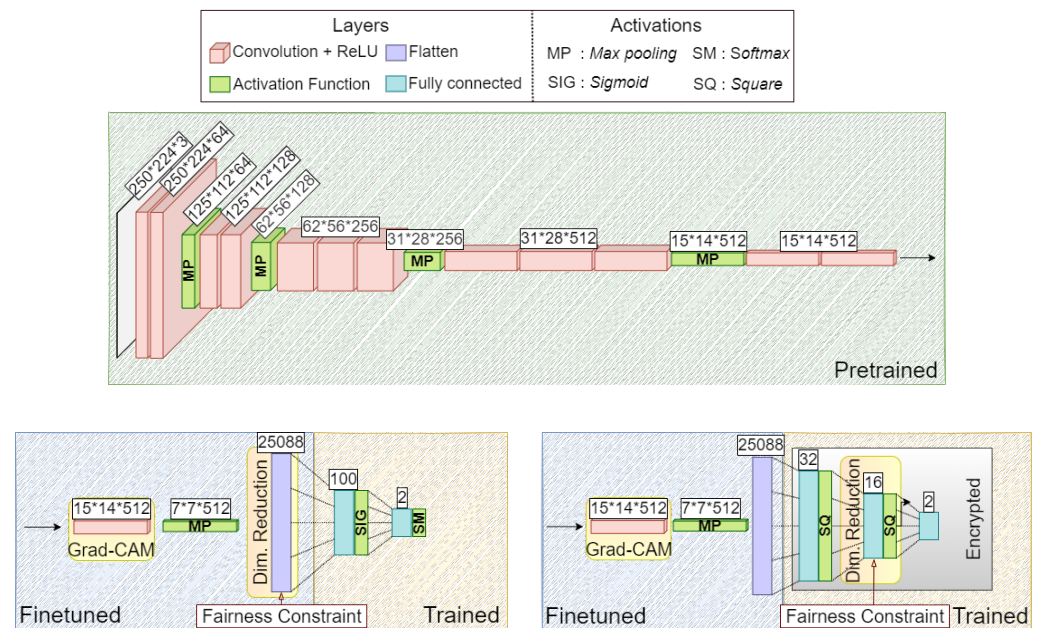
4.2. The Architecture of the Deep Model

In this work, we will exploit the VGGNet-16 [129] (Configuration D) as the architecture for face recognition. VGG-based networks exploit deep architectures, leading to quite accurate results for a variety of different tasks while maintaining relatively low computational requirements thanks to the use of small filters. In fact, stacking convolution layers with small kernels seems to be preferable to using a single layer with larger kernels, providing computational and generalization advantages [129]. Moreover, the use of multiple stacked layers allows to easily increase the nonlinearity harnessed by the network by adding an activation function at each intermediate step. The VGGNet-16 embeds the input data in a 25,088-dimensional vector space by means of 14 million parameters, which allow linear state-of-the-art results to be achieved in multiple facial-recognition-related tasks [122,131–133]. Because of such complexity, the VGGNet-16 deployed in this work has been pretrained on the face recognition dataset VGG-Face [151].

However, in this work, we will need to depart from the standard end-to-end use of deep learning models due to the limitation imposed by HE as expressed in Section 4.1. The convolutional layers need to be kept fixed, i.e., used just to extract the embedding. In fact, fine-tuning them would require using end-to-end HE for the whole architecture, resulting in an intractable problem from the computational point of view. Actually, the 25,088-dimensional embedding cannot be directly used, for the same reason above, but we have to shrink it in a much smaller embedding, i.e., a 32-dimensional, by means of a dense layer with Sigmoidal activation. The parameters of this layer, along with those of the last convolutional layer, are initially pretrained on the FairFace dataset. To give an idea of why we chose 32 as the dimensions of the embeddings, we point out that managing 1000 images embedded in this 32-dimensional vector requires approximately 30 GB of memory (see details in Section 5).

The actual learning phase was conducted starting from these fixed 32-dimensional embedding using a single hidden layer architecture empowered with HE. The 32-dimensional vector is fed into a 16-dimensional dense layer with a square activation function, which complies with the CKKS limitation. The output of this layer is fed into a 2-dimensional dense network with linear activation. We did not use a single output neuron for being able to exploit the Grad-CAM visualization algorithm (see Section 4.4). The parameters of these last layers are randomly initialized according to a Gaussian distribution $N(0, 0.01)$.

Figure 1 represents both a classical architecture [45] (Figure 1a) and the proposed one (Figure 1b) for facial recognition.



(a) Classical VGGNet-16-based face recognition architecture. (b) Proposed VGGNet-16-based face recognition architecture.

Figure 1. Classical and proposed architectures for facial recognition. Figure shows the pretrained, the trained, and the fine tuned layers, where the HE is employed, the fairness constraints, and the visualization methods are applied.

4.3. Making the Model Fair

As reported previously, many different approaches exist to impose the fairness constraint of Equation (1). In particular, following the fair representation principle [45], we propose the formulation of a Tikhonov regularizer $F(h)$ for balancing the possibly biased performance index $P(h)$ (Section 3) in the cost function as follows:

$$h^* = \arg \min_h P(h) + \lambda F(h), \tag{4}$$

where $\lambda \in [0, +\infty)$ trades off accuracy and fairness, as we will also see in Section 5. Note that the constraint could have been imposed using the Ivanov philosophy [152], and the results would be the following optimization problem

$$h^* = \arg \min_h P(h), \quad \text{s.t. } F(h) \leq \eta, \tag{5}$$

where $\eta \in [0, 1]$ regulates the level of accepted fairness, which is cognitively more close to the problem of imposing a certain level of fairness to the final model. Nevertheless, note that, for some values of η and λ , the two problems of Equations (4) and (5) are equivalent, but Problem (4) is much less computationally demanding with respect to Problem (5) [153]. Note also that setting $\eta = 0$ in Problem (5) (or $\lambda \rightarrow +\infty$ in Problem (4)) to impose the DP does not guarantee fairness in terms of generalization since Problem (5) (or Problem (4)) exploits empirical quantities. Setting $\eta \in [0, 1], \lambda \in [0, +\infty)$ allows one to avoid overfitting the particular sample.

The concept of learning a fair representation is expressed, in this context, by imposing the regularizers on the last layer of the representation, namely $F(h) \rightarrow F(r)$. This translates, in the classical architecture, into imposing the constraint in the last convolutional layer as the most effective strategy [45] (see Figure 1a), while, for the proposed architecture,

in imposing the constraint in the last tunable representation layer, i.e., the 16-dimensional last hidden layer (see Figure 1b).

Unfortunately, the fairness constraint of Equation (3) is practically hard to handle, and it is necessary to approximate it by defining effective yet computationally efficient alternative regularizers, which also have to meet the HE limitations. In the literature, different approaches have been proposed, and the most effective ones appear to be the one reported in [64]. The authors of [64] propose three different regularizers: one based on convex approximation and relaxation of the constraint of Equation (1), one based on the squared Maximum Mean Discrepancy [154], and one based on the Sinkhorn divergence [155].

Because of the limitations imposed by HE, the approaches based Maximum Mean Discrepancy and Sinkhorn divergence cannot be effectively employed. We rely on the convex approximation and relaxation of the constraint of Equation (1) proposed by [64] where the regularizer assumes the following form:

$$\text{AVG}(r) = \left\| \frac{1}{|\mathcal{D}^1|} \sum_{(Z,y) \in \mathcal{D}^1} r(Z) - \frac{1}{|\mathcal{D}^2|} \sum_{(Z,y) \in \mathcal{D}^2} r(Z) \right\|^2. \quad (6)$$

Note that this convex approximation and relaxation is simply the first-order approximation of Equation (1). Note that if the chosen architecture can be handled with H, this regularizer added in the cost function simply adds a term which can be computed with sum and multiplications, such as its derivatives.

4.4. Making the Model Interpretable

In order to provide both local and global explanations and to visualize how the CNNs react to the input images, we analyze the attention regions obtained through Grad-CAM [79]. Specifically, Grad-CAM extracts attention maps (i.e., heatmap images) that highlight the most influential features for a particular supervised task. They can be used as a local explanation method if applied to a single instance or global if the result is averaged over a subset of the instances (e.g., over all the men older than 30 years old). When dealing with fairness, attention maps can underline any divergence in the representations between different protected groups.

By fixing a classification prediction target $y \in \mathcal{Y}$ (i.e., the output neuron of the network corresponding to particular class), a non-normalised network score y_n for y (i.e., prior to the final Softmax activation for classical architecture or simply the output for the proposed one in Figure 1), and a convolutional layer output $A \in \mathbb{R}^{K \times U \times V}$ (where we extract the matrix $A_k \in \mathbb{R}^{U \times V}$ relative to the channel $k \in \{1, \dots, K\}$ - U, V are the output matrices dimensions for any of the k channels), then the gradient $G_{y_n, A_k} \in \mathbb{R}^{U \times V}$ of y_n with respect to A_k is defined as

$$G_{y_n, A_k} = \frac{\partial y_n}{\partial A_k}. \quad (7)$$

The importance weight of the channel k with respect to the class y is naturally obtained as the average $\alpha_{y, k}$ across the convolutional layer matrix entries

$$\alpha_{y, A_k} = \frac{1}{UV} \sum_{i=1}^U \sum_{j=1}^V G_{y_n, A_k, i, j}. \quad (8)$$

Finally, the Grad-CAM map with respect to a target y is defined as L_y , namely the weighted sum across all the dimensions k , is

$$L_y = \text{ReLU} \left(\sum_{k=1}^K \alpha_{y, A_k} A_k \right), \quad (9)$$

where the ReLU function [127] simply suppresses all the negative values highlighting the interest for the features that have only a positive influence towards a certain target.

Since the first part of the architecture is unencrypted (see Figure 1b), the network can be inspected easily by either the user (using its own private data) or the model designer (using a set of data not constrained by privacy issues) both when the network parameters are encrypted or unencrypted.

In our work, we extract the Grad-CAM attention maps relative to the last convolutional layer of VGGNet. Usually, earlier convolutions extract low-level features (e.g., edges or corners), while deeper convolutional layers are able to describe more abstract features, such as geometrical shapes or complex connected regions [79], which are extremely significant for tasks such as facial or image recognition. Note that Grad-CAM allows inspecting the network perception even if some of the deep layers are kept fixed and just the last layers are modified or fine-tuned (like in our case, see Figure 1b). In fact, the perception is propagated from the output to the inner convolutional layers, which allows to track back changes on the last weights [79].

Although gradient-based methods might not be the optimal solution for visual explanation (e.g., saturation, zero-gradient image regions, and false confidence in the output score phenomena [156]), the computational cost of Grad-CAMs is negligible when compared to other methods that require multiple network forward-passes per image [156,157]. Moreover, Grad-CAM is considered the reference method in several recent works [157–161].

The second implemented approach to globally explain the deep network behavior consists in observing whether the network maps the input data into a feature space able to both preserve performances and mask the membership in a protected population. Fixing an internal network layer, this task can be performed by reducing the dimensionality of the layer's original space to a lower-dimensional (possibly two) and more interpretable one. In our work we rely on the t-SNE algorithm [80,81] for effectively carrying out this dimensionality reduction. As an unsupervised approach, it allows one to evaluate the statistical distribution of the extracted features hiding task-related information, which may produce undesired distortions.

t-SNE firstly calculates the similarity between points both in the high-dimensional space and in the corresponding low-dimensional one. The similarity is calculated as the conditional probability that a point P_1 would choose point P_2 as its neighbor following a Gaussian distribution centered at P_1 . Then, it tries to minimize the difference between these conditional probabilities in the higher-dimensional and lower-dimensional spaces by minimizing the sum of Kullback–Leibler divergence of overall data points using a gradient descent method.

In this work, we applied directly t-SNE on the 16-dimensional embedding (since it is the only one which varies with the training phase) for the proposed architecture (see Figure 1b).

For what concerns the classical architecture (see Figure 1a), instead, the 25,088-dimensional embedding is too big to be fed directly to the t-SNE algorithms. For this reason, we will adopt a two-step approach for effectively reducing its dimensionality. The first step of this feature reduction is supervised (by means of L_1 -regularized Logistic Regression [162]), while the second one is un-supervised (by means of the t-SNE). The first step allows us to remove the features with zero contributions to the specific task under examination. The second step allows us instead to evaluate the statistical distribution of the remaining features hiding task-related information which may produce unwanted distortions. As usually happens in deep networks, the representation vector has a large number of elements (to allow its use in multiple tasks) but only a subset of them is needed to solve a specific task. Exploiting a L_1 -regularized Logistic Regression allows to discard the features with no contribution to the task solution (by means of the L_1 -regularization [163]) reducing the dimension of the space to just the informative features for the considered task. Since t-SNE is a computationally demanding algorithm, usually a PCA-based [164] pre-dimensionality reduction step is adopted.

5. Experimental Results

In this section we present the results of applying the methodology presented in Section 4 on the FairFace real-world dataset [65]. In particular, in Section 5.1 describes the FairFace dataset. Then, Section 5.2 describes the architectural configurations tested in the study (i.e., with and without HE and/or fairness constraints). Section 5.3 reports the performance of this architecture in terms of accuracy and fairness, while Section 5.4 focuses on their computational requirements. Finally Sections 5.5 and 5.6 focus on local and global explainability, respectively, to give more insights into what the different architecture actually learned from the data and what the effects of introducing privacy and fairness requirements are. All the codes for producing the results are made freely available to the community (https://github.com/lucaoneto/ENTROPY_2021 (accessed on 11 August 2021)).

5.1. The Dataset

The FairFace dataset [65] is a collection of ≈ 100 thousand facial images extracted from the YFCC-100M Flickr dataset [165]. Automated models trained on FairFace can exploit age group (age ranges of [0–2], [3–9], [10–19], [20–29], [30–39], [40–49], [50–59], [60–69], and [70+]), gender (which, for this dataset, refers to the perceived binary gender (Male and Female) of an individual), and ethnicity (Western White, Middle Eastern White, East Asian, Southeast Asian, Black, Indian, and Latinx.) as sensitive information. Our task consists in predicting whether a face belongs to a person with more (1) or less (0) than 30 years old measuring the discrimination between individuals with different gender. Table 2 reports some statistics about the FairFace dataset with respect to the selected sensitive attribute.

Table 2. Fairface: label distribution (gender is the sensitive features).

	Age ≥ 30	Age < 30	<i>Sensitive Marginals</i>
Females	18.60%	28.40%	47.00%
	18,174	27,746	45,920
Males	27.21%	25.79%	53.00%
	26,587	25,191	51,778
<i>Class</i>	45.82%	54.18%	
<i>Marginals</i>	44,761	52,937	97,698

For this work, the training and test sets are composed, respectively, of 86.7 thousand and 10.9 thousand images (same split of the original FairFace dataset [65]).

5.2. Tested Configurations

In this section, we summarize all the architecture that we tested in the experiments:

- The classical VGGNet-16-based face recognition architecture (see Figure 1a) under the following settings:
 - The architecture was trained with a random selection of 20,000 training and 10,000 test images from the training and test sets, respectively. We train every model for a total of 10 epochs using the ADADELTA [166] gradient descent method with mini batches of 150 images. The layers before the last convolution one (excluded) were not fine-tuned and would benefit from the parameters pre-trained on the VGG-Face dataset (see Section 4.2);
 - We investigated the case when the fairness constraint (see Section 4.3) is or is not imposed in the last convolutional layer;

- We also derived the attention maps and the dimensionality reduction with respect to the last convolutional layer.
- The proposed VGGNet-16-based facial recognition architecture (see Figure 1b) under the following setting:
 - The architecture was trained with a random selection of 1000 training (because of the limitation imposed by HE; see Section 4.1) and 10,000 test images from the training and test sets, respectively. We trained every model for a total of 10 epochs using gradient descent [66]. Before the actual training could take place, the embeddings needed to be reduced to a much smaller representation vector due to the computational limitation imposed by HE (see Section 4.1). To perform this task, we trained the architecture depicted in Figure 1b without applying HE. We chose a 32-dimensional representation since it represents a good tradeoff between information compression (due to the HE limitations) and utility (the accuracy of the whole network remains unaltered). This preliminary phase observes the same settings imposed for training the classical architecture. Once the network parameters are trained for extracting the 32-dimensional representation vector, we reset the weights of the last two dense layers following again the original Gaussian distribution $N(0, 0.01)$. This simulates the case when a new network is trained from scratch by applying privacy guarantees through HE. The layers before the last convolution one (excluded) were fine-tuned and could benefit from the parameters pre-trained on the VGG-Face dataset (see Section 4.2).
 - We investigate the case when HE was or was not exploited (see Section 4.1) in the last three layers of the network (see Figure 1b).
 - We investigated the case when the fairness constraint (see Section 4.3) was or was not imposed in the last hidden layer;
 - We derived the attention map with respect to the last convolutional layer. We applied, instead, the dimensionality reduction to the last hidden layer.

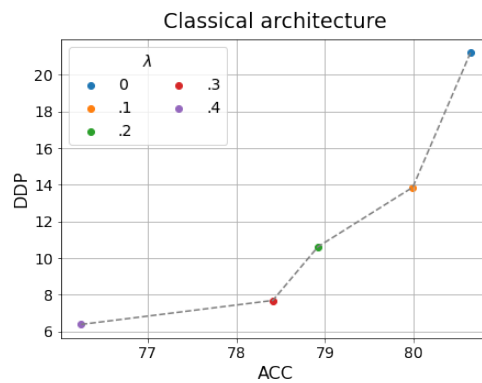
5.3. Accuracy vs. Difference of Demographic Parity

In this section, we evaluate the different architectures in terms of accuracy ACC and fairness DDP, on the test set.

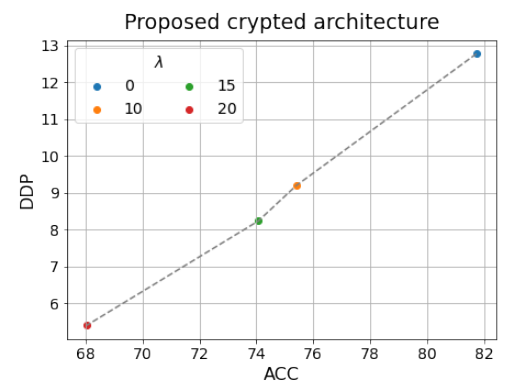
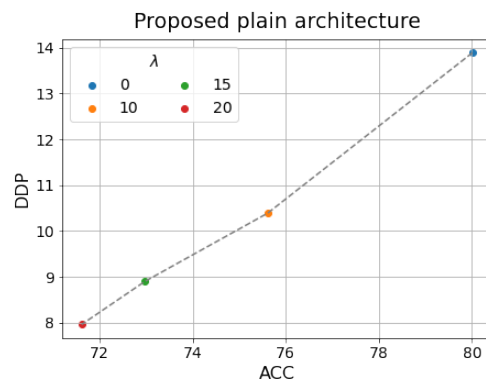
In particular, Figure 2 reports the ACC against the DDP for the different architectures (see Section 5.2) when different values of λ are exploited.

Two tendencies can be observed in Figure 2. The first one refers to the tension between accuracy and fairness: the more fair we want the model to be (the higher value for the regularization parameter λ), the less accurate the model will be on the available data (i.e., data are biased and then, trying to remove these bias, not fully trustable). The second one refers to the tension between accuracy and privacy: enforcing privacy with HE actually reduces our ability to use large amounts of data, computation, and architectural choices and hence reduces our ability to learn accurate models. Nevertheless, the results of the proposed architecture gives similar results as expected from the theory (see Section 4.1) whether HE is present or not, while the small differences are obviously due to the noise introduced by HE in the computation.

Figure 2 clearly shows the effectiveness of the proposed approaches in learning private and fair models.



(a) Classical Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint.



(b) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and no HE. (c) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and HE.

Figure 2. Comparison between the different architectures in terms of accuracy ACC and fairness DDP.

5.4. Computational Requirements

In this section, we would like to report and underline the computational requirements for training the different architectures under examination (see Section 5.2).

In particular, Table 3 reports the training time and memory requirements averaged over different runs for the different architectures described in Section 5.2. Experiments were run on a machine equipped with Windows Server 2019, 4 Intel Xeon CPU E5-4620, 256GB DDR3 of RAM, 1 TB SSD disk, Python 3.7, scikit-learn 0.24.2, PyTorch 1.8, and TenSeal 0.3.4.

Table 3. Training time and memory requirements averaged over different rounds for the different architectures.

Architecture	Fairness Constraint	Homomorphic Encryption	Number of Training Samples	Training Time (sec/epoch)	Memory Requirements (GB)
Classical			20,000	6200	≈10
(Figure 1a, full)	x		(batches of 150)	6100	≈10
Proposed				< 1	≈1
(Figure 1b, training last two dense layers)	x		1000	< 1	≈1
		x		2100	≈30
	x	x		3000	≈30

From Table 3, it is possible to note the explosion, in terms of computational requirements, when the HE is employed. This is expected from the theory (see Section 4.1), and this is the reason behind the architectural choices (the reduction of the embedding dimension from 25,088 to 32) and the limitation in the size of the training set from 20,000 to 1000). Nevertheless, the results of Section 5.3 have shown how these limitations actually do not compromise the ability to learn fair and accurate models.

5.5. Attention Maps

In this section, we aim at assessing a possible discriminatory attention behavior carried out by the different architectures (see Section 5.2) tested in Sections 5.3 and 5.4.

In fact, following the method presented in Section 4.4, we wish to observe whether the application of the fairness constraint produces less discriminatory attention mechanisms, namely more similar attention maps between different subgroups. In order to standardize the image face regions, we exploited a set of 50,000 images of frontal faces extracted from the Diversity in Faces dataset [167], where, again, gender was exploited as the sensitive feature.

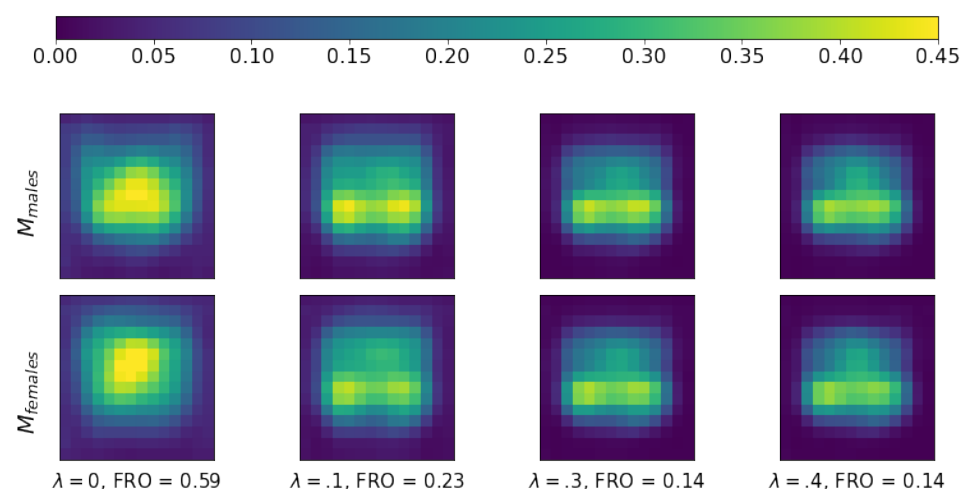
Firstly, we took the average attention maps of both males and females, and we computed the difference between these two average attention maps through the Frobenius distance [168]. More formally, for each image in the dataset, let us compute the grad-CAM attention map L_Y (see Section 4.4). Then, let us define $M_s \in \mathbb{R}^{U \times V}$, with $s \in \{\text{males, females}\}$, as the dataset averaged L_Y for each subgroup in the population. Finally, the Frobenius distance of M_{males} and M_{females} is computed as

$$\text{FRO}(M_{\text{males}}, M_{\text{females}}) = \sqrt{\sum_{i=1}^U \sum_{j=1}^V (M_{\text{males},i,j} - M_{\text{females},i,j})^2}.$$

Figure 3 reports M_{males} , M_{females} , and $\text{FRO}(M_{\text{males}}, M_{\text{females}})$ for the different architectures with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint. Figure 3 clearly shows the positive effect of the fairness regularizes in reducing the networks' discriminatory attention mechanism, which is quite evident if compared to the no-regularized case.

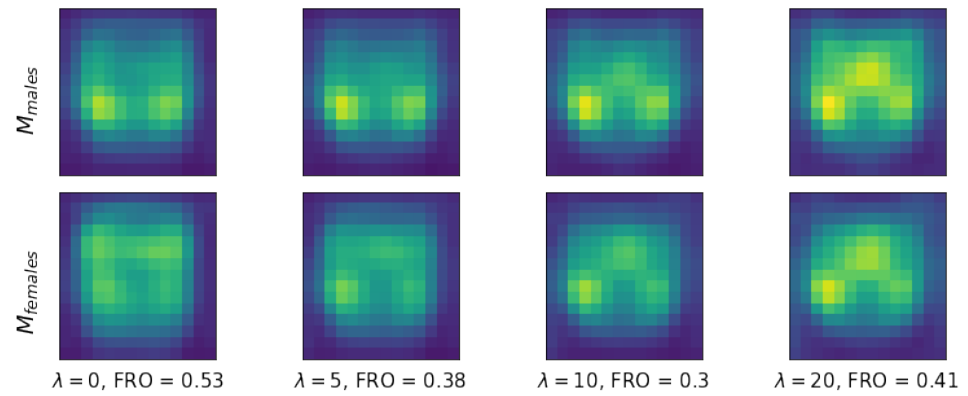
For sake of completeness, we also report in Figure 4 the attention map for the different architectures with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint for a young male, a young female, an old male, and an old female. Due to space limitations, we report just the results with a λ that showed the best accuracy/fairness in the results of Section 5.3.

Figure 4 clearly shows how the fairness regularizer is able to restrict the networks' receptive field to class-specific face regions.

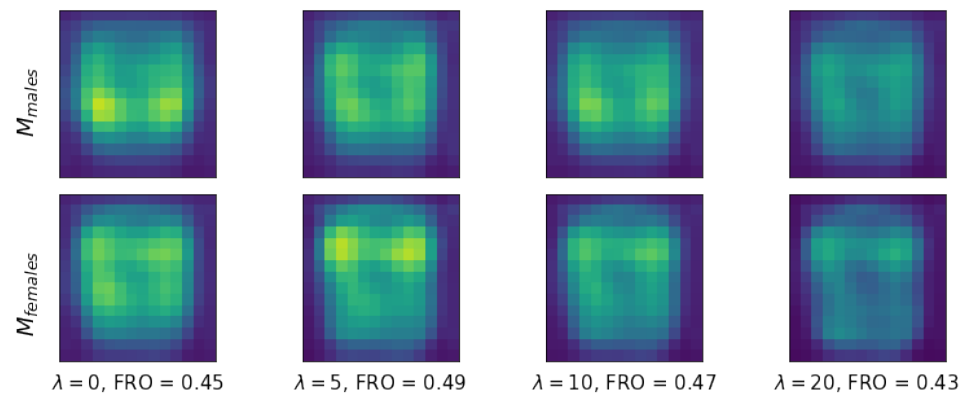


(a) Classical Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint.

Figure 3. Cont.



(b) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and no HE.



(c) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and HE.

Figure 3. Comparison between the different architectures using the average attention map extracted with Grad-CAM.

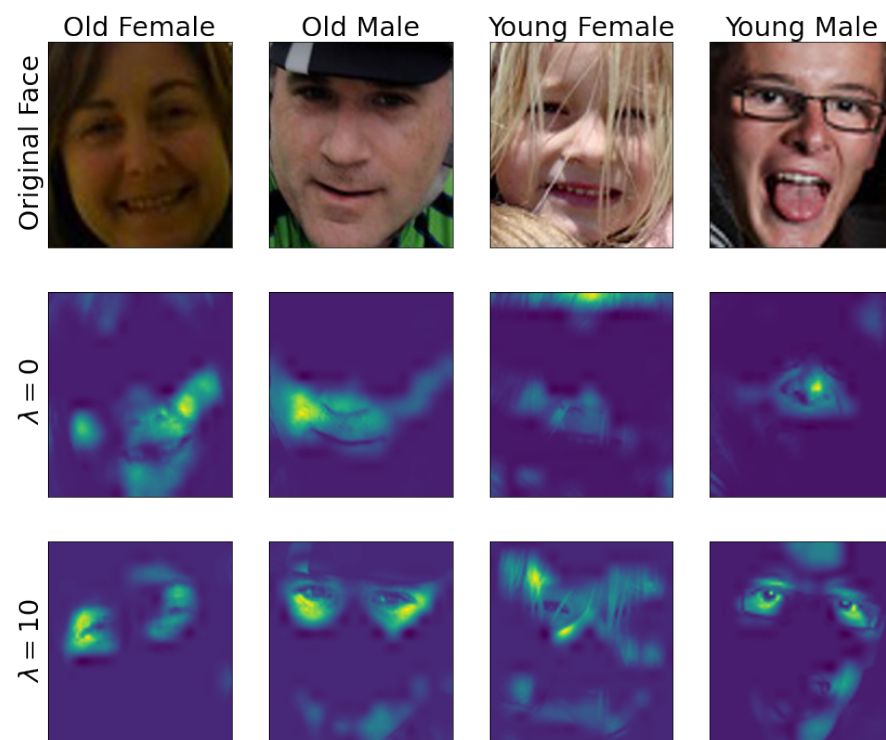


(a) Classical Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint.

Figure 4. Cont.



(b) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and no HE.



(c) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and HE.

Figure 4. Comparison between the different architectures using the attention map of sample images extracted with Grad-CAM.

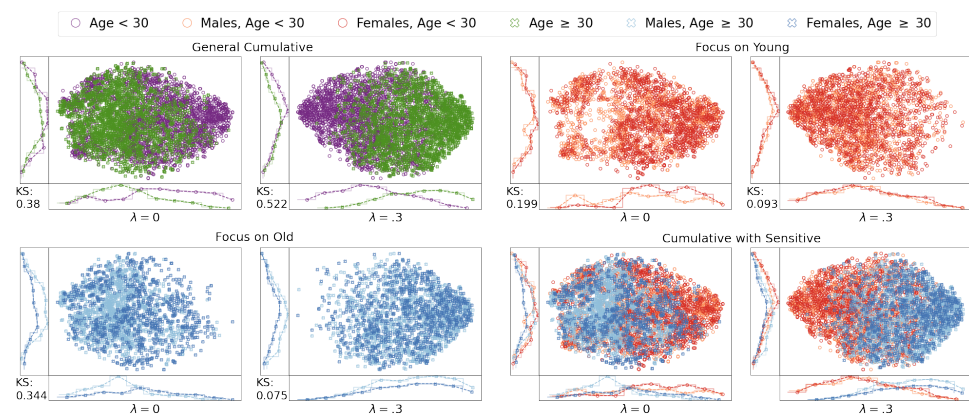
5.6. Dimensionality Reduction

In this last set of experiments, we analyzed the distribution of the representation vectors (see Section 5.2 and Figure 1) by means of dimensionality reduction using the pipeline presented in Section 4.4.

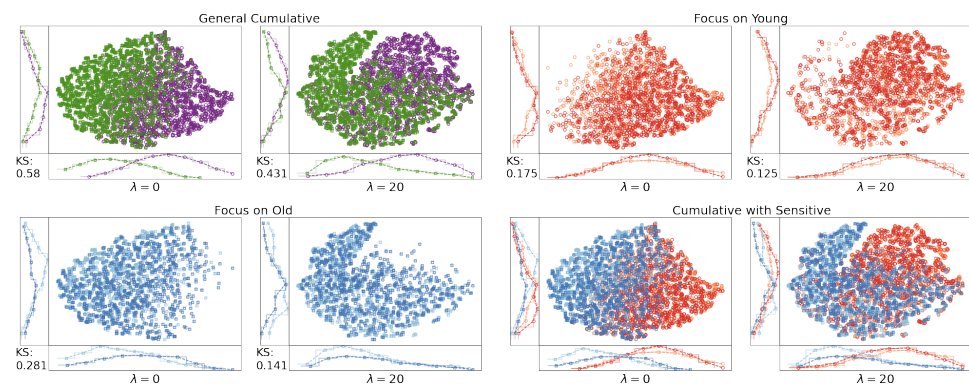
For what concerns the classical architecture (Section 4.4), we cross-validated the L_1 -regularization strength in the L_1 -regularized logistic regression obtaining a top accuracy above 80% for the binary classification task. This step discarded most of the representation vector features extracted by the VGGNet-16, keeping just ≈ 450 features (i.e., those with weights different from 0). Then, the PCA further reduced the dimensionality of the space from ≈ 450 to 50 features. Finally, the t-SNE has been exploited to map this 50-dimensional space into a 2-dimensional space.

For what concerns the proposed architectures (Section 4.4), instead, the t-SNE was exploited to directly map the hidden 16-dimensional embedding into a 2-dimensional space.

Figure 5 displays the results for the different architectures, with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint, on a random selection of 3000 samples from the test set to allow a friendly visualization. Due to space limitations, we report just the results with a λ , which showed the best accuracy/fairness in the results of Section 5.3. For the sake of completeness, Figure 5, also reports the two-dimensional Kolmogorov–Smirnov (KS) distance [169,170] between the distributions of Males and Females in this low-dimensional space. Figure 5 clearly shows how the application of the fairness regularizer reduces the amount of discrimination: Males and Females are distributed similarly in the space after the application of the fairness regularizer, while before they either were clustered in different sub-spaces (Figure 5a) or presented a higher KS distance (Figure 5a–c). This means that the regularizer reduced the ability to identify Males and Females simple based on their position in the space defined by the representation vector learned by the architecture.

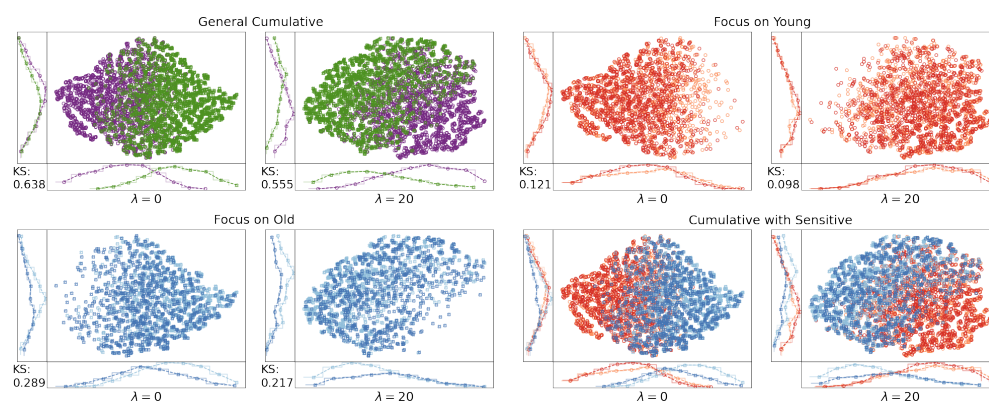


(a) Classical Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint.



(b) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and no HE.

Figure 5. Cont.



(c) Proposed Architecture with ($\lambda > 0$) and without ($\lambda = 0$) the fairness constraint and HE.

Figure 5. Comparison between the different architecture using the t-SNE dimensionality reduction algorithm on the learned embedding.

6. Conclusions

The use of artificial intelligence coupled with the ability to learn from historical data is becoming ubiquitous. For this reason, the social and ethical implications of the widespread use of tools empowered with such intelligence cannot be ignored any longer. The increasing concerns regarding these issues are not only increased by the population or by the institutions, but also by researchers who have shown potential discriminatory behavior, by threats to privacy and to the right of explanation, and by risks of attacks in current artificial intelligence systems. Institutions such as the European Union have created a high-level expert group on this subject drawing guidelines for more trustworthy intelligent systems (<https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai> (accessed on 11 August 2021)).

For this purpose, in this work, we work toward the development of systems able to ensure trustworthiness by delivering privacy, fairness, and explainability by-design. In particular, we have shown that it is possible to simultaneously learn from data while preserving the privacy of the individuals thanks to the use of Homomorphic Encryption, ensure fairness by learning a fair representation from the data, and ensure explainable decisions with local and global explanations without compromising the accuracy of the final models. We then tested the practicality of our approach on a widespread and controversial application, namely the face recognition, using the recent FairFace dataset to prove the validity of our approach.

To the best knowledge of the authors this is one of the first results in this framework with actual practical results. Nevertheless, this work is just a step forward toward the design of fully trustworthy intelligent systems. For example, in the future, more applications could be explored. Moreover, we need to address the requirement of robustness, which demands making the approach more robust to the presence of adversarial attacks (i.e., adversarial samples or poisoning methods). For this aspect, our framework is already designed to encapsulate robustness requirements since adversarial defense mechanisms are mostly based on gradient-based methods, which marry well with our framework. Finally, while a strong theoretical framework has been developed for the different methods employed in this work, a theoretical framework able to simultaneously offer statistical guarantees of privacy, fairness, and explainability still needs to be designed.

Author Contributions: Conceptualization, D.F., L.O. and N.N.; methodology, D.F., L.O. and N.N.; software, D.F.; validation, D.F., L.O. and N.N.; formal analysis, D.F. and L.O.; investigation, D.F., L.O. and N.N.; resources, D.F.; data curation, D.F.; writing—original draft preparation, D.F. and L.O. writing—review and editing, D.F., L.O., N.N., and D.A.; visualization, D.F.; supervision, L.O. and D.A.; project administration, L.O. and D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this work we exploited the FairFace dataset [65]. The code is available at the following link https://github.com/lucaoneto/ENTROPY_2021 (accessed on 11 August 2021).

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

References

1. Winfield, A.F.; Michael, K.; Pitt, J.; Evers, V. Machine ethics: The design and governance of ethical AI and autonomous systems. *Proc. IEEE* **2019**, *107*, 509–517. [CrossRef]
2. Schneider, F.B. *Trust in Cyberspace*; National Academy Press: Washington, DC, USA, 1999.
3. Jiang, R. A trustworthiness evaluation method for software architectures based on the principle of maximum entropy (POME) and the Grey decision-making method (GDMM). *Entropy* **2014**, *16*, 4818–4838. [CrossRef]
4. European Commission. Ethics Guidelines for Trustworthy AI. Available online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 21 June 2021).
5. Borrellas, P.; Unceta, I. The Challenges of Machine Learning and Their Economic Implications. *Entropy* **2021**, *23*, 275. [CrossRef]
6. Resource. How Robots are Reshaping ‘Dirty, Dull and Dangerous’ Recycling Jobs. Available online: <https://resource.co/article/how-robots-are-reshaping-dirty-dull-and-dangerous-recycling-jobs> (accessed on 21 June 2021).
7. Smuha, N.A. The EU approach to ethics guidelines for trustworthy artificial intelligence. *Comput. Law Rev. Int.* **2019**, *20*, 97–106. [CrossRef]
8. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.
9. Merkert, J.; Mueller, M.; Hubl, M. A Survey of the Application of Machine Learning in Decision Support Systems. In *European Conference on Information Systems*; Association for Information Systems: Atlanta, GA, USA, 2015.
10. Hekler, A.; Utikal, J.S.; Enk, A.H.; Solass, W.; Schmitt, M.; Klode, J.; Schadendorf, D.; Sondermann, W.; Franklin, C.; Bestvater, F.; et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur. J. Cancer* **2019**, *118*, 91–96. [CrossRef] [PubMed]
11. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
12. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *J. Artif. Intell. Res.* **2018**, *62*, 729–754. [CrossRef]
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
14. Flores, A.W.; Bechtel, K.; Lowenkamp, C.T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. *Fed. Probat.* **2016**, *80*, 38.
15. Propublica. Machine Bias. Available online: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (accessed on 21 June 2021).
16. Lum, K.N. Limitations of mitigating judicial bias with machine learning. *Nat. Hum. Behav.* **2017**, *1*, 1. [CrossRef]
17. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
18. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, in press. [CrossRef]
19. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv* **2018**, arXiv:1802.07228.
20. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial Attacks on Medical Machine Learning. *Science* **2019**, *363*, 1287–1289. [CrossRef]
21. Comiter, M. *Attacking Artificial Intelligence*. *Belfer Center Paper*; Belfer Center for Science and International Affairs: Cambridge, MA, USA, 2019.
22. Microsoft. Failure Modes in Machine Learning. Available online: <https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning> (accessed on 21 June 2021).
23. Schneier, B. Attacking Machine Learning Systems. *IEEE Ann. Hist. Comput.* **2020**, *53*, 78–80. [CrossRef]
24. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
25. Verma, S.; Rubin, J. Fairness definitions explained. In Proceedings of the IEEE/ACM International Workshop on Software Fairness, Gothenburg, Sweden, 29 May 2018; pp. 1–7.
26. Oneto, L.; Chiappa, S. Fairness in Machine Learning. In *Recent Trends in Learning From Data*; Oneto, L., Navarin, N., Sperduti, N., Anguita, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2020.

27. Corbett-Davies, S.; Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* **2018**, arXiv:1808.00023.
28. Al-Rubaie, M.; Chang, J.M. Privacy-preserving machine learning: Threats and solutions. *IEEE Secur. Priv.* **2019**, *17*, 49–58. [[CrossRef](#)]
29. Biggio, B.; Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.* **2018**, *84*, 317–331. [[CrossRef](#)]
30. Gunning, D. Explainable artificial intelligence (XAI). *Sci. Robot.* **2019**, *4*, eaay7120. [[CrossRef](#)]
31. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [[CrossRef](#)]
32. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, 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](#)]
33. Oneto, L.; Donini, M.; Pontil, M.; Shawe-Taylor, J. Randomized Learning and Generalization of Fair and Private Classifiers: From PAC-Bayes to Stability and Differential Privacy. *Neurocomputing* **2020**, in press. [[CrossRef](#)]
34. Kilbertus, N.; Gascón, A.; Kusner, M.; Veale, M.; Gummadi, K.; Weller, A. Blind justice: Fairness with encrypted sensitive attributes. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
35. Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; Ullman, J. Differentially private fair learning. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
36. Cummings, R.; Gupta, V.; Kimpara, D.; Morgenstern, J. On the compatibility of privacy and fairness. In Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization, Larnaca, Cyprus, 9–12 June 2019.
37. Mozannar, H.; Ohanessian, M.; Srebro, N. Fair learning with private demographic data. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020.
38. Xu, D.; Yuan, S.; Wu, X. Achieving differential privacy and fairness in logistic regression. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019.
39. Bagdasaryan, E.; Poursaeed, O.; Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 15479–15488.
40. Ding, J.; Zhang, X.; Li, X.; Wang, J.; Yu, R.; Pan, M. Differentially private and fair classification via calibrated functional mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
41. Lyu, L.; Li, Y.; Nandakumar, K.; Yu, J.; Ma, X. How to democratise and protect AI: Fair and differentially private decentralised deep learning. *IEEE Trans. Dependable Secur. Comput.* **2020**, in press. [[CrossRef](#)]
42. Adel, T.; Valera, I.; Ghahramani, Z.; Weller, A. One-network adversarial fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
43. Xu, D.; Yuan, S.; Zhang, L.; Wu, X. Fairgan: Fairness-aware generative adversarial networks. In Proceedings of the IEEE International Conference on Big Data, Seattle, WA, USA, 10–13 December 2018.
44. Wadsworth, C.; Vera, F.; Piech, C. Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv* **2018**, arXiv:1807.00199.
45. Franco, D.; Navarin, N.; Donini, M.; Anguita, D.; Oneto, L. Deep Fair Models for Complex Data: Graphs Labeling and Explainable Face Recognition. *Neurocomputing* **2021**, in press. [[CrossRef](#)]
46. Fu, Z.; Xian, Y.; Gao, R.; Zhao, J.; Huang, Q.; Ge, Y.; Xu, S.; Geng, S.; Shah, C.; Zhang, Y.; et al. Fairness-aware explainable recommendation over knowledge graphs. In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi’an, China, 25–30 July 2020.
47. Schumann, C.; Foster, J.; Mattei, N.; Dickerson, J. We need fairness and explainability in algorithmic hiring. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, Auckland, New Zealand, 9–13 May 2020.
48. Fidel, G.; Bitton, R.; Shabtai, A. When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020.
49. Goldberg, S.; Pinsky, E.; Galitsky, B. A bi-directional adversarial explainability for decision support. *Hum.-Intell. Syst. Integr.* **2021**, *3*, 1–14. [[CrossRef](#)]
50. Huang, C.; Kairouz, P.; Chen, X.; Sankar, L.; Rajagopal, R. Context-aware generative adversarial privacy. *Entropy* **2017**, *19*, 656. [[CrossRef](#)]
51. Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; Jana, S. Certified robustness to adversarial examples with differential privacy. In Proceedings of the Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2019.
52. Nasr, M.; Shokri, R.; Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018.
53. Wu, Z.; Wang, Z.; Wang, Z.; Jin, H. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018.
54. Meden, B.; Emeršič, Ž.; Štruc, V.; Peer, P. k-Same-Net: K-Anonymity with generative deep neural networks for face deidentification. *Entropy* **2018**, *20*, 60. [[CrossRef](#)] [[PubMed](#)]
55. Fitzsimons, J.; Al Ali, A.; Osborne, M.; Roberts, S. A general framework for fair regression. *Entropy* **2019**, *21*, 741. [[CrossRef](#)]

56. Cooley, E.; Hester, N.; Cipolli, W.; Rivera, L.I.; Abrams, K.; Pagan, J.; Sommers, S.R.; Payne, K. Racial biases in officers' decisions to frisk are amplified for Black people stopped among groups leading to similar biases in searches, arrests, and use of force. *Soc. Psychol. Personal. Sci.* **2020**, *11*, 761–769. [[CrossRef](#)]
57. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [[CrossRef](#)]
58. Cox, L.A. Information Structures for Causally Explainable Decisions. *Entropy* **2021**, *23*, 601. [[CrossRef](#)] [[PubMed](#)]
59. Käding, C.; Rodner, E.; Freytag, A.; Denzler, J. Fine-tuning deep neural networks in continuous learning scenarios. In Proceedings of the Asian Conference on Computer Vision (ACCV 2016), Taipei, Taiwan, 20–24 November 2016.
60. Peters, M.E.; Ruder, S.; Smith, N.A. To tune or not to tune? Adapting pretrained representations to diverse tasks. *arXiv* **2019**, arXiv:1903.05987.
61. Rivest, R.L.; Adleman, L.; Dertouzos, M.L. On data banks and privacy homomorphisms. *Found. Secur. Comput.* **1978**, *4*, 169–180.
62. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning fair representations. In Proceedings of the International Conference on Machine Learning (ICML 2013), Atlanta, GA, USA, 16–21 June 2013.
63. McNamara, D.; Ong, C.S.; Williamson, B. Costs and Benefits of Fair Representation Learning. In Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, Honolulu, HI, USA, 27–28 January 2019.
64. Oneto, L.; Donini, M.; Luise, G.; Ciliberto, C.; Maurer, A.; Pontil, M. Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020.
65. Kärkkäinen, K.; Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv* **2019**, arXiv:1908.04913.
66. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
67. Hutchinson, M.L.; Antono, E.; Gibbons, B.M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming data scarcity with transfer learning. *arXiv* **2017**, arXiv:1711.05099.
68. Wu, Y.; Ji, Q. Constrained deep transfer feature learning and its applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5101–5109.
69. Gentry, C. Computing arbitrary functions of encrypted data. *Commun. ACM* **2010**, *53*, 97–105. [[CrossRef](#)]
70. Acar, A.; Aksu, H.; Uluagac, A.S.; Conti, M. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.* **2018**, *51*, 1–35. [[CrossRef](#)]
71. Naehrig, M.; Lauter, K.; Vaikuntanathan, V. Can homomorphic encryption be practical? In Proceedings of the ACM Workshop on Cloud Computing Security Workshop, Chicago, IL, USA, 21 October 2011.
72. Gilad-Bachrach, R.; Dowlin, N.; Laine, K.; Lauter, K.; Naehrig, M.; Wernsing, J. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
73. Pulido-Gaytan, B.; Tchernykh, A.; Cortés-Mendoza, J.M.; Babenko, M.; Radchenko, G.; Avetisyan, A.; Drozdov, A.Y. Privacy-preserving neural networks with Homomorphic encryption: C challenges and opportunities. *Peer-Netw. Appl.* **2021**, *14*, 1666–1691. [[CrossRef](#)]
74. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Rambachan, A. Algorithmic Fairness. *AEA Pap. Proc.* **2018**, *108*, 22–27. [[CrossRef](#)]
75. McNamara, D.; Ong, C.S.; Williamson, R.C. Provably fair representations. *arXiv* **2017**, arXiv:1710.04394.
76. Obla, S.; Gong, X.; Aloufi, A.; Hu, P.; Takabi, D. Effective activation functions for homomorphic evaluation of deep neural networks. *IEEE Access* **2020**, *8*, 153098–153112. [[CrossRef](#)]
77. Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; Pontil, M. Empirical risk minimization under fairness constraints. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018.
78. Calders, T.; Kamiran, F.; Pechenizkiy, M. Building classifiers with independency constraints. In Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, Miami, FL, USA, 6 December 2009.
79. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 27–29 October 2017.
80. Hinton, G.; Roweis, S.T. Stochastic neighbor embedding. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002.
81. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
82. Chouldechova, A.; Roth, A. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* **2020**, *63*, 82–89. [[CrossRef](#)]
83. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
84. Lahoti, P.; Gummadi, K.P.; Weikum, G. iFair: Learning individually fair data representations for algorithmic decision making. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 8–11 April 2019.
85. Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; Ver Steeg, G. Invariant representations without adversarial training. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018.
86. Xie, Q.; Dai, Z.; Du, Y.; Hovy, E.; Neubig, G. Controllable invariance through adversarial feature learning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

87. 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]
88. De Cristofaro, E. An overview of privacy in machine learning. *arXiv* **2020**, arXiv:2005.08679.
89. Sweeney, L. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
90. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkatasubramanian, M. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data (TKDD)* **2007**, *1*, 3-es. [CrossRef]
91. Li, N.; Li, T.; Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007.
92. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
93. Thorgeirsson, A.T.; Gauterin, F. Probabilistic Predictions with Federated Learning. *Entropy* **2021**, *23*, 41. [CrossRef]
94. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
95. Asad, M.; Moustafa, A.; Ito, T. FedOpt: Towards communication efficiency and privacy preservation in federated learning. *Appl. Sci.* **2020**, *10*, 2864. [CrossRef]
96. Hao, M.; Li, H.; Xu, G.; Liu, S.; Yang, H. Towards efficient and privacy-preserving federated deep learning. In Proceedings of the ICC 2019—2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019.
97. Fang, H.; Qian, Q. Privacy Preserving Machine Learning with Homomorphic Encryption and Federated Learning. *Future Internet* **2021**, *13*, 94. [CrossRef]
98. Cai, Y.; Tang, C.; Xu, Q. Two-Party Privacy-Preserving Set Intersection with FHE. *Entropy* **2020**, *22*, 1339. [CrossRef]
99. Rivest, R.L.; Shamir, A.; Adleman, L. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **1978**, *21*, 120–126. [CrossRef]
100. Paillier, P. Public-key cryptosystems based on composite degree residuosity. classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 1999.
101. Fan, J.; Vercauteren, F. Somewhat practical fully homomorphic encryption. *IACR Cryptol. EPrint Arch.* **2012**, *2012*, 144.
102. Cheon, J.H.; Kim, A.; Kim, M.; Song, Y. Homomorphic encryption for arithmetic of approximate numbers. In *International Conference on the Theory and Application of Cryptology and Information Security*; Springer: Cham, Switzerland, 2017; pp. 409–437.
103. Gentry, C. *A Fully Homomorphic Encryption Scheme*; Stanford University: Stanford, CA, USA, 2009.
104. Li, J.; Li, Y.; Xiang, X.; Xia, S.; Dong, S.; Cai, Y. TNT: An Interpretable Tree-Network-Tree Learning Framework Using Knowledge Distillation. *Entropy* **2020**, *22*, 1203. [CrossRef]
105. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
106. Barocas, S.; Selbst, A.D. Big data’s disparate impact. *Calif. Law Rev.* **2016**, *104*, 671. [CrossRef]
107. Valiant, L.G. A theory of the learnable. *Commun. ACM* **1984**, *27*, 1134–1142. [CrossRef]
108. Zhang, J.; Zhang, Z.; Xiao, X.; Yang, Y.; Winslett, M. Functional mechanism: Regression analysis under differential privacy. *arXiv* **2012**, arXiv:1208.0219.
109. Kairouz, P.; Oh, S.; Viswanath, P. Extremal Mechanisms for Local Differential Privacy. *J. Mach. Learn. Res.* **2016**, *17*, 1–51.
110. 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 Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016.
111. Choraś, M.; Pawlicki, M.; Puchalski, D.; Kozik, R. Machine learning—The results are not the only thing that matters! what about security, explainability and fairness? In Proceedings of the International Conference on Computational Science, Amsterdam, The Netherlands, 3–5 June 2020.
112. Vellido, A. Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis.* **2019**, *5*, 11–17. [CrossRef]
113. Bonnefon, J.; Černý, D.; Danaher, J.; Devillier, N.; Johansson, V.; Kovacicova, T.; Martens, M.; Mladenovic, M.; Palade, P.; Reed, N. *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*; EU Commission: Brussels, Belgium, 2020.
114. New York Times. A Case for Banning Facial Recognition. Available online: <https://www.nytimes.com/2020/06/09/technology/facial-recognition-software.html> (accessed on 21 June 2021).
115. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018.
116. Raji, I.D.; Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA, 27–28 January 2019.
117. The Verge. A Black Man Was Wrongfully Arrested Because of Facial Recognition. Available online: <https://www.theverge.com/2020/6/24/21301759/facial-recognition-detroit-police-wrongful-arrest-robert-williams-artificial-intelligence> (accessed on 21 June 2021).

118. CNN. Portland Passes Broadest Facial Recognition Ban in the US. Available online: <https://edition.cnn.com/2020/09/09/tech/portland-facial-recognition-ban/index.html> (accessed on 21 June 2021).
119. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8–14 December 2001.
120. Liu, C.; Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **2002**, *11*, 467–476.
121. Ahonen, T.; Hadid, A.; Pietikainen, M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [[CrossRef](#)]
122. Wang, M.; Deng, W. Deep face recognition: A survey. *Neurocomputing* **2021**, *429*, 215–244. [[CrossRef](#)]
123. Jayaraman, U.; Gupta, P.; Gupta, S.; Arora, G.; Tiwari, K. Recent development in face recognition. *Neurocomputing* **2020**, *408*, 231–245. [[CrossRef](#)]
124. Dhillon, A.; Verma, G.K. Convolutional neural network: A review of models, methodologies and applications to object detection. *Prog. Artif. Intell.* **2020**, *9*, 85–112. [[CrossRef](#)]
125. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [[CrossRef](#)]
126. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
127. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
128. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
129. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
130. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
131. Masi, I.; Wu, Y.; Hassner, T.; Natarajan, P. Deep Face Recognition: A Survey. In Proceedings of the Conference on Graphics, Patterns and Images, Paraná, Brazil, 29 October–1 November 2018.
132. Guo, G.; Zhang, N. A survey on deep learning based face recognition. *Comput. Vis. Image Underst.* **2019**, *189*, 102805. [[CrossRef](#)]
133. Du, H.; Shi, H.; Zeng, D.; Mei, T. The Elements of End-to-end Deep Face Recognition: A Survey of Recent Advances. *arXiv* **2020**, arXiv:2009.13290.
134. Dwork, C.; Immorlica, N.; Kalai, A.T.; Leiserson, M.D.M. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018.
135. Oneto, L.; Donini, M.; Elders, A.; Pontil, M. Taking Advantage of Multitask Learning for Fair Classification. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019.
136. Edwards, H.; Storkey, A.J. Censoring Representations with an Adversary. In Proceedings of the International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016.
137. Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; Zemel, R.S. The Variational Fair Autoencoder. In Proceedings of the International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016.
138. Madras, D.; Creager, E.; Pitassi, T.; Zemel, R. Learning Adversarially Fair and Transferable Representations. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
139. Johansson, F.; Shalit, U.; Sontag, D. Learning representations for counterfactual inference. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
140. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [[CrossRef](#)]
141. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
142. Oneto, L. *Model Selection and Error Estimation in a Nutshell*; Springer: Berlin/Heidelberg, Germany, 2020.
143. Lyubashevsky, V.; Peikert, C.; Regev, O. On ideal lattices and learning with errors over rings. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 2010.
144. Rosasco, L.; De Vito, E.; Caponnetto, A.; Piana, M.; Verri, A. Are loss functions all the same? *Neural Comput.* **2004**, *16*, 1063–1076. [[CrossRef](#)]
145. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
146. Orlandi, C.; Piva, A.; Barni, M. Oblivious neural network computing via homomorphic encryption. *EURASIP J. Inf. Secur.* **2007**, *2007*, 1–11. [[CrossRef](#)]
147. Benaissa, A.; Retiat, B.; Cebere, B.; Belfedhal, A.E. TenSEAL: A Library for Encrypted Tensor Operations Using Homomorphic Encryption. *arXiv* **2021**, arXiv:2104.03152.
148. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
149. Microsoft Research. Microsoft SEAL. 2020. Available online: <https://github.com/Microsoft/SEAL> (accessed on 11 August 2021).

150. Halevi, S.; Shoup, V. Algorithms in helib. In Proceedings of the Annual Cryptology Conference, Santa Barbara, CA, USA, 17–21 August 2014.
151. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference (BMVC 2015), Swansea, UK, 7–10 September 2015.
152. Ivanov, V.V. *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*; Springer: Berlin/Heidelberg, Germany, 1976.
153. Oneto, L.; Ridella, S.; Anguita, D. Tikhonov, Ivanov and Morozov regularization for support vector machine learning. *Mach. Learn.* **2015**, *103*, 103–136. [[CrossRef](#)]
154. Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; Smola, A. A kernel method for the two-sample-problem. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006.
155. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013.
156. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020.
157. Desai, S.; Ramaswamy, H.G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2020), Snowmass Village, CO, USA, 1–5 March 2020.
158. Rebuffi, S.; Fong, R.; Ji, X.; Vedaldi, A. There and Back Again: Revisiting Backpropagation Saliency Methods. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA 14–19 June 2020.
159. Taha, A.; Yang, X.; Shrivastava, A.; Davis, L. A Generic Visualization Approach for Convolutional Neural Networks. In Proceedings of the IEEE European Conference on Computer Vision (ECCV 2020), Online, 23–28 August 2020.
160. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv* **2020**, arXiv:2003.07631.
161. Sattarzadeh, S.; Sudhakar, M.; Lem, A.; Mehryar, S.; Plataniotis, K.; Jang, J.; Kim, H.; Jeong, Y.; Lee, S.; Bae, K. Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation. *arXiv* **2020**, arXiv:2010.00672.
162. Lee, S.; Lee, H.; Abbeel, P.; Ng, A.Y. *Efficient L_1 Regularized Logistic Regression*; AAAI: Menlo Park, CA, USA, 2006.
163. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **1996**, *58*, 267–288. [[CrossRef](#)]
164. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [[CrossRef](#)]
165. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L. YFCC100M: The new data in multimedia research. *Commun. ACM* **2016**, *59*, 64–73. [[CrossRef](#)]
166. Zeiler, M.D. Adadelta: An adaptive learning rate method. *arXiv* **2012**, arXiv:1212.5701.
167. Merler, M.; Ratha, N.; Feris, R.S.; Smith, J.R. Diversity in faces. *arXiv* **2019**, arXiv:1901.10436.
168. Van Loan, C.F.; Golub, G.H. *Matrix Computations*; Johns Hopkins University Press: Baltimore, MD, USA, 1983.
169. Peacock, J.A. Two-dimensional goodness-of-fit testing in astronomy. *Mon. Not. R. Astron. Soc.* **1983**, *202*, 615–627. [[CrossRef](#)]
170. Fasano, G.; Franceschini, A. A multidimensional version of the Kolmogorov–Smirnov test. *Mon. Not. R. Astron. Soc.* **1987**, *225*, 155–170. [[CrossRef](#)]