applied sciences

MDPI

Review

Privacy-Preserving and Explainable AI in Industrial Applications

Iulian Ogrezeanu *, Anamaria Vizitiu, Costin Ciușdel, Andrei Puiu, Simona Coman, Cristian Boldișor, Alina Itu, Robert Demeter, Florin Moldoveanu, Constantin Suciu and Lucian Itu

Automation and Information Technology, "Transilvania" University of Brașov, 500036 Brașov, Romania; anamaria.vizitiu@unitbv.ro (A.V.); costin.ciusdel@unitbv.ro (C.C.); andrei.puiu@unitbv.ro (A.P.); simona.coman@unitbv.ro (S.C.); cristian.boldisor@unitbv.ro (C.B.); alina.itu@unitbv.ro (A.I.); rdemeter@unitbv.ro (R.D.); moldof@unitbv.ro (F.M.); suciuc@unitbv.ro (C.S.); lucian.itu@unitbv.ro (L.I.)
* Correspondence: iulian.ogrezeanu@unitbv.ro

Abstract: The industrial environment has gone through the fourth revolution, also called "Industry 4.0", where the main aspect is digitalization. Each device employed in an industrial process is connected to a network called the industrial Internet of things (IIOT). With IIOT manufacturers being capable of tracking every device, it has become easier to prevent or quickly solve failures. Specifically, the large amount of available data has allowed the use of artificial intelligence (AI) algorithms to improve industrial applications in many ways (e.g., failure detection, process optimization, and abnormality detection). Although data are abundant, their access has raised problems due to privacy concerns of manufacturers. Censoring sensitive information is not a desired approach because it negatively impacts the AI performance. To increase trust, there is also the need to understand how AI algorithms make choices, i.e., to no longer regard them as black boxes. This paper focuses on recent advancements related to the challenges mentioned above, discusses the industrial impact of proposed solutions, and identifies challenges for future research. It also presents examples related to privacy-preserving and explainable AI solutions, and comments on the interaction between the identified challenges in the conclusions.

Keywords: artificial intelligence; industrial applications; privacy preservation; explainability; bias; fairness

check for updates

Citation: Ogrezeanu, I.; Vizitiu, A.; Ciușdel, C.; Puiu, A.; Coman, S.; Boldișor, C.; Itu, A.; Demeter, R.; Moldoveanu, F.; Suciu, C.; et al. Privacy-Preserving and Explainable AI in Industrial Applications. *Appl. Sci.* **2022**, *12*, 6395. https://doi.org/10.3390/app12136395

Academic Editor: Federico Divina

Received: 27 May 2022
Accepted: 22 June 2022
Published: 23 June 2022

1. Introduction

Industry 4.0 [1] has introduced advanced technology in manufacturing, to make it more client-driven and customizable, leading to manufacturers striving toward a continuous improvement in quality and productivity. To achieve smart manufacturing, which enables variable product demand, intelligent systems were introduced in industrial units.

Recent developments in Internet of things (IOT) [2], Cyber-Physical Production Systems (CPPS) [3], and big data [4] led to major improvements in productivity, quality, and monitoring of industrial processes. Artificial intelligence (AI) plays an important role in industry, as more and more manufacturers are implementing AI in their processes.

Developed and employed with the purpose of performing tasks that normally require human discernment, AI is currently a popular topic. Having the capability of interpreting data for solving complex problems [5], AI is also a good fit for factories [6]. It enables industrial systems to process data, perceive their environment, and learn, while building up experience, in order to become better at a task by dealing with it and its data repeatedly.

Artificial intelligence [7] is a subject that researchers have been preoccupied with almost since computers were invented. AI includes every algorithm that enables machines to perform tasks that require discernment, not just by applying a formula or following a strict rule-based logic. Thus, if we provide datasets with inputs and outputs to an AI algorithm, it will be capable of yielding a logic which maps the inputs to the outputs. In contrast, in classic programming, humans provide the logic. Of course, in many situations it is not necessary to use AI (e.g., if the problem can be solved through a mathematical formula). In the last two decades, thanks to the increase in computational power, AI has become very popular, and it has been used in several domains (medicine, marketing, industry, etc.), with various subdomains of algorithms such as machine learning (ML) [8] and deep learning (DL).

Machine learning is a popular subdomain of AI, and it is composed of statistical algorithms that can learn from data to create mathematical models for intelligent systems. Today, we have recommendation systems that use ML to suggest aspects that we like on the basis of our preferences (e.g., music, ads, and shopping). In the medical environment, we have ML models which help clinicians during diagnostic processes (decision support systems) [9].

Deep learning [10] is a widely used type of learning algorithm that relies on defining neural networks [11] with more than one hidden layer of neurons. Neural networks are inspired by the human brain, having computational units (named neurons) which are interconnected and exchange information to extract features from input data, thus enabling the mapping of input data to output data. However, neural networks used in AI do not work in the same way as human neural networks, because they exchange information using real numbers, whilst our neurons exchange information through electrical impulses. Neurons are organized in layers, where the first and the last layer are those that interact with the external environment, also named the input layer and output layer. Intermediate layers are called hidden layers.

Multiple review papers have described the multitude of approaches on the basis of which AI is employed in manufacturing. In [12], Sharma et al. presented a theoretical framework for machine learning in manufacturing, which guides researchers in elaborating a paper in this field. They pointed to several review papers that targeted the use of ML in the industrial environment. Rai et al. [13] discussed the use of AI in the context of the fourth industrial revolution. To highlight the potential advantages and potential flaws of using AI in industry, Bertolini et al. [14] reviewed the literature and classified research on the basis of the algorithm and application domain. Sarker [15] also reviewed the use of machine learning in real-world applications such as cyber security, agriculture, smart cities, and healthcare. In [16], Rao summarized the use of AI in different domains such as healthcare and travel.

In [17], four important challenges were identified: data availability, data quality, cybersecurity and privacy preservation, and interpretability/explainability. While the former two have been extensively discussed in the past and are well known, in this paper, we focus on topics related to the latter two challenges.

AI/ML relies extensively on existing and future data to deliver accurate and reliable results. The collection of large volumes of data for centralized processing poses severe privacy concerns. Thus, the first challenge refers to the fact that, while industrial data are abundant, they are hard to circulate and access due to privacy/IP constraints, also affecting the development of computer-based solutions. Industrial AI systems are difficult to realize, as data to develop and train them exist, but are not accessible. If training datasets lack diversity, algorithms may be biased or skewed to certain types of data/events [18].

Secondly, AI algorithms should be explainable and interpretable. ML algorithms are, in general, related to the concept of 'black box', i.e., the rationale for how the outputs are inferred from the input data is unclear [19]. Algorithmic decisions should, however, ideally provide a form of explainability [20]. In general, explanations are about the attribution of the worth of input features toward the final model predictions, whereas interpretability refers to the deterministic propagation of information from the input to the response function.

ML is usually regarded as a 'black box' unit; once a model is trained, its logic for determining the outputs on the basis of the inputs is not available, and further experiments and methods need to be performed to understand the way a trained model analyzes and processes the data. For stakeholders, however, it is important to understand how and why a solution is being proposed. Hence, explainable AI, with its interpretability tools, is key.

Model-agnostic methods [21] were the subject of past research that yielded good results. Most of them targeted local interpretable model-agnostic explanations (LIMEs) [22] and Shapley additive explanations [23]. An important advantage of these methods is that they are compatible with a multitude of ML models. On the other hand, there are model-specific interpretation methods [24], which have the disadvantage of being compatible only with specific model types.

This paper highlights the recent developments related to privacy preservation and explainability in industrial AI applications and discusses the potential impact of existing solutions in the industrial domain. Several examples are presented, related to explainable AI methods and privacy preservation techniques. Section 2 addresses aspects related to privacy preservation in industrial AI applications, while the explainability and interpretability requirements of an AI model are discussed in Section 3. In the context of the approaches described herein, Section 4 focuses on the impact of AI in industry and identifies remaining challenges. Final conclusions are drawn in Section 5. Given the focus on the two challenges, the paper should be regarded as an argumentative review; the literature is examined selectively in order to support the arguments of the necessity of both explainability and privacy preservation in industrial AI applications. Furthermore, new challenges are identified, and their interaction is discussed.

2. Privacy Preservation in Industrial AI Applications

2.1. State of the Art in Privacy-Preserving AI

In this section, we briefly present various approaches for performing privacy-preserving AI.

One of the most used solutions in privacy-preserving AI is homomorphic encryption (HE). HE allows users to perform computations on encrypted data, yielding results that are also encrypted (results are identical to those obtained by performing the operations on unencrypted data). This type of encryption is necessary when processing sensitive data (e.g., healthcare data). Homomorphic encryption has been introduced and developed independently from AI, but the large computational overhead limits its real-world usage. Since AI-based methods provide results in near real time, i.e., the computational cost during inference is small, extending AI with HE allows for privacy-preserving data processing, while obtaining results in a reasonable amount of time.

One of the first notable approaches in using homomorphic encryption with neural networks was proposed by Orlandi et al. [25]. They developed an approach to process encrypted data using a neural network, ensuring that not only the data are protected, but also the neural network itself (weight values and hyperparameters). Figure 1 illustrates the data exchange between server and client, in an encrypted format.

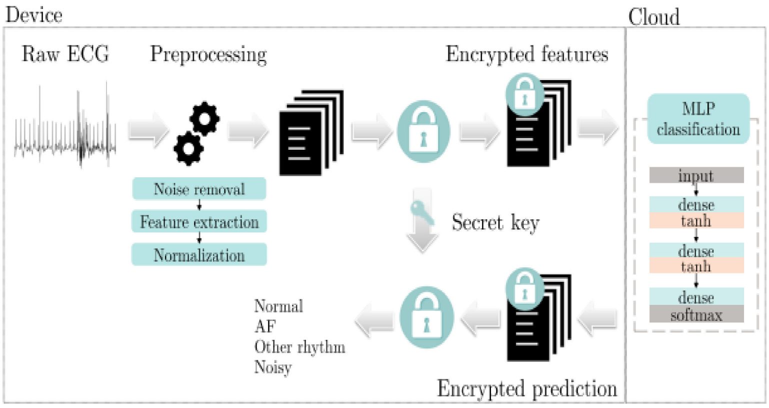


Figure 1. Workflow of privacy-preserving feature-based ECG classification method [26].

HE comes in many forms, and one of them is fully homomorphic encryption (FHE), which refers to a cryptosystem able to support arbitrary computation on ciphertexts. Sun et al. [27] used FHE to implement a private decision tree classification of user data. Aslett et al. [28]

performed a review of homomorphic encryption techniques successfully applied in machine learning, and they also documented an R package implementing a homomorphic scheme. Deep neural networks were also employed in studies with homomorphic encryption. For example, Takabi et al. [29] used HE for multiparty machine learning; multiple parties participated in training the deep neural network, while maintaining data privacy.

A novel homomorphic encryption framework was proposed by Li et al. [30] to protect the data and the expertise of the algorithm when using cloud computing for model training. For healthcare and bioinformatics applications, Wood et al. [31] reviewed the use of FHE together with machine learning models.

The main disadvantage of FHE is its large computational cost. To address this aspect, partially homomorphic encryption (PHE) schemes were proposed and used by Fang et al. [32] to transmit encrypted gradients from all learning parties, thus speeding up the training by 25–28%, while maintaining the same level of accuracy in comparison with a classic approach.

Distributing the training process to multiple servers or decentralized edge devices with local data is a preferred approach to address privacy and scalability issues. This type of training is known as federated learning [33] and enables the collaboration of industrial nodes for training a model without exchanging sensitive data. However, reverse engineering may still be employed to extract from the model sensitive information regarding the datasets [34]. Hence, further research is needed for addressing privacy preservation.

Cloud-based implementations that can be used to run homomorphic encryption frameworks are available. One of them is Google's Cloud Platform [35], which runs on the same infrastructure as Google Search, YouTube, and Google Drive. Another widely used ML service is made available by Microsoft. Microsoft Azure Machine Learning [36] provides ML services which can also help in deploying models and managing them efficiently.

2.2. Review of Privacy-Preserving AI in Industrial Applications

In this section, we focus on research that targeted privacy-preserving artificial intelligence applied in industrial applications. Some of the main subjects in industry-oriented research are industrial Internet of things (IIOT) and Industry 4.0. A new method termed verifiable federated learning (VFL) was proposed by Fu et al. [37] for privacy preservation in industrial IOT, which employs federated learning, while also allowing for information extraction from the shared gradients. Figure 2 illustrates the proposed federated learning framework.

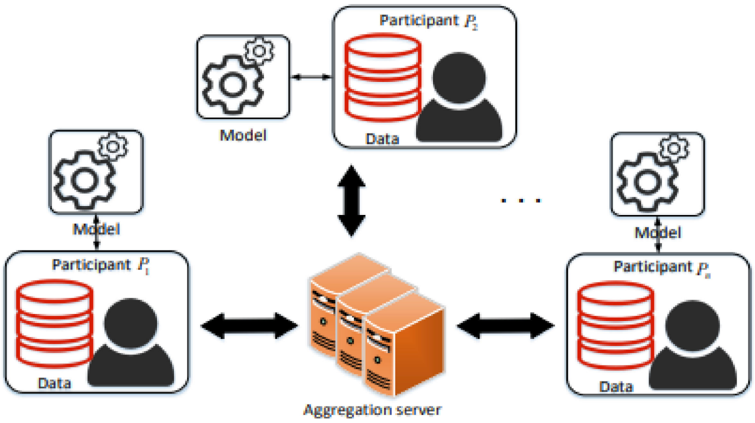


Figure 2. Federated learning framework proposed by Fu et al.

As mentioned above, a promising solution for privacy preservation is homomorphic encryption, but there are also other techniques for data encryption and/or anonymization. For example, on the basis of homomorphic data space transformation, Girka et al. [38] proposed an anonymization algorithm to protect data, while still allowing neural network training. They analyzed the effects that this method has on the neural network performance. By adding new frozen layers to the neural network, they succeeded in achieving anonymization, while the performance was slightly lower when compared to that of the original model.

Blockchain can also be used instead of simple federated learning. Zhao et al. [39] employed blockchain to transfer models trained by customers, thus eliminating the need for federated learning for gradient updates. Because blockchain records are not altered, malicious manufacturers or customer activities are traceable. Another study highlighted the need for privacy preservation to ensure data protection when exchanging information between multiple owners of renewable energy power plants. The main goal of the data exchange is to increase the forecast performance. Gonçalves et al. [40] proposed a privacy-preserving framework that combines the alternating direction method of multipliers with data transformation techniques. Their method proved to be successful, being robust to privacy breaches and communication failures, while the forecast performance was only marginally lower than that obtained using a model without privacy protection.

Generative adversarial network (GAN) is a machine learning algorithm that is widely used to generate synthetic data. Being first proposed in 2014 by Goodfellow et al. [41], GAN is currently popular and comes in different forms, one of them being least square generative adversarial network (LSGAN), which was used by Li et al. [42] together with federated learning to generate renewable scenarios. Through federated learning, a model was trained by gathering knowledge from different renewable sites, and then LSGAN was employed to generate renewable scenarios from the same distribution as the historical data, thanks to the capability of capturing the spatiotemporal characteristics of renewable powers.

Below, we provide a concrete example of a privacy-preserving AI application for casting [43]: a manufacturing process in which a liquid material is usually poured into a mold, which contains a hollow cavity of the desired shape, and then allowed to solidify. Defects may appear during the casting process, e.g., blow holes, pinholes, burr, shrinkage defects, mold material defects, pouring metal defects, and metallurgical defects, which have to be detected, and the corresponding parts have to be removed. Typically, this process is performed by a human operator, who may not be 100% accurate and consistent in their decisions. A fully automated, AI-based approach may reach 100% accuracy and remove inter- and intra-user variability, i.e., improve the robustness of the detection. The manufacturer would typically decide to externalize the development of the AI model, which means that a large dataset containing photos of both acceptable and nonacceptable parts would have to be shared with the entity developing the AI model. However, the manufacturer may not feel comfortable with externalizing photos of nonacceptable parts. In a privacy-preserving setting, the photos would first be homomorphically encrypted or obfuscated, such that the external party cannot reconstruct the original images. The AI model would be trained on the encrypted or obfuscated images, and the trained model would be deployed as an AI service. During inference, the same encryption or obfuscation method would be employed to ensure that the AI model is not fed with out-of-distribution data. A possible technical solution was recently published for a healthcare application [44], which could be similarly applied in the industrial domain for the casting application. Therein, an image obfuscation algorithm was proposed that combines a variational autoencoder with random non-bijective pixel intensity mapping to protect the content of medical images, which are subsequently employed in the development of DL-based solutions. Although a drop in accuracy could be observed when the classifier was trained on obfuscated images, the performance was deemed satisfactory in the context of a privacy–accuracy tradeoff.

3. Explainable Industrial AI Applications

3.1. The Black-Box Aspect of AI

Artificial intelligence algorithms are, in general, regarded as black-box algorithms, i.e., it is not possible to determine or infer why the model has generated a certain output. A representative example was described in [45]; a robotics graduate student tried in 1991 to train a military vehicle to self-drive. The training of the system was accomplished by him manually driving the vehicle while the system (the algorithm was a neural network) memorized the moves for different situations. After a few training sessions, the approach seemed to be working well; however, when the vehicle reached a bridge, it did not know how to handle the situation, and the model would have crashed the vehicle if the user had not intervened. Further testing revealed that the model was relying on grassy roadsides to be guided along the road; hence, the appearance of the bridge caused confusion.

The black-box problem [46] has represented a concern since the very beginnings of neural network research. Currently, very complex neural network architectures are employed, which deepen the black-box problem. The advancements of the technology and of computing power have also further increased its importance. It has become obvious both as a developer and as a user that, to trust an algorithm for making important decisions, one needs to make sure that the algorithm relies on the right properties and reasons.

3.2. State of the Art in Explainable AI

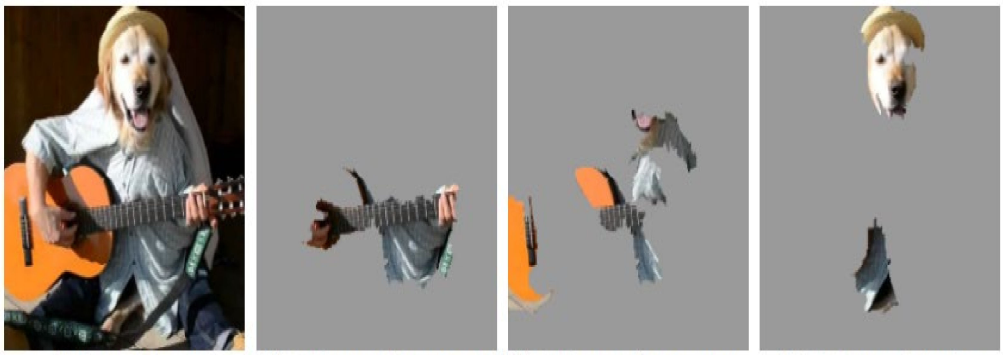
To make it easier for a user to understand the logic behind the decision taken by an algorithm, a user interface (UI) should accompany it. Using UI, automation bias can be mitigated. We can categorize explainable AI methods into the following:

- saliency maps: elements in the input that have the largest influence in the prediction are identified (e.g., LIME);
- feature attribution: attributing the classification to a small number of numeric/semantic features [47,48];
- metric learning [49]: mapping out data structures by deriving a metric from a classifier (explicit Siamese networks are very popular);
- activation maximization: methods that are based on GAN.

There are several methods of explainable AI that can be applied, depending on the data type. These are discussed below.

Even though there are several explainable AI techniques, only certain methods can be applied on tabular data [50]. Techniques designed for images or text data are typically not applicable to tabular data. Tabular data can present characteristics such as correlations between features or temporal aspects, and they may contain categorical features along with continuous features. Poulin et al. [51] proposed one of the first methods of explainable AI, named ExplainD. It measures the importance of each input feature related to the prediction of a classifier. LIME proposed by Ribeiro et al. [52] is a model-agnostic technique used to explain the predictions of the classifier. Many other methods have been developed on the basis of LIME. The core principle is that one or more models are trained to approximate the predictions of the classification model, to determine why the classification model outputted a certain prediction. These models are trained with a dataset containing perturbed data points, which are close to the instance of interest. The newly trained model (named the surrogate model) is used to compute the proximity of each sample instance to the instance of interest. In Figure 3, the explanations for three predictions can be observed.

For time series, saliency maps can be extracted to highlight the importance of a sequence from the input, which is related to the prediction. Class activation mapping (CAM) is a method used on convolutional neural networks (CNN) to identify input features which are representative for a class. Oviedo et al. [53] applied this method to explain model decisions when classifying small X-ray diffraction. Thrun et al. [54] proposed a new method of explainable artificial intelligence (XAI) in which a data-driven approach is used to exploit distance-based data structures, without the need for making any assumption related to the data. LIME can also be employed, as well as any method that yields a heatmap, e.g., deep learning important features (DeepLIFT). Layer-wise relevance propagation (LRP) [55] can also be used, which is a method that computes feature importance by backpropagating a relevance score through the model. Figure 4 contains an example of how features are highlighted to justify model predictions.



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 3. Explaining Google's Inception neural network predictions: electric guitar (**b**), acoustic guitar (**c**) and Labrador (**d**). Highlighted parts from the (**a**) original image are those that contributed the most to each prediction. Adapted with permission from [52], 2016.

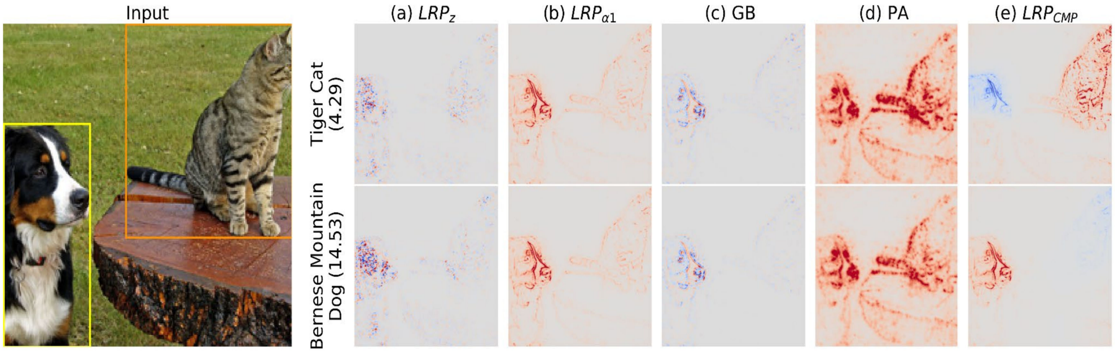


Figure 4. Heatmaps yielded by different techniques for two animals identified in a picture. Adapted with permission from [55], 2020.

It is widely known that XAI presents a wide variety of methods to be used for imaging data. All methods described in the categories above can be applied to gain insight into the logic used by the model to perform the prediction. Yang et al. [56] discussed the use of XAI for medical images, and Xu et al. [57] described approaches and future challenges after summarizing the history of XAI.

3.3. Review of Explainable Industrial AI Applications

Manufacturers are now more interested in the use of AI to improve the overall quality of industrial applications, while at the same time unboxing the original black-box models. One of the domains in which AI is not widely used is air traffic management (ATM); however, introducing AI is a need which can help ATM in the future. As they are increasing in complexity, there is also a need for explainable AI techniques, to identify the most important features in model predictions [58]. Gade et al. [59] introduced a tutorial in which they presented an overview of model explainability and interpretability, alongside techniques which are helpful in providing explainability for AI systems. Their examples were mainly applications from the industrial environment. Longo et al. [60] and Ahmed et al. [61] also reviewed XAI applied in industry. In the automotive field, explainable AI is used to create transparency and understand model decisions [62], but it turns out that XAI itself is not sufficient to increase the trust [63]. Other explanations need to be provided, not only for developers, but also for the end-user.

Krishnamurthy et al. [64] proposed a new XAI framework for predicting maintenance applications for automotive applications and others. Brito et al. [65] used XAI for diagnosis and fault detection in rotating machinery. Anomaly detection was used for performing, while, for interpreting fault diagnosis models, they employed Shapley additive explanations (SHAPs). To reduce energy consumption in mineral processing, Chelgani et al. [66] highlighted the importance of using high-pressure grinding rolls (HPGRs). The main problem is HPGR modeling; they proposed to expand the existing conscious laboratory (CL) and used XAI systems to innovate powder technology industries. SHAP and extreme gradient boosting (XGBoost) were the selected methods for achieving model explainability.

4. Industrial Impact and Remaining Challenges

4.1. Industrial Impact

The use of artificial intelligence in industrial applications, and the use of techniques for privacy preservation and model explainability can impact industries in many ways:

- improving productivity: by predicting the quality parameters of the product [67], manufacturers can swiftly modify the industrial process setup to fit the updated requirements. Thus, they can save time by using an AI method to provide the best setup which meets their needs;
- improving maintenance: AI algorithms can be used to identify anomalies, and they can also handle large quantities of data [68]. By training an AI model to behave like a device in its normal state, it will be able to identify events that are abnormal (anomalies), which are dangerous, and which can lead to accidents. Prevention is a key factor in reducing them;
- increasing security: the use of privacy-preserving methods for artificial intelligence algorithms will increase security for the client, as well as the provider, making sure that no entity can have access to the model expertise while using the model to generate predictions [69];
- increasing trust in predictions: to improve the accuracy, the model complexity must be increased, and this leads to models being regarded as black boxes. To identify the logic behind a prediction, explainability methods have been developed, and they can be used to identify key features from the input data, which lead to a certain prediction and, thus, an understanding of the model logic.

4.2. Remaining Challenges

In this section, we present and discuss other remaining challenges related to the use of AI in industrial applications. AI models come with inherent risks posed by factors which are outside our control or governance, such as biased datasets or lack of robustness. These are discussed briefly below.

4.2.1. Bias and Fairness

Data remain the base of every learning algorithm, and any issue in the underlying dataset will be reflected in the algorithm performance. One such example is data bias [70]. If datasets present biases, then these will be learned and reflected by the model predictions. Biases can also appear when data are not biased, e.g., due to design choices. Considering the training of a neural network for identifying animals, like dogs, for example, if 95% of dogs in the dataset have brown fur, then the model will not have a good performance on dogs with white fur, and the model might focus on color instead of key features to identify dogs in images.

In the industrial environment, it is crucial to store data yielded by each hardware component (sensors, motors, etc.), so that an algorithm can extract relevant features and learn to predict accurately. Typically, equipment is running in a normal state, and breakdowns are very rare, e.g., once a week, month, or even year. Hence, omitting breakdown events will result in a biased dataset, because the model will react poorly when an abnormal event happens, thus decreasing the trust in its predictions.

Another issue that can be encountered in datasets related to industrial applications is data noise [71]. Training only on noisy data may result in a biased model because, when presenting data without noise or with a decreased level of noise, the model may have a weaker performance. Trying to filter the noise from a dataset may not be a good choice either, since noise happens uncontrollably and is typically present in real-world applications. To address these issues, one needs to make sure that datasets include both normal and noisy data, representative for the actual use of the models.

Another overarching topic is that of fairness [72]. There is no general definition since it is a term which spans across multiple domains such as computer science, psychology, and philosophy. A fair algorithm will be one that is not biased and does not discriminate against individuals, groups, or subgroups. If biases are identified in datasets, it is necessary to eliminate them before performing the training, to ensure that the model will not perpetuate them.

4.2.2. Robustness

Robustness refers to the property that characterizes how effective the AI model is when being tested on a new independent dataset. Specifically, robustness can be linked to the topic of confidence and out-of-distribution detection. It is known that the output of classic deep neural networks may be unreliable when applied on out-of-domain, noisy, or uncertain input data. Many methods have been proposed for assessing model output confidence.

Normalizing flows (NFs) are a family of generative models with tractable distributions, where both sampling and density evaluation can be efficient and exact [73]. The goal is to model $p(x)$, where x denotes samples from a training set and $p(x)$ is the probability distribution. An NF model can answer the following question: given a new set of x, how likely are they to be from the same $p(x)$ distribution (as observed in the training set)? The NF framework employs two components: a bijective encoder (usually employing deep neural networks) and a prior probability distribution (usually a fixed multivariate normal distribution). In contrast with other methods such as variational inference (which only offer a lower bound of $p(x)$ named "evidence lower bound-ELBO"), NFs are capable of fast density estimation, given a suitable choice of model architecture. In that aspect, coupling layers have recently been proposed [74,75] which offer a simple and efficient mechanism for computation of both forward (for density estimation) and backward passes (for sampling). Training can be performed end-to-end in an unsupervised manner.

An NF model can be deployed to detect out-of-distribution (OoD) input samples which should be excluded from the downstream deep neural network (DNN) pipeline. For example, given a model which was trained on a supervised task on a trainset T, an NF model can be trained on the same trainset. If, for new samples, the NF model computes low probability estimates, then those samples are outside the training manifold T of the supervised model, and its predictions may be regarded as unreliable.

Another approach to OoD detection in multiclass classification tasks is to employ softmax logits to compute energy scores, which have been shown to be aligned with the probability density of the inputs and be less susceptible to the overconfidence issue [76]. This approach can distinguish between in- and out-of-distribution samples, even when employing out-of-the-box models which have not been specifically tuned for this purpose. Naturally, OoD detection can be improved by employing outliers and an additional loss term which encourages the network to maximize the separation between energy scores for true samples vs. outliers. They have also stated that "methods relying on the softmax confidence score suffer from overconfident posterior distributions for OoD data". This means that output softmax probabilities tend to be erroneously high for outliers. It is shown that a classification model having a softmax final activation implicitly contains an input density estimator. The energy score can also be incorporated into the training objective, along with the categorical cross-entropy classification loss. The training dataset would consist of two parts: (i) an in-distribution subset, on which the classification loss is computed, and for which the energy score is minimized, and (ii) an OoD subset, on which only the energy score is computed and maximized. The energy method is applicable to an already trained model, extending it as an OoD classifier. Moreover, fine-tuning energy scores during main training can boost OoD detection performance.

Other approaches for analyzing and improving robustness is to use uncertainty estimation techniques. An established method for uncertainty quantification is to employ

Gaussian processes (GPs). Figure 5 illustrates the performance of two methods using GPs for uncertainty quantification: red dots are the predictions, deep blue lines represent the label values, and light blue areas represent the certainty limits. In their original formulation, they lack efficiency for large dimensional datasets, but recent studies showcased that using variational approaches and certain architectural constraints, highly efficient models can be obtained, which offer high task-specific performance and uncertainty estimates using only a single forward pass of the neural network [77]. A smooth and sensitive feature extractor feeds a sparse variational Gaussian process, which outputs both the expected prediction and the output variance.

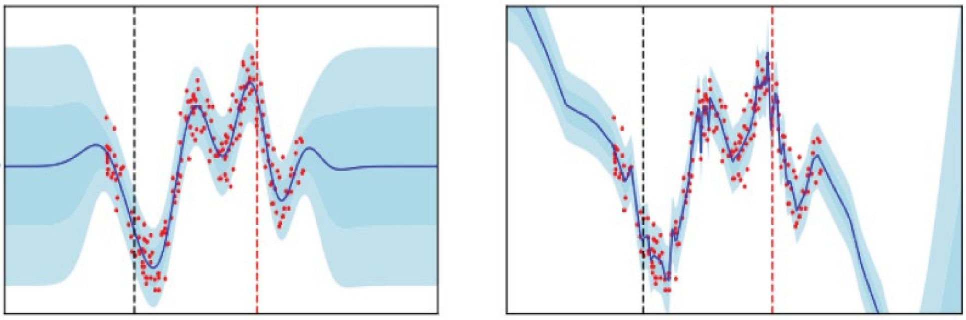


Figure 5. Predictions and attached uncertainties on a toy dataset from a squared exponential (SE) kernel (**left**) and a deep kernel learning (DKL) kernel (**right**). Outside the training points, the SE kernel model reverts to its zero mean with high uncertainty, while the DKL extrapolates confidently (adapted with permission from [77], 2021).

To improve robustness for deep neural network models, Kwon [78] proposed the concept of a backdoor attack, to retrain the model with additional altered images that were intentionally introduced to mislead the algorithm. A backdoor attack relies on introducing inputs that contain certain triggers which determine the model to output a wrong prediction. After model training, triggers may be introduced in the test input data to try to manipulate the predictions. The advantage of the proposed method is that the classification accuracy is maintained at a high level, even when the model is under attack.

Text datasets are also subject to potential robustness issues. These may be addressed similarly to imaging datasets, i.e., by adding noise over some of the samples used in the training process. These samples with noise are also known as adversarial examples; noise is represented by modified characters or words in a paragraph. As for images with noise, text altered by noise should have the same meaning, i.e., a human reader should perceive it similarly to text without noise. By introducing adversarial examples in the training dataset, Kwon and Lee [79] increased the model accuracy over altered samples by 13.3% (from 9.2% to 22.5%), while the overall accuracy dropped only by 0.9% (from 88.1% to 87.2%).

5. Conclusions

Artificial intelligence is widely used and brings benefits in every domain in which it is applied. Industry 4.0 has created the conditions to apply artificial intelligence through digitalization. Large quantities of data are now available to be used to generate knowledge, which is exactly what AI algorithms have been designed for.

Even though it can improve industrial processes, AI needs to be applied with caution, because the access to large quantities of data also has downsides, such as the risk of data theft. To prevent this, privacy-preserving methods have been developed to be used along with AI algorithms, to ensure that knowledge is generated, while maintaining data safety. Privacy-preserving solutions proposed to date have the disadvantage of either increasing the runtime by a prohibitive amount or decreasing the accuracy significantly. Thus, the tradeoff between privacy preservation and usability is still too large. Further research is warranted to develop solutions which can be considered both secure and accurate enough.

Another aspect related to artificial intelligence is the black-box nature of the models. Increasing the accuracy means increasing the model complexity, thus making it harder to interpret how a model takes the decision starting from the given inputs. Explainable AI has been introduced to fill this gap and to help understand the way a model maps inputs to outputs. Current solutions are limited to certain models, and state-of-the-art AI approaches, such as deep neural networks, require further research to ensure levels of transparency that would allow the user to fully trust AI model decisions.

As AI techniques evolve, newly developed concepts will be translated into the various application domains, including industry. Similarly, new challenges will be identified, which will need to be addressed first at a core or theoretical level, and then within the application domains. Two such current challenges were described herein: bias/fairness and robustness. AI model robustness is closely linked to AI model explainability; a robust model will perform well even when being presented with a data sample that has distinct properties from those of the training data samples. A robust AI model performs well on such out-of-distribution data, specifically because it is capable of recognizing and interpreting certain characteristics of the data sample, even if they have a slightly different appearance than in the training dataset. Such model capability is also crucial for achieving high model explainability; a model that generalizes well takes the decisions on the basis of the right characteristics of the data samples, which in turn means that it can potentially explain its decisions correctly, i.e., generating trust. Furthermore, AI model robustness is also linked to AI model bias. A model without or with low bias is likely to achieve a superior robustness by removing or at least reducing so-called 'blind spots' in the data processing and interpretation. We also note that privacy-preserving methods increase complexity, since they introduce an additional layer of data manipulation. Specifically, the data manipulation methodology itself should not introduce any bias in the data and maintain the same level of robustness as if the model was trained on the original data. Lastly, we note that privacy preservation and explainability requirements apparently have opposite effects. Privacy preservation is typically achieved by encrypting/obfuscating/altering the model input data, which in turn diminishes the explainability and interpretability capabilities. Hence, at least with current approaches, the user has to choose which of these aspects should be prioritized.

As a limitation, we note that this argumentative review does not represent an exhaustive attempt at discussing the application of AI in industrial applications. We focused on specific challenges and highlighted recent developments related to these challenges, and we identified other challenges to be considered in future research.

Author Contributions: Conceptualization, I.O. and L.I.; methodology, I.O., A.V., C.C. and A.P.; validation, S.C., C.B., A.I., R.D., F.M. and C.S.; resources, L.I.; writing—original draft preparation, I.O., C.C., A.P. and L.I.; writing—review and editing, R.D., F.M. and C.S.; visualization, I.O.; supervision, C.S.; project administration, L.I.; funding acquisition, L.I. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the Romanian National Authority for Scientific Research and Innovation, CCCDI–UEFISCDI, project number ERANET-PERMED-PROGRESS, within PNCDI III.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not alicable.

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

References

1. Ghobakhloo, M. Industry 4.0, digitization, and opportunities for sustainability. *J. Clean Prod.* **2020**, *252*, 119869. [CrossRef]
2. Kumar, S.; Tiwari, P.; Zymbler, M. Internet of Things is a revolutionary approach for future technology enhancement: A review. *J. Big Data* **2019**, *6*, 111. [CrossRef]

3. Cardin, O. Classification of cyber-physical production systems applications: Proposition of an analysis framework. *Comput. Ind.* **2019**, *104*, 11–21. [CrossRef]
4. Wang, J.; Yang, Y.; Wang, T.; Sheratt, R.S.; Zhang, J. Big data service architecture: A survey. *J. Internet Technol.* **2020**, *21*, 393–405.
5. Chen, Z.; Ye, R. Principles of Creative Problem Solving in AI Systems. *Sci. Educ.* **2022**, *31*, 555–557. [CrossRef]
6. Fahle, S.; Prinz, C.; Kuhlenkotter, B. Systematic review on machine learning (ML) methods for manufacturing processes–Identifying artificial intelligence (AI) methods for field application. *Procedia CIRP* **2020**, *93*, 413–418. [CrossRef]
7. Zhang, C.; Lu, Y. Study on artificial intelligence: The state of the art and future prospects. *J. Ind. Inf. Integr.* **2021**, *23*, 100224. [CrossRef]
8. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2022**, *54*, 1–35. [CrossRef]
9. Varghese, J.; Kleine, M.; Gessner, S.I.; Sandmann, S.; Dugas, M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: A systematic review. *J. Am. Med. Inform. Assn.* **2018**, *25*, 593–602. [CrossRef]
10. Kotsiopoulos, T.; Sarigiannidis, P.; Ioannidis, D.; Tzovaras, D. Machine Learning and Deep Learning in smart manufacturing: The Smart Grid paradigm. *Comput. Sci. Rev.* **2021**, *40*, 100341. [CrossRef]
11. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, e00938. [CrossRef] [PubMed]
12. Sharma, A.; Zhang, Z.; Rai, R. The interpretive model of manufacturing: A theoretical framework and research agenda for machine learning in manufacturing. *Int. J. Prod. Res.* **2021**, *59*, 4960–4994. [CrossRef]
13. Rai, R.; Tiwari, M.K.; Ivanov, D.; Dolgui, A. Machine learning in manufacturing and industry 4.0 applications. *Int. J. Prod. Res.* **2021**, *59*, 4773–4778. [CrossRef]
14. Bertolini, M.; Mezzogori, D.; Neroni, M.; Zammori, F. Machine Learning for industrial applications: A comprehensive literature review. *Expert Syst. Appl.* **2021**, *175*, 114820. [CrossRef]
15. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef] [PubMed]
16. Rao, N.T. A review on industrial applications of machine learning. *Int. J. Disast. Recov. Bus. Cont.* **2018**, *9*, 1–9.
17. Peres, R.S.; Jia, X.; Lee, J.; Sun, K.; Colombo, A.W.; Barata, J. Industrial Artificial Intelligence in Industry 4.0–Systematic Review, Challenges and Outlook. *IEEE Access* **2020**, *8*, 220121–220139. [CrossRef]
18. Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **2019**, *28*, 231–237. [CrossRef]
19. Rai, A. Explainable AI: From black box to glass box. *Acad. Mark. Sci. Rev.* **2020**, *48*, 137–141. [CrossRef]
20. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18. [CrossRef]
21. Messalas, A.; Kanellopoulos, Y.; Makris, C. Model-Agnostic Interpretability with Shapley Values. In Proceedings of the 10th International Conference on Information, Intelligence, Systems and Applications (IISA 2019), Patras, Greece, 15–17 July 2019; pp. 1–7.
22. Palatnik de Sousa, I.; Maria Bernardes Rebuzzi Vellasco, M.; Costa da Silva, E. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors* **2019**, *19*, 2969. [CrossRef]
23. Antwarg, L.; Miller, R.M.; Shapira, B.; Rokach, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst. Appl.* **2021**, *186*, 115736. [CrossRef]
24. Liang, Y.; Li, S.; Yan, C.; Li, M.; Jiang, C. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* **2019**, *419*, 168–182. [CrossRef]
25. Orlandi, C.; Piva, A.; Barni, M. Oblivious Neural Network Computing via Homomorphic Encryption. *Eurasip J. Inf.* **2007**, *2007*, 1–11. [CrossRef]
26. Vizitiu, A.; Nita, C.I.; Toev, R.M.; Suditu, T.; Suciu, C.; Itu, L.M. Framework for Privacy-Preserving Wearable Health Data Analysis: Proof-of-Concept Study for Atrial Fibrillation Detection. *Appl. Sci.* **2021**, *11*, 9049. [CrossRef]
27. Sun, X.; Zhang, P.; Liu, J.K.; Yu, J.; Xie, W. Private Machine Learning Classification Based on Fully Homomorphic Encryption. *IEEE Trans. Emerg. Top. Comput.* **2020**, *8*, 352–364. [CrossRef]
28. Aslett, L.J.; Esperança, P.M.; Holmes, C.C. A review of homomorphic encryption and software tools for encrypted statistical machine learning. *arXiv* **2015**, arXiv:1508.06574.
29. Takabi, H.; Hesamifard, E.; Ghasemi, M. Privacy preserving multi-party machine learning with homomorphic encryption. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
30. Li, J.; Kuang, X.; Lin, S.; Ma, X.; Tang, Y. Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. *Inf. Sci.* **2020**, *526*, 166–179. [CrossRef]
31. Wood, A.; Najarian, K.; Kahrobaei, D. Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. *ACM Comput. Surv.* **2021**, *53*, 1–35. [CrossRef]
32. Fang, H.; Qian, Q. Privacy Preserving Machine Learning with Homomorphic Encryption and Federated Learning. *Future Internet* **2021**, *13*, 94. [CrossRef]
33. Khan, L.U.; Saad, W.; Han, Z.; Hossain, E.; Hong, C.S. Federated Learning for Internet of Things: Recent Advances, Taxonomy, and Open Challenges. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1759–1799. [CrossRef]

34. Oh, S.J.; Schiele, B.; Fritz, M. Towards Reverse-Engineering Black-Box Neural Networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 1st ed.; Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R., Eds.; Springer: Cham, Switzerland, 2019; Volume 11700, pp. 121–144. ISBN 978-3-030-28954-6.
35. Google Cloud. Accelerate Your Transformation with Google Could. 2022. Available online: https://cloud.google.com/ (accessed on 10 March 2022).
36. Azure Machine Learning. An Enterprise-Grade Service for the End-to-End Machine Learning Lifecycle. 2022. Available online: https://azure.microsoft.com/en-us/services/machine-learning/ (accessed on 10 March 2020).
37. Fu, A.; Zhang, X.; Xiong, N.; Gao, Y.; Wang, H.; Zhang, J. VFL: A Verifiable Federated Learning with Privacy-Preserving for Big Data in Industrial IoT. *IEEE Trans. Industr. Inform.* **2020**, *18*, 3316–3326. [CrossRef]
38. Girka, A.; Terziyan, V.; Gavriushenko, M.; Gontarenko, A. Anonymization as homeomorphic data space transformation for privacy-preserving deep learning. *Procedia Comput. Sci.* **2021**, *180*, 867–876. [CrossRef]
39. Zhao, Y.; Zhao, J.; Jiang, L.; Tan, R.; Niyato, D.; Li, Z.; Lyu, L.; Liu, Y. Privacy-Preserving Blockchain-Based Federated Learning for IoT Devices. *IEEE Internet Things J.* **2021**, *8*, 1817–1829. [CrossRef]
40. Gonçalves, C.; Bessa, R.J.; Pinson, P. Privacy-Preserving Distributed Learning for Renewable Energy Forecasting. *IEEE Trans. Sustain. Energy* **2021**, *12*, 1777–1787. [CrossRef]
41. Goodfellow, I.; Jean, P.A.; Mehdi, M.; Bing, X.; David, W.F.; Sherjil, O.; Aaron, C.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
42. Li, Y.; Li, J.; Wang, Y. Privacy-Preserving Spatiotemporal Scenario Generation of Renewable Energies: A Federated Deep Generative Learning Approach. *IEEE Trans. Industr. Inform.* **2021**, *18*, 2310–2320. [CrossRef]
43. Kaggle. Casting Product Image Data for Quality Inspection–Dataset. Available online: https://www.kaggle.com/datasets/ravirajsinh45/real-life-industrial-dataset-of-casting-product (accessed on 20 May 2022).
44. Popescu, A.B.; Taca, I.A.; Vizitiu, A.; Nita, C.I.; Suciu, C.; Itu, L.M.; Scafa-Udriste, A. Obfuscation Algorithm for Privacy-Preserving Deep Learning-Based Medical Image Analysis. *Appl. Sci.* **2022**, *12*, 3997. [CrossRef]
45. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef]
46. Holm, E.A. In defense of the black box. *Science* **2019**, *364*, 26–27. [CrossRef]
47. Fedotova, A.; Romanov, A.; Kurtukova, A.; Shelupanov, A. Authorship Attribution of Social Media and Literary Russian-Language Texts Using Machine Learning Methods and Feature Selection. *Future Internet* **2021**, *14*, 4. [CrossRef]
48. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent individualized feature attribution for tree ensembles. *arXiv* **2018**, arXiv:1802.03888.
49. Kaya, M.; Bilge, H.Ş. Deep metric learning: A survey. *Symmetry* **2019**, *11*, 1066. [CrossRef]
50. Sahakyan, M.; Aung, Z.; Rahwan, T. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access* **2021**, *9*, 135392–135422. [CrossRef]
51. Poulin, B.; Eisner, R.; Szafron, D.; Lu, P.; Greiner, R.; Wishart, D.S.; Fushe, A.; Pearcy, B.; MacDonell, C.; Anvik, J. Visual explanation of evidence with additive classifiers. In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI 2006), Boston, MA, USA, 16–20 July 2006; pp. 1822–1829.
52. Ribeiro, M.T.; Singh, S.; Guestrin, C. 'Why should i trust you? ': Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
53. Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N.T.P.; Ramasamy, S.; DeCost, B.L.; Tian, S.I.; Romano, G.; et al. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *NPJ Comput. Mater.* **2019**, *5*, 1–9. [CrossRef]
54. Thrun, M.C.; Ultsch, A.; Breuer, L. Explainable AI Framework for Multivariate Hydrochemical Time Series. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 170–204. [CrossRef]
55. Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; Pushkin, S. Towards Best Practice in Explaining Neural Network Decisions with LRP. In Proceedings of the International Joint Conference on Neural Networks (IJCNN 2020), Glasgow, UK, 19–24 July 2020; pp. 1–7.
56. Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* **2022**, *77*, 29–52. [CrossRef]
57. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In Proceedings of the Natural Language Processing and Chinese Computing (NLPCC 2019), Dunhuang, China, 9–14 October 2019; pp. 564–574.
58. Degas, A.; Islam, M.R.; Hurter, C.; Barua, S.; Rahman, H.; Poudel, M.; Ruscio, D.; Ahmed, M.U.; Begum, S.; Rahman, M.A.; et al. A Survey on Artificial Intelligence (AI) and Explainable AI in Air Traffic Management: Current Trends and Development with Future Research Trajectory. *Appl. Sci.* **2022**, *12*, 1295. [CrossRef]
59. Gade, K.; Geyik, C.; Kenthapadi, K.; Mithal, V.; Taly, A. Explainable AI in Industry. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), Ankorage, AK, USA, 4–8 August 2019; pp. 3203–3204.

60. Longo, L.; Goebel, R.; Lecue, F.; Kieseberg, P.; Holzinger, A. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2020), Dublin, Ireland, 25–28 August 2020; pp. 1–16.
61. Ahmed, I.; Jeon, G.; Piccialli, F. From Artificial Intelligence to eXplainable Artificial Intelligence in Industry 4.0: A survey on What, How, and Where. *IEEE Trans. Industr. Inform.* **2022**, *18*, 5031–5042. [CrossRef]
62. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *arXiv* **2021**, arXiv:2112.11561.
63. Glomsrud, J.A.; Ødegårdstuen, A.; Clair, A.L.S.; Smogeli, Ø. Trustworthy versus Explainable AI in Autonomous Vessels. In Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV 2019), Helsinki, Finland, 17–18 September 2019; pp. 37–47.
64. Krishnamurthy, V.; Nezafati, K.; Stayton, E.; Singh, V. Explainable AI Framework for Imaging-Based Predictive Maintenance for Automotive Applications and Beyond. *Data-Enabled Discov. Appl.* **2020**, *4*, 7. [CrossRef]
65. Brito, L.C.; Susto, G.A.; Brito, J.N.; Duarte, M.A.V. An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. *Mech. Syst. Signal Pr.* **2022**, *163*, 108105. [CrossRef]
66. Chelgani, S.C.; Nasiri, H.; Tohry, A. Modeling of particle sizes for industrial HPGR products by a unique explainable AI tool- A "Conscious Lab" development. *Adv. Powder Technol.* **2021**, *32*, 4141–4148. [CrossRef]
67. Ahmed, A.N.; Othman, F.B.; Afan, H.A.; Ibrahim, R.K.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084. [CrossRef]
68. Himeur, Y.; Ghanem, K.; Alsalemi, A.; Bensaali, F.; Amira, A. Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Appl. Energy* **2021**, *287*, 116601. [CrossRef]
69. Asad, M.; Moustafa, A.; Ito, T. FedOpt: Towards Communication Efficiency and Privacy Preservation in Federated Learning. *Appl. Sci.* **2020**, *10*, 2864. [CrossRef]
70. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, U.; Nejdl, W.; Vidal, M.E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems–An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, 1356. [CrossRef]
71. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–19. [CrossRef]
72. Madaio, M.; Egede, L.; Subramonyam, H.; Vaughan, J.W.; Wallach, H. Assessing the Fairness of AI Systems: AI Practitioners Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–26. [CrossRef]
73. Kobyzev, I.; Prince, S.J.; Brubaker, M.A. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3964–3979. [CrossRef]
74. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using Real NVP. *arXiv* **2016**, arXiv:1605.08803.
75. Kingma, D.P.; Dhariwal, P. Glow: Generative Flow with Invertible 1×1 Convolutions. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2018), Montreal, QC, Canada, 3–8 December 2018.
76. Liu, W.; Wang, X.; Owens, J.; Li, Y. Energy-based Out-of-distribution Detection. *Adv. Neural. Inf. Process. Syst.* **2020**, *33*, 21464–21475.
77. Ober, S.W.; Rasmussen, C.E.; van der Milk, M. The Promises and Pitfalls of Deep Kernel Learning. In Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021), Toronto, ON, Canada, 26–30 July 2021.
78. Kwon, H.; Kim, Y. BlindNet backdoor: Attack on deep neural network using blind watermark. *Multimed. Tools Appl.* **2022**, *81*, 6217–6234. [CrossRef]
79. Kwon, H.; Lee, S. Textual Adversarial Training of Machine Learning Model for Resistance to Adversarial Examples. *Secur. Commun. Netw.* **2022**, *2022*, 4511510. [CrossRef]