applied sciences

MDPI

Review

Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review

Thi-Thu-Huong Le [1,2,*], Aji Teguh Prihatno [3], Yustus Eko Oktian [1,2], Hyoeun Kang [3] and Howon Kim [3,*]

[1] Blockchain Platform Research Center, Pusan National University, Busan 609735, Republic of Korea; yustus@islab.re.kr
[2] IoT Research Center, Pusan National University, Busan 609735, Republic of Korea
[3] School of Computer Science and Engineering, Pusan National University, Busan 609735, Republic of Korea; ajiteguh@pusan.ac.kr (A.T.P.); hyoeun405@gmail.com (H.K.)
* Correspondence: lehuong7885@gmail.com (T.-T.-H.L.); howonkim@pusan.ac.kr (H.K.)

Abstract: In recent years, numerous explainable artificial intelligence (XAI) use cases have been developed, to solve numerous real problems in industrial applications while maintaining the explainability level of the used artificial intelligence (AI) models to judge their quality and potentially hold the models accountable if they become corrupted. Therefore, understanding the state-of-the-art methods, pointing out recent issues, and deriving future directions are important to drive XAI research efficiently. This paper presents a systematic literature review of local explanation techniques and their practical applications in various industrial sectors. We first establish the need for XAI in response to opaque AI models and survey different local explanation methods for industrial AI applications. The number of studies is then examined with several factors, including industry sectors, AI models, data types, and XAI-based usage and purpose. We also look at the advantages and disadvantages of local explanation methods and how well they work in practical settings. The difficulties of using local explanation techniques are also covered, including computing complexity and the trade-off between precision and interpretability. Our findings demonstrate that local explanation techniques can boost industrial AI models' transparency and interpretability and give insightful information about them. The efficiency of these procedures must be improved, and ethical concerns about their application must be resolved. This paper contributes to the increasing knowledge of local explanation strategies and offers guidance to academics and industry professionals who want to use these methods in practical settings.

Keywords: machine learning; explainable artificial intelligence; local explanation techniques; industrial application; trustworthiness

Citation: Le, T.-T.-H.; Prihatno, A.T.; Oktian, Y.E.; Kang, H.; Kim, H. Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review. *Appl. Sci.* **2023**, *13*, 5809. https://doi.org/10.3390/app13095809

Academic Editors: Esteban García-Cuesta, Manuel Castillo-Cara and Ricardo Aler Mur

Received: 12 April 2023
Revised: 1 May 2023
Accepted: 5 May 2023
Published: 8 May 2023

1. Introduction

Machine learning (ML) and deep learning (DL) models have achieved remarkable success in a variety of domains, including healthcare [1–3], financial systems [4,5], criminal justice [6,7], and cybersecurity [8,9]. While accuracy is critical, the emphasis on accuracy has frequently resulted in developers sacrificing interpretability in favor of better accuracy by making models more complex and difficult to comprehend [10]. When the learning model has the authority to make critical decisions that influence people's well-being, this lack of interpretability becomes a major concern.

To overcome this issue, explainable artificial intelligence (XAI) approaches must provide end-users with coherent explanations of these models' decision-making processes. Indeed, because these models are black boxes, it is difficult to understand how they arrive at their conclusions. XAI technologies provide visualization techniques to comprehend the decision-making processes of these models, making forecast explanations easier to understand and communicate. Although the concept of XAI has recently received significant

attention in studies [11,12], it has only recently received significant attention in academia, with an increasing number of research papers being published on the topic [13].

The increasing use of ML and DL models in various applications has highlighted the need to explain the decision-making processes to gain end-users' trust in industrial applications. However, it is essential to investigate the effectiveness and limitations of local explanation techniques in industrial settings. To address this research gap, we conducted a literature review that focused on XAI studies within various industrial sectors, AI models, data types, and the usage and purpose of XAI. Our study stands out from the existing literature survey papers by covering a comprehensive set of criteria, including data types and industrial sectors, as illustrated in Table 1. As shown in Table 1, our survey is more comprehensive and distinctive than previously published surveys in two main criteria: data types and industrial sectors. While many existing surveys only cover one industrial sector or one or two data types, our analysis aims to better understand the local explanation in industrial settings and identify future research directions.

Table 1. Comparison of existing surveys with our work.

Survey Paper	Year	Data Type				Industrial Sector							
		T	*I*	*Ta*	*Te*	*ASR*	*EBM*	*E*	*F*	*H*	*IE*	*C*	*SA*
[14]	2020	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
[15]	2021	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
[16]	2021	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
[17]	2021	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
[18]	2022	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
[19]	2022	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
[20]	2022	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
[21]	2022	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
[22]	2022	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
[23]	2022	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗
[24]	2022	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
[25]	2022	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
[26]	2022	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
[27]	2022	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
[28]	2022	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗
[29]	2023	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
[30]	2023	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
[31]	2023	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
Ours	2023	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Legend: ✓ means included, ✗ means not included. T: time series; I: image; Ta: table; Te: text; ASR: autonomous systems and robotics; EBM: energy building management; E: environment; F: finance; H: healthcare; IE: industrial engineering; C: cybersecurity; SA: smart agriculture

This paper's main goal is to assess the state of local explanation strategies today thoroughly and explore how they are used in the marketplace. To accomplish this, we present a methodical literature analysis that identifies the gaps and limits in the existing research and analyzes and examines the efficacy and limitations of various local explanation strategies in deriving insights from complex models in the industrial setting. Furthermore, we provide recommendations for future research, including areas where further investigation is needed to improve the effectiveness and applicability of local explanation techniques in the

industry. The research motivation emphasizes the precision of the study and the void in the research and highlights the research objectives and questions that need to be addressed throughout the paper. The main contributions of the paper are as follows:

- Identifying and categorizing local explanation techniques for industrial AI applications based on different criteria categories, including XAI usage and purpose, industrial applications and AI models, and data types.
- Analyzing the advantages and disadvantages of different local explanation techniques in explaining complex models in the industrial context.
- Identifying the current challenges of local explanation for industrial AI applications and suggesting recommendations for addressing these challenges, such as developing more effective and efficient local explanation techniques and exploring the ethical implications of using these techniques.
- Summarizing the study's main findings, identifying gaps and limitations in the current literature, and suggesting future research directions, which can guide future research efforts and lead to more reliable and trustworthy AI systems in industry.

The paper is structured as follows: In Section 2, we describe the methodology we used for the literature review. Section 3 presents the survey results on local explanation techniques and their practical applications in industry. We first provide an overview of local explanation techniques and then analyze their distribution based on different criteria categories, including XAI usage and purpose, industrial applications and AI models, and data types. We then highlight the advantages and disadvantages of these techniques in explaining complex models in the industrial context and identify several challenges. Furthermore, we provide recommendations for addressing these challenges. In Section 4, we summarize the main findings of our study, identify gaps and limitations in the current literature, and suggest future research directions. Finally, in Section 5, we conclude the paper and recommend future research on local explanation techniques and their practical applications in industry.

2. Literature Review Methodology

This section outlines our methodology for conducting a comprehensive literature review on local explanation techniques and their practical applications in industry. Our discussion of local explanation techniques and their practical applications in industry is built upon a thorough and objective literature study, made possible by a rigorous search strategy, inclusion and exclusion criteria, data extraction and analysis, and quality assessment. The workflow of the survey methodology is shown in Figure 1. We used a clear process that included the following phases to attain this goal:

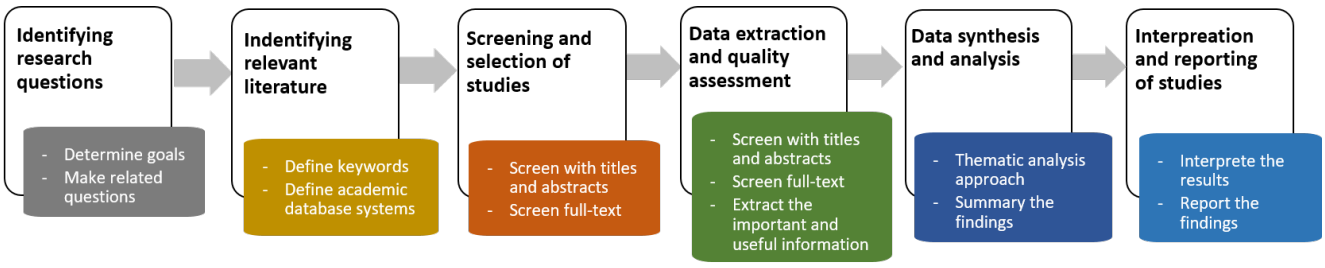


Figure 1. The workflow of the survey on the local explanation for industrial AI applications.

1. *Identifying the research question*: Our research question focused on understanding the different types of local explanation techniques used in industry, their benefits and limitations, and their effectiveness in explaining complex machine learning models. Our review aims to answer the following research questions:
   - (Q1) What local explanation techniques are used in industrial applications?
   - (Q2) How widespread are practical industrial applications of local explanation techniques?

- (Q3) What are the benefits and limitations of local explanation techniques for industrial applications?
- (Q4) How to build effective local explanation techniques in practical settings?

By answering these questions, we hope to provide a comprehensive overview of the current state of the art in local explanation techniques for practical industrial applications.

2. *Identifying relevant literature*: We conducted a comprehensive search of relevant academic and industry sources, including peer-reviewed journals, conference proceedings, technical reports, and gray literature. The search included studies published from 2020 to Mar 2023. We used a combination of keywords and controlled vocabulary terms related to machine learning, interpretability, local explanations, and industrial applications to identify relevant articles. We specifically examined academic resources such as ACM Digital Library, IEEE Xplore, ScienceDirect, Google Scholar, MDPI, and others. The search terms used include "local explanation techniques", "model interpretability", "explainable artificial intelligence", "XAI", "machine learning", and "industrial applications". We also manually searched relevant journals and conference proceedings to ensure comprehensive coverage of the literature.
3. *Screening and selection of studies*: We used a two-stage screening process to identify articles for inclusion in our review. In the first stage, we screened titles and abstracts to identify potentially relevant articles. In the second stage, we screened the full text of articles to determine their eligibility based on our inclusion and exclusion criteria. We included studies focusing on local explanation techniques and their applications in industrial settings. Studies investigating the effectiveness and limitations of different explanation techniques and their comparative analysis were also considered. We excluded studies focusing on global explanations, theoretical aspects of model interpretability, or applications in non-industrial settings.
4. *Data extraction and quality assessment*: Data from the selected papers were extracted, including study design, sample size, research question, methods, results, and limitations. We extracted data from the selected studies, including the authors, publication year, research question, methodology, datasets used, and main findings. The data were analyzed thematically to identify patterns, trends, and research gaps in the literature. We also performed a qualitative synthesis of the studies, highlighting the benefits and limitations of local explanation techniques and their effectiveness in industrial applications.
5. *Data synthesis and analysis*: We synthesized the data from the selected articles using a thematic analysis approach. We identified common themes and patterns across the articles and summarized the findings in a narrative synthesis.
6. *Interpretation and reporting of results*: We interpreted the results of our review in light of our research question and the existing literature. We reported our findings in a structured manner, highlighting the key themes and patterns that emerged from our analysis.

Overall, our literature review methodology was designed to ensure our review's comprehensiveness, validity, and rigor. We set out to present a thorough and trustworthy evaluation of the available literature on local explanation strategies in industrial applications, using a well-defined methodology. One hundred and one papers were chosen using the aforementioned six procedures. The literature review was completed by 31 March 2023. These papers were categorized according to the industrial sectors, data types, applied AI model, used dataset, and practical local explanation techniques in terms of usage and purpose. After that, we analyzed the growth and distribution of local explanations for industrial AI applications following several important factors, such as the AI model, industrial sector, data types, usage, and purpose. Finally, several representative AI industrial application methods were chosen to analyze the pros and cons in depth, as represented in Section 3.

3. Survey Results on Local Explanation in Industrial AI Applications

The taxonomy of XAI techniques, as mentioned in [32–35], follows four aspects: (1) scope or level explanation; (2) purpose explanation; (3) implement or usage explanation; (4) data types. First, both global and local explanations are included at the explanation level. The models' operational and decision-making processes and their ability to be articulated are generally emphasized by each explanation level. For many purposes, selecting the explanation level depends on specific cases or subgroups. The survey scope with the local explanation level is the main priority of this paper. Second, there are post hoc and intrinsic (ante hoc) approaches regarding XAI's purpose. The characteristics, decision-making procedures, and rules of the models, such as linear models, decision trees, Bayesian networks, etc., provide internal justifications for the model type. Meanwhile, post hoc reasoning holds that black-box models' innermost workings and decision-making procedures are revealed after training. Numerous post hoc XAI solutions have been created, because post hoc explainers include anything from black-box models to interpretable models. Third, usage or implementation includes both agnostic and specific models. Model-specific approaches use the model's characteristics and properties to make it [36] interpretable. Such approaches' strengths come at a cost, because applying them to different models is challenging. What gives them their power is their access to model internals such as weights or structures. Meanwhile, post hoc explanations can be provided after the model has been trained using model-agnostic techniques, which can be used for any ML model. Since these methods can only examine input–output pairings, they have the drawback of being unable to benefit from model internals. Fourth, from an XAI-based data perspective, a data type can be tabular, text, image, time series, or any other type.

Most proposed explainable AI systems from industrial applications have been tested, including autonomous systems and robotics, energy and building management, security and privacy, environment, finance, industrial engineering, healthcare, and smart agriculture. The year-by-year distribution of publications on the local explanation of industrial AI applications from 2020 to March 2023 is depicted in Figure 2. According to the literature review, researchers increasingly use local XAI approaches in industrial applications. In particular, the links between XAI approaches and different industrial application areas are shown in Table 2.

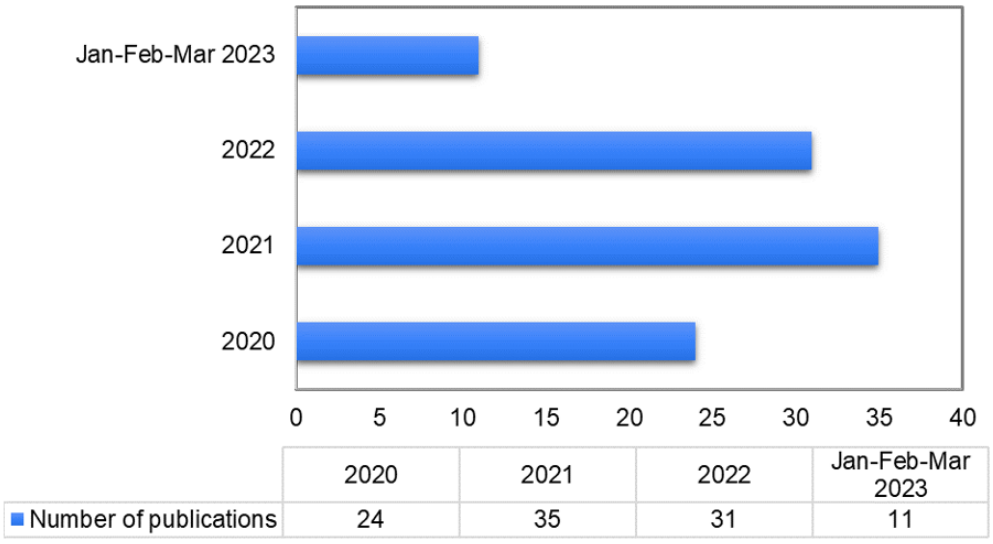


Figure 2. Year-wise distribution of publications on the local explanation of industrial AI applications.

Table 2. Local explanation on AI industrial applications.

Industrial Sector	Data Type	Ref.	Year	AI Model and Dataset		Practical Local Explanation		
				AI Model	*Dataset*	*XAI Method*	*Usage*	Purpose
Autonomous systems and robotics	Time series	[37]	2020	Deep Q-Networks	Voltage load	SHAP	A	P
		[38]	2022	FL	QoE forecasting	FED-XAI	S	An
	Image	[39]	2020	Q-learning	Agent data	Interaction data	A	P
		[40]	2021	DRL	Depth image, UAV states	SHAP-CAM	A	P
		[41]	2022	Faster R-CNN, ResNet-50	BDD-OID, PSI	t-SNE	S	An
	Tabular	[42]	2022	RF	Processed dataset (D*)	SHAP	A	P
	Text	[43]	2021	DT, RF	Lyft Level 5	SHAP	A	P
Energy and building management	Time series	[44]	2020	Seq2seq	EnergyPlus	AM	A	P
		[45]	2020	MLRi, GBT	CBECS	SHAP	A	P
		[46]	2020	XGBoost	PLUTO	SHAP	A	P
		[47]	2020	RF	GEFCOM	SHAP, LIME, ELI5	A	P
		[48]	2021	XGBoost	Historical climate	SHAP	A	P
		[49]	2021	DNN	IPCC's SRES A2 climate	SHAP	A	P
		[50]	2021	DNN	Energy data	IG, DeepLIFT	A	P
		[51]	2021	LSTM	Energy consumption	AM	S	An
		[52]	2021	LSTM	Heat demand by FMEDH	LIME	A	P
		[53]	2022	XGBoost	UNICON dataset	SHAP	A	P
		[54]	2022	CatBoost	Energy disclosure	LIME	A	P
		[55]	2022	DNN	Enerate synthetic	LIME	A	P
		[56]	2022	LSTM	Electricity load	LIME	A	P
		[57]	2022	GS-XGBoost	Electricity consumption	SHAP	A	P
		[58]	2022	Cubist regression	Electricity two buildings	FI	S	An
		[59]	2022	QLattice	Residential building	FI	S	An
		[60]	2022	Bi-LSTM, CNN-LSTM	SCADA	LIME	A	P
	Tabular	[61]	2021	XGBoost	Singapore's building	LIME	A	P
		[62]	2022	RF, CNN, IT	CER	LIME, SHAP	A	P
		[63]	2021	XGBoost, SVR, LightGBM, LSTM	Energy data in Seoul	SHAP	A	P

Table 2. *Cont.*

Industrial Sector	Data Type	Ref.	Year	AI Model and Dataset		Practical Local Explanation		
				AI Model	*Dataset*	*XAI Method*	*Usage*	Purpose
Environment	Time series	[64]	2020	XGBoost, LSTM-RNN	Escherichia coli	SHAP	A	P
		[65]	2022	DFNN	Building-level damage	LIME	A	P
	Image	[66]	2020	RF	Occurrence data	LIME	A	P
		[67]	2021	LSTM, CNN-BiLSTM	Monthly rainfall data	SHAP	A	P
		[68]	2023	XGBoost	Land-cover, topographic	LIME	A	P
Finance	Tabular	[69]	2020	XGBoost	Sports and travel	SHAP	A	P
		[70]	2021	GBDT	Daily observations	SHAP	A	P
		[71]	2021	LSTM	OHLC	LIME	A	P
		[72]	2021	XGBoost	Financial indicators	SHAP	A	P
		[73]	2022	LightGBM	Proprietary	SHAP	A	P
	Time series	[74]	2020	XGBoost	Credit risk	SHAP	A	P
		[75]	2022	DQN	SENSEX, DJIA	SHAP	A	P
Healthcare	Tabular	[76]	2020	DT	Cervical cancer risk	SHAP	A	P
		[77]	2020	Stacked LSTM-CNN-MLP	PhysioNet	SHAP, occlusion maps	A	P
		[78]	2020	DT	MIMIC-III	Doctor XAI	S	An
		[79]	2021	XGBoost	Diabetes	PDP, ICE, ALE, LIME, SHAP, Anchors	A	P
		[80]	2021	XGBoost	ROSMAP	SHAP	A	P
		[81]	2021	LightGBM	K-attention	SHAP	A	P
		[82]	2021	RF, GBDT, XGBoost	UCI CKD	SHAP	A	P
		[83]	2021	RF	Retrospective study	SHAP	A	P
	Time series	[84]	2020	Bi-LSTM ensemble	PhysioNet 2017	Attention	S	An
		[85]	2021	CNN	PhysioNet 2017	LIME, guided saliency	A	P
		[86]	2021	WCPH-RNN	Retrospective study	Saliency	A	P
		[87]	2021	CNN	Gait dataset	LRP	A	P
		[88]	2022	CNN-CRF	Sleep-EDF	Grad-CAM	A	P

Table 2. *Cont.*

Industrial Sector	Data Type	Ref.	Year	AI Model and Dataset		Practical Local Explanation		
				AI Model	*Dataset*	*XAI Method*	*Usage*	Purpose
	Image	[89]	2020	VGG-16	Chest X-ray	Grad-CAM	S	An
		[90]	2021	ResNet-50	CT	LIME, SHAP	A	P
		[91]	2021	Inception-v3	Diagnosis of retinal images	GBP, SHAP	A	P
		[92]	2021	CNN	IVCM	Grad-CAM, guided Grad-CAM	S	An
		[93]	2021	EfficientNet	Chest X-ray images	Grad-CAM	S	An
		[94]	2021	CNN, LSTM	ISIC 2017 and 2018	Grad-CAM	S	An
		[95]	2021	VGG, ResNet, DenseNet	COVID-19 chest X-ray	Grad-CAM	S	An
		[96]	2021	VGG-16	Chest CT, X-ray image	Grad-CAM, Grad-CAM++, LRP	S	An
		[97]	2022	VGG-19	Oral images	Grad-CAM	S	An
		[98]	2022	COMiT-Net	ChestXray-14, CheXpert	Grad-CAM	S	An
		[99]	2022	CNN	LC25000, NSCLC	Grad-CAM, OS	A	P
		[100]	2022	DNN	CheXpert, MIMIC, NIH	CAM	S	An
		[101]	2023	DNN	COVID-QU, QaTa-Cov19	Uncertain-CAM	S	An
		[102]	2023	VGG-16	CTs	Grad-CAM	A	P
		[103]	2023	EfficientNet, DenseNet, ResNet	Tooth areas	Grad-CAM	A	P
		[104]	2023	ResNet-50	COVIDNet	LIME	A	P
		[105]	2023	ResNet152	KVASIR	Grad-CAM	S	An
Industrial engineering	Time series	[106]	2020	CNN	Machinery fault	LRP	A	P
		[107]	2020	RF	Bushings testbed	LIME	A	P
		[108]	2020	CNN, VAE	Ford Motor	CAM	A	P

Table 2. *Cont.*

Industrial Sector	Data Type	Ref.	Year	AI Model and Dataset		Practical Local Explanation		
				AI Model	*Dataset*	*XAI Method*	*Usage*	Purpose
		[109]	2020	Deep-SincNet	Motor currents	t-SNE, SincNet filters	A	P
		[110]	2020	CNN, LSTM, Bi-LSTM	C-MAPSS	SHAP	A	P
		[111]	2021	1D-CNN	Normal and fault conditions	FG-CAM	S	An
		[112]	2021	DNN	Prismatic cantilever steel beam	LIME, SHAP	A	P
		[113]	2021	TScatNet	CWRU, DDS	t-SNE	S	An
		[114]	2022	kNN, OCSVM, etc.	Bearing, Gearbox Fault	SHAP, Local-DIFFI	A	P
		[115]	2022	WaveletKernelNet	Bearing, Gearbox Fault	CWConv layer	S	An
		[116]	2023	SVM, kNN	Bearings	SHAP	A	P
	Image	[117]	2020	CNN	Bearings	GradCAM	A	P
		[118]	2020	CNN	Image fault diagnosis	CAM	A	P
Cybersecurity	Tabular	[119]	2020	SVM	NSL-KDD	SHAP	A	P
		[120]	2021	RF, XGBoost, Sequence Model	ISCX-URL2016, CICMalDroid 2020	SHAP, LIME	A	P
		[121]	2021	Autoencoder	CID-IDS2017	SHAP	A	P
		[122]	2021	DT, ANN	Private dataset	SHAP, LIME	A	P
		[123]	2022	DT	NF-BoT-IoT-v2, NF-ToN-IoT-v2	SHAP	A	P
		[124]	2022	DNN	UNSW-NB15	SHAP	A	P
		[125]	2023	CNN	ToN_IoT	SHAP	A	P
		[126]	2023	ANN	WUSTL-IIoT, NSL-KDD	TRUST, LIME	A	P
Smart agriculture	Tabular	[127]	2021	RF	Wheat, maize, olive groves	LIME	A	P
		[128]	2022	DT, RF	Maize crop yield	LIME	A	P
		[129]	2022	LSTM, Bi-LSTM, Bi-GRU-LSTM-CNN	ProductReview	SHAP, LIME	A	P
		[130]	2022	XGB, MLP, SVM	Crop Recommendation	SHAP, LIME	A	P
	Time series	[131]	2020	CNN	Meteorological, wheat yield data	RAM	A	P
		[132]	2022	LightGBM	Diverse physical agricultural	SHAP, LIME	A	P
		[133]	2023	GRU	Plant SSPs	ISM	S	An

Table 2. *Cont.*

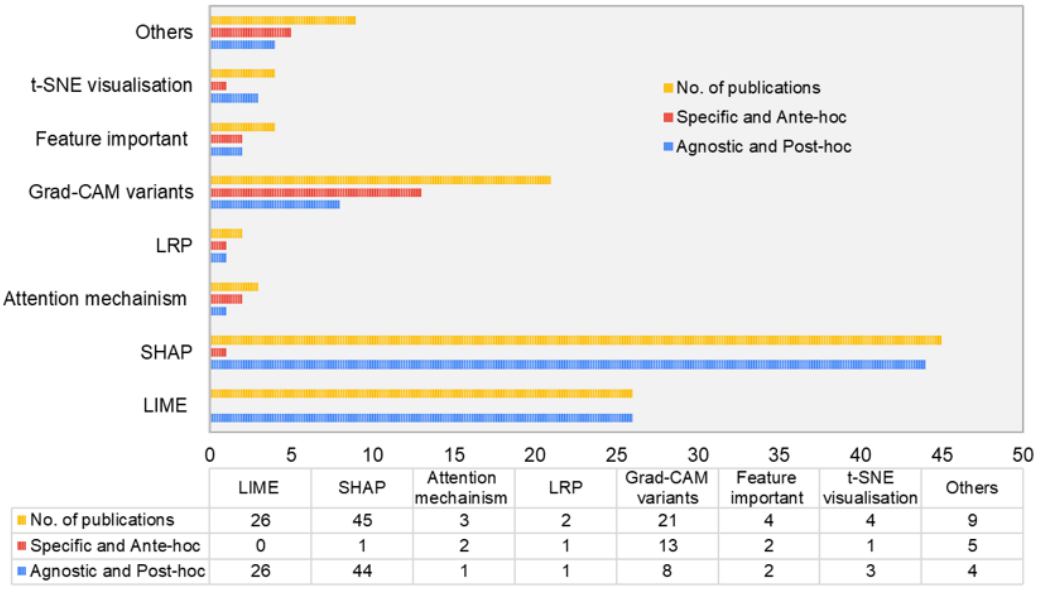
Industrial Sector	Data Type	Ref.	Year	AI Model and Dataset		Practical Local Explanation		
				AI Model	***Dataset***	***XAI Method***	***Usage***	**Purpose**
	Image	[134]	2021	ResNet-V2, VGG-19, VGG-16, Inception-V3	Diseased leaves of pearl millet	Grad-CAM	A	P
		[135]	2022	LightGBM	Agri-worker motion	ELI5, PDPbox, Skater	A	P
		[136]	2022	CNN	Fire and smoke	LIME, Grad-CAM++	A	P
	Text	[137]	2022	OAK4XAI	Graph database	AgriComO	S	An

A: agnostic, P: post hoc, S: specific, An: ante hoc.

3.1. Quantitative Analysis

3.1.1. Analysis Based on Usage and Purpose

Figure 3 presents the distribution of publications based on local explanation methods' usage and purpose for industrial AI applications, as derived from Table 2. The survey shows that the SHAP (Shapley additive explanations), LIME (local interpretable model-agnostic explanations), and various Grad-CAM (gradient-weighted class activation mapping) methods are the most commonly used local explanation techniques, accounting for 39.47%, 22.81%, and 18.42% of the studies, respectively. Other local explanation methods have been used less frequently. Regarding usage and purpose, most of the SHAP and LIME applications in the surveyed literature are agnostic and post hoc. At the same time, the specific and ante hoc of local explanations are primarily used in various Grad-CAM methods for industrial AI applications.



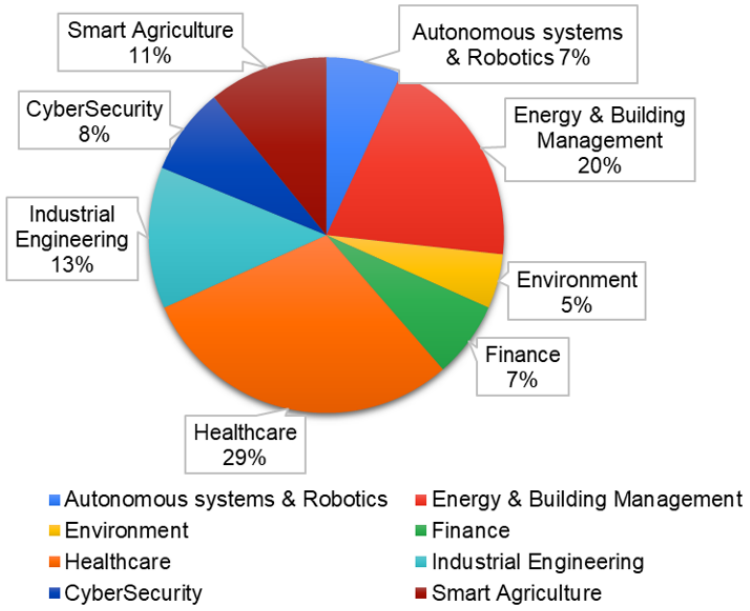
	LIME	SHAP	Attention mechainism	LRP	Grad-CAM variants	Feature important	t-SNE visualisation	Others
No. of publications	26	45	3	2	21	4	4	9
Specific and Ante-hoc	0	1	2	1	13	2	1	5
Agnostic and Post-hoc	26	44	1	1	8	2	3	4

Figure 3. Distribution of studies based on local explanation with usage and purpose for industrial AI applications.

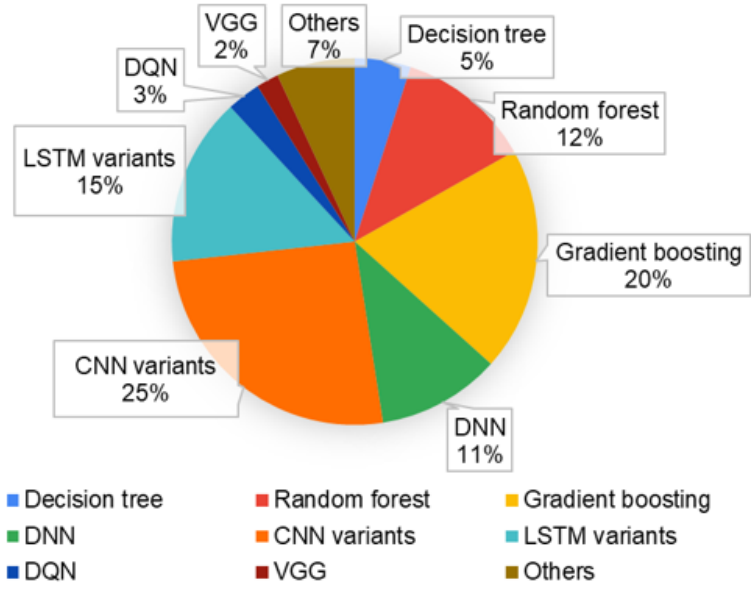
3.1.2. Analysis Based on Industrial Applications and AI Model Implementation

In this section, we present statistics on the distribution of publications on local explanation in the context of industrial AI applications and the AI models employed, based

on the findings in Table 2. Figure 4 illustrates the classification of the research papers we reviewed according to their industrial application domain. As depicted in Figure 4a, healthcare and energy and building management are the most frequently studied sectors, accounting for 29% and 20% of the surveyed XAI techniques, respectively. Conversely, cybersecurity, autonomous systems and robotics, finance, and environment constituted less than 10% of the surveyed techniques, suggesting that these domains have yet to see the widespread application of XAI methods.



(**a**) Based on AI industrial applications



(**b**) Based on AI models

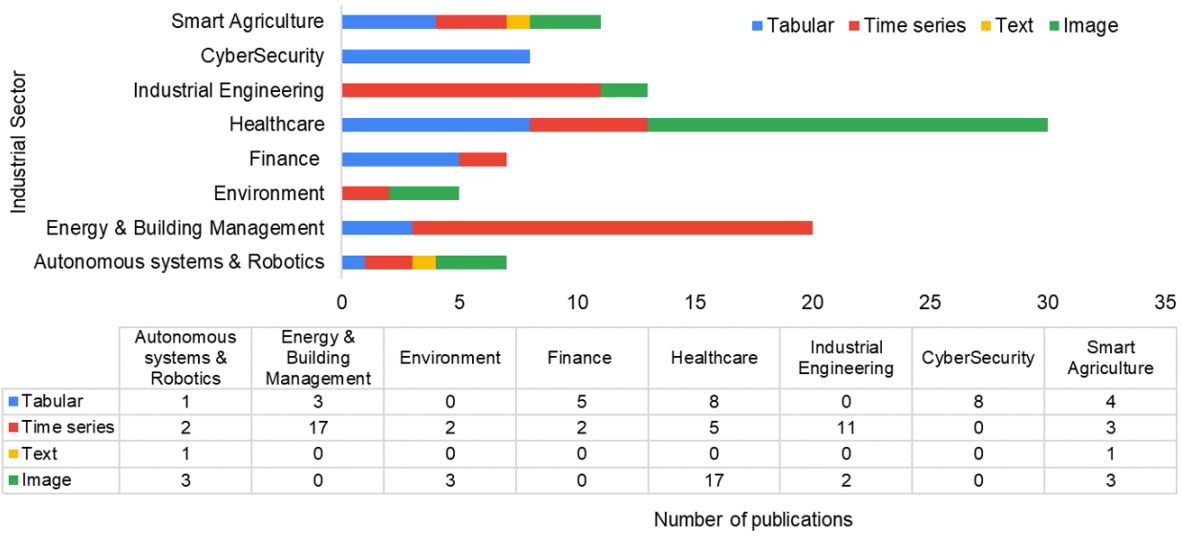
Figure 4. Distribution of studies based on local explanations for industrial applications and AI models.

Additionally, we analyzed the distribution of AI models used in the surveyed works. We found that most local explanation techniques were applied to various CNN models, accounting for 25% of the publications, followed by gradient boosting techniques, particularly XGBoost, at 20%. Decision tree, DQN (deep Q-network), and VGG (visual geometry group) models, among others, were used in less than 10% of the surveyed papers, indicating a lack

of widespread use of these models with XAI for other industrial applications. These results are presented in Figure 4b.

3.1.3. Analysis Based on Data Types

Based on the information presented in Table 2, we analyzed the distribution of publications on local explanation techniques used for different types of data in industrial applications. Figure 5 breaks down the studies based on data types, including tabular, time series, text, and image. The results show an increasing interest in interpretable ML for healthcare with image data (16.83%) and energy and building management with time-series data (16.83%). However, fewer applications of local explanation techniques were found for text data (1.98%), consistent with the findings for the entire industrial sector. It is worth noting that since this review covers publications up to 31 March 2023, the number of publications for 2023 is expected to exceed those of 2022.



	Autonomous systems & Robotics	Energy & Building Management	Environment	Finance	Healthcare	Industrial Engineering	CyberSecurity	Smart Agriculture
Tabular	1	3	0	5	8	0	8	4
Time series	2	17	2	2	5	11	0	3
Text	1	0	0	0	0	0	0	1
Image	3	0	3	0	17	2	0	3

Figure 5. Breakdown of publications based on local explanation for industrial applications, with data types.

3.2. Qualitative Analysis

This section discusses the effectiveness and limitations of local explanation techniques in industrial applications from the results of Table 2.

3.2.1. *Autonomous Systems and Robotics*

According to the survey results in Table 2, several publications have focused on XAI local explanations for autonomous systems and robotics.

Zhang et al., 2020 [37] employed the SHAP method and the probability of SHAP values to explain emergency control based on DRL in power systems, in the context of time-series data. Their study combined a post hoc XAI technique with agnostic usage. As an alternative, Renda et al. [38] proposed the FED-XAI idea, which suggests federated learning of XAI models for AI-pervasive 6G networks while proposing an ante hoc XAI strategy. The concept is anticipated to enhance the performance, intelligence, and trustworthiness of automated vehicle networking and improve user experience, fostering end-users' protection in network AI processes.

In the context of image data, He et al., 2021 [40] proposed a new DRL scheme for model explainability using an agnostic and post hoc XAI approach. They developed a saliency map generation method that merges CAM and SHAP values to create visual and textual action explanations for non-expert users. The model can be refined and enhanced based on the explanation generated. On the other hand, Zhang et al., 2022 [41] employed a distinctive and ante hoc XAI strategy to improve the performance and transparency

of autonomous driving systems' decision making by offering multi-modal explanations, particularly when interacting with pedestrians.

Cui et al., 2022 [42] used SHAP and random forest (RF) approaches regarding tabular data to encourage transparent DRL-based decision making. They adopted an agnostic and post hoc XAI approach. Regarding text datam and employing an agnostic and post hoc approach, Nahata et al., 2021 [43] developed interpretable machine learning models to evaluate and forecast collision risk based on various sensor data features.

3.2.2. *Energy and Building Management*

As presented in Table 2, our survey results show that XAI local explanation research has concentrated on time-series and tabular data with energy and building management applications.

Much research has used an agnostic and post hoc design for time-series data and two main approaches, SHAP and LIME. For example, Arjunan et al. [45] proposed an approach that improves the current Energy Star calculation method and employs extra model output processing to illuminate how a building can reach a particular score. Li et al. [57] used SHAP to analyze and interpret the XGBoost-based grid search power load forecasting model. While Zdravkovi et al. [52] presented an AI-assisted control of district heating systems (DHS) for effective heat distribution. Kim et al.'s explanation of the energy demand forecast model employing feature importance and attention methods was published in [44]. To support and elucidate the forecasts, Grzeszczyk et al. [56] suggested a strategy based on the LIME method. Both Wenninger et al. [59] and Moon et al. [58] used distinct and ante hoc XAI techniques with feature-important analysis. While Wenninger et al. proposed the QLattice for estimating the final energy performance of residential buildings, Moon et al. evaluated the most crucial aspects of electrical load forecasting. However, several of these studies' weaknesses must be addressed in future research. For example, the generalizability of the procedures and outcomes only relates to the input data tested. The explanations may also be difficult for average consumers to comprehend [45]. Future studies could concentrate on developing technical research and apps to provide more intuitive explanations to general consumers [57]. Furthermore, robust model interpretability can sometimes be difficult, and future studies should investigate alternate brain models and ways of explaining black-box models [59].

The application of various XAI techniques to tabular data, with an agnostic and post hoc approach, has been explored. Specifically, LIME and SHAP have been investigated for generating local explanations. Examples include Srinivasan's SHAP-based XAI-FDD [61], Sim's SHAP-based analysis of input variables for energy consumption forecasting [63], and Wastensteiner's LIME and SHAP-based visualizations for personalized feedback on electricity consumption time-series data [62]. Srinivasan et al. [61] proposed XAI-FDD (explainable artificial intelligence-fault detection and diagnosis), which uses explanations generated for each data instance to detect incipient faults. By combining human expertise with the explanations provided by the XAI system, this approach could potentially improve accuracy in classification. The XAI-FDD may examine air-handling systems, renewable energy sources, and other building energy components. Sim et al. [63] used SHAP-based XAI to examine how input variables affect energy use forecasting. Their research divided the input variables into three categories—strong, ambiguous, and weak—providing insights into which variables had the most significant impact. However, they only analyzed certain targets, input factors, and predictive models.

Future studies may need to consider a more comprehensive range of socioeconomic variables for different types of buildings. Wastensteiner et al. [62] developed five visualizations using ML and XAI methods, to provide personalized feedback on electricity consumption time-series data, incorporating domain-specific knowledge. However, their approach only considered SHAP-based XAI visualizations, and future research could explore other XAI methods, such as LIME, for additional visualizations.

3.2.3. *Environment*

Researchers in the field of environmental science have focused on XAI techniques applied to time-series and image datasets, as summarized in Table 2.

Researchers have seen the application of SHAP and LIME approaches with agnostic and post hoc designs in time-series data, as shown in Table 2. Graham et al. [64] employed deep neural networks (DNN) and XAI algorithms on the "Dynomics" platform to explore patterns in transcriptional data on a genome-scale and identify the genes contributing to these patterns. Gao et al. [65] used LIME to interpret the DFNN (dynamic fuzzy neural networks) models to assess the danger levels of a building in Cameron County, Louisiana, in response to a fictitious impending hurricane, with updated weather predictions. Despite these studies' achievements, there are still limitations, and future research can expand the analytical targets and models employed in these investigations.

The use of SHAP and LIME techniques with agnostic and post hoc designs on image data has been explored by various research groups, as described in [66–68]. Masahiro et al. [66] discuss the importance of XAI in ecological modeling and list tools that can be employed to understand complex model behavior at different scales. Integrating XAI with ecological and biogeographical knowledge can enhance the accuracy of machine learning models. Kim et al. [68] used the XGBoost model and SHAP to analyze urban expansion, where land-cover characteristics were identified as the primary factor, followed by topographic attributes. However, the XGBoost-SHAP model's accuracy, such as using AutoML algorithms, should be assessed compared to other XAI methods. Dikshit et al. [67] built and compared an XAI model to physical-based models using an explainable deep learning technique. Their study examined how predictors interacted locally for distinct drought conditions and timeframes, providing insight into how the model produced specific findings at different spatiotemporal intervals. Future research should look at SHAP plots for long-term forecasting and other additive SHAP properties.

3.2.4. *Finance*

In the finance sector, the majority of researchers have focused on using LIME and SHAP methods with an agnostic and post hoc approach for tabular and time-series datasets (as shown in Table 2). Some researchers, including Gramegna et al. [69], Benhamou et al. [70], Babaei et al. [72], and de Lange et al. [73], have applied the SHAP method on tabular datasets, while Kumar et al. [75] and Bussmann et al. [74] have developed their SHAP methods on time-series datasets. Additionally, Gite et al. [71] have utilized the LIME method for a tabular dataset.

Several researchers have focused on applying explainable AI models to tabular datasets in finance. For example, Gramegna et al. [69] proposed a model that uses Shapley values and similarity clustering to explain why customers either buy or abandon non-life insurance coverage. In contrast, Benhamou et al. [70] utilized SHAP to identify important variables in stock market crises and provide local explanations of the probability of a crisis at each date. Babaei et al. [72] developed a Shapley-based model to predict the expected return of small and medium-sized enterprises based on their credit risk and expected profitability. Similarly, de Lange et al. [73] combined SHAP with a LightGBM (light gradient-boosting machine) model to interpret explanatory variables affecting credit scoring, which outperformed the bank's logistic regression model. Additionally, Gite et al. [71] proposed a model that uses long short-term memory (LSTM) and efficient machine learning techniques to accurately predict stock prices based on user sentiments derived from news headlines. They suggest further research directions to include automated prediction of financial news headlines and adding emotion-based GIFs to enhance the model's appeal. This model can be used as a decision maker for algorithmic trading. Future research may explore applying these methods to other situations in the finance sector, such as underwriting and claims management.

Kumar et al. [75] have suggested using SHAP with a popular deep reinforcement learning architecture, DQN, to explain an agent's actions in financial stock trading, in time-

series data. Additionally, they advised expanding the method to continuous action space and using DRL models, such as deep deterministic policy gradient (DDPG) and advantage actor–critic (A2C), to improve the explanation by adding more technical indicators as features. In another study, Bussmann et al. [74] created an explainable AI model for fintech risk management, particularly for assessing the risks associated with peer-to-peer lending platforms' credit borrowing. They used Shapley values to interpret AI predictions based on underlying explanatory variables. Future research could be conducted on enhancing prediction understanding by clustering the Shapley values.

3.2.5. *Healthcare*

Table 2 shows that in the healthcare sector, most research studies have focused on applying Grad-CAM, SHAP, and LIME methods.

Numerous studies have utilized the SHAP method to interpret tabular data. For example, while Kim et al. [81] developed an interpretable machine learning model for predicting early neurological deterioration related to stroke, Beebe et al. [80] developed an approach that captures non-linear feature effects in personalized risk predictions. Rashed et al. [82] developed models for diagnosing chronic kidney disease (CKD) that identified important features consistent with clinical knowledge. Zhang et al. [83] also used SHAP to develop machine learning models for CKD diagnosis that provided physicians with richer information on feature importance to improve decision making. These studies highlight the usefulness of SHAP in developing interpretable models for medical applications and the potential to improve clinical decision making.

Three XAI approaches were used for time-series data: attention, layer-wise relevance propagation (LRP), and Grad-CAM. Mousavi et al. [84] introduced HAN-ECG. This bidirectional-recurrent-neural network-based technique employs three attention mechanism levels to detect atrial fibrillation (AF) patterns in an ECG (electrocardiogram). However, the success of the strategy is dependent on the preprocessing stage. This method should be applied to different ECG leads and arrhythmias in the future to extract novel patterns that may be beneficial in detecting arrhythmias. An interpretability method was proposed by Filtjens et al. in their paper [87] to explain DNN decisions for recognizing the movement preceding the freezing of gait (FOG) in Parkinson's disease (PD). The recommended pipeline can help physicians to explain DNN conclusions and let ML experts check the generalizability of their models. This pipeline could be used to start FOG treatment pipelines. In these circumstances, the interpreters can promote the provision of external stimuli and evaluate the effectiveness of the intervention by picturing diminished significance for FOG. Dutt et al.'s SleepXAI [88], an explainable unified method for multi-class sleep stage categorization utilizing modified Grad-CAM, was first proposed in 2012. This technique explains multi-class categorization of univariate EEG (electroencephalogram) signals while significantly increasing overall sleep stage classification accuracy. During the anticipated sleep stage, SleepXAI generates a heatmap depicting relevant features learned from univariate EEG data. It enables sleep specialists to link observed characteristics to traditional manual sleep grading techniques, boosting confidence in opaque systems with justifications.

Various studies have applied the Grad-CAM method for interpreting image data, focusing on medical imaging. For example, Figueroa et al. [97] proposed a two-stage training process to enhance classification performance and guide the CNN's attention to lesion areas. To detect COVID-19 on chest X-ray images, Chetoui et al. [93] developed DeepCCXR (deep COVID-19 CXR detection), while Barata et al. proposed a hierarchical neural network with spatial attention modules for skin cancer diagnosis. To help COVID-19 patients be triaged more quickly, Singh et al. [95] proposed an AI-based solution; Malhotra et al. [98] presented COMiT-Net; and Oztekin et al. [103] proposed an explainable deep learning model. These investigations show that the model's decision-making process is now more accurate and comprehensible and can be directed to focus on particular areas of interest. However, limitations include the need for more accurate labeled datasets,

limited data availability, and extending the models to process other types of diseases with radiography images.

3.2.6. *Industrial Engineering*

In industrial engineering, time-series and image datasets are commonly used for local explanation methods such as Grad-CAM and SHAP, as seen in Table 2.

Brusa et al. [116] and Hong et al. [110] have both employed Grad-CAM in their ML models to interpret their predictions for time-series data. Hong et al. employed XAI algorithms to find the most important sensors in predicting the remaining useful life of a turbofan engine, and Brusa et al. used SHAP values to diagnose flaws in industrial bearings. Other interpretability approaches, such as LRP and LIME, were used by Grezmak et al. [106] and by Serradilla et al. [107]. The effectiveness of a CNN trained on pictures of the time-frequency spectra of vibration signals measured on an induction motor was evaluated by Grezmak et al. using LRP. Serradilla et al., in contrast, employed XAI methods to direct the development, improvement, and interpretation of a model for estimating the remaining life during fatigue testing, based on condition and setting variables.

For image data, Grad-CAM and CAM (class activation mapping) are common agnostic and post hoc explanation methods used in the industrial engineering sector. Chen et al. [117] used Grad-CAM to interpret the predictions of their CNN (convolutional neural network) model for bearing fault classification using time-frequency spectra. Similarly, Sun et al. [118] employed CAM to diagnose faults and recognize images in the cantilever beam case. Both studies demonstrated the feasibility of using explainable deep learning to diagnose faulty components from images. Future research directions may include testing different equipment settings to determine the minimum requirements for successfully implementing these techniques.

3.2.7. *Cybersecurity*

Local explanations based on XAI techniques in cybersecurity are mainly applied to tabular data. This is evident from the industrial sector analysis presented in Table 2, where most methods employed are based on SHAP.

An ML-based intrusion detection system (IDS) employing an ensemble trees technique was proposed in work by Le et al. [123]. The method employs decision trees and random forest classifiers, and it does not require much computational power to train the models. The SHAP approach was utilized to explain and interpret the models' classification conclusions, allowing cybersecurity specialists to optimize and evaluate the validity of their judgments swiftly. Karn et al. [122] developed an ML-based detection strategy for anomalous pods in a Kubernetes cluster in a different study. To identify and explain crypto-mining applications, the system uses auto-encoding-based techniques for LSTM models, SHAP, LIME, and LIME. The system's explainability is critical for system administrators to grasp system-level rationales for supporting disruptive administrative decisions. Wang et al. [119] also used SHAP to improve the interpretation of IDSs by integrating local and global explanations. Their proposed architecture can improve IDS transparency, allowing cybersecurity experts to make better decisions and optimize IDS structure. Finally, Alenezi et al. [120] employed RFC (random forest classifier), XGBoost, and the sequential algorithm to study two big cybersecurity datasets and used three SHAP approaches to explain the feature contributions. The study emphasizes the need to understand the value of data in order to increase the explanatory capabilities of cybersecurity threats data using SHAP methodologies, which can lead to future data collection operations in cybersecurity or other fields.

The benefits of employing SHAP for XAI in cybersecurity include improved model transparency and interpretability, the ability for cybersecurity specialists to make better decisions, and the optimization of model structures. However, drawbacks include the method's intricacy and the possibility of misleading explanations if the model is not thoroughly understood.

3.2.8. *Smart Agriculture*

The smart agriculture sector is explained locally using various datasets, including tabular, time-series, and text data, as illustrated in Table 2.

Using publicly available tabular data, Ryo et al. [128] used XAI and interpretable ML to study the influence of no-tillage on agricultural yield compared to conventional tillage. The authors assessed the importance of factors for prediction, variable interactions, and the relationship between relevant variables and the response variable. Adak et al. [129] used sentiment analysis to assess customer evaluations in the food delivery services (FDSs) domain, and they justified their predictions using SHAP and LIME. Viana et al. [127] proposed a machine learning model to discover the factors influencing agricultural land usage at the regional level for wheat, maize, and olive grove plantings. Using a model-agnostic methodology, they presented global and local interpretations of the significant elements. XAI technologies were used by Cartolano et al. [130] on the 'Crop Recommendation' dataset to make ML models clear and trustworthy. Their research focused on sensitivity analysis, comparing what the models discovered to what farmers and agronomists already knew. Future research could look into other XAI methodologies and visualization techniques in sectors as diverse as computational creativity and emotion recognition.

Several studies have looked into the usage of deep neural networks in time-series data research. For example, Wolanin et al. [131] used a deep neural network to predict wheat yield in the Indian wheat belt. They used regression activation maps (RAMs) to improve interpretability to show the model's learned features and yield drivers. Similarly, Kawakura et al. [132] created an XAI-based technique for agri-directors to train agri-workers by analyzing varied data and merging agri-informatics, statistics, and human dynamics. Li et al. [133] created the ExamPle model, which employed a Siamese network and multi-view representation to forecast plants' small secreted peptides (SSPs) and revealed the SSPs' sequential pattern. Additionally, Kundu et al. [134] presented the AIDCC (automatic and intelligent data collector and classifier) framework for automating the collection of imaging and parametric datasets from farms producing pearl millet, disease prediction, and feature visualization using deep learning and IoT.

Finally, Apostolopoulos et al. [136] showed that the Xception network outperforms other CNNs in recognizing suspicious situations in various photos. It improved its post hoc explainability by using the Grad-CAM++ and LIME algorithms. They recommended that future studies look at various methodologies, such as fuzzy logic and fuzzy cognitive maps (FCM), to examine timely fire and smoke incident detection.

Ngo et al. [137] have presented OAK4XAI (model towards out-of-box explainable artificial intelligence), an XAI system for text data analysis that combines domain knowledge semantics via an ontology and knowledge map model. To describe the knowledge mined in agriculture, they developed the agriculture computer ontology (AgriComO), a well-structured framework suitable for agriculture and computer domains. In future research, the authors intend to create an explanation interface as a service for user engagement and to expand the model by integrating multiple ML algorithms for prediction utilizing explainable methodologies.

3.3. Current Challenges and Recommendations of Local Explanation for Industrial Applications

Based on our study and analysis of the surveyed papers, we identify several challenges and derive possible recommendations to solve those issues.

3.3.1. Challenges

Local explanations are essential to industrial AI applications but face two significant challenges. The first challenge is that local explanations require computational resources and only provide a narrow view of the model's behavior. The second issue is the trade-off between interoperability and performance accuracy, as a human-centric strategy is still required.

Local explanation techniques aim to provide insight into the decision-making processes of complex ML models by highlighting the contribution of individual features to the final prediction for a specific input instance. Some of the popular local explanation techniques used in industrial applications include LIME [138], SHAP [139], and Anchors [140]. These techniques have been applied to a wide range of applications such as credit risk assessment [73], image classification [141], and text classification [142]. These techniques provide transparency, increase trustworthiness, and can improve model performance. However, the effectiveness of these techniques may vary depending on the type of model, dataset, and task [138,139]. Several studies have demonstrated the effectiveness of local explanation techniques in industrial applications. For example, Selvaraju et al. [141] used LIME to explain the predictions of a deep neural network for image classification and showed that the technique improved the model's performance and helped to detect biases in the dataset. De Lange et al. [73] used SHAP to interpret the predictions of a credit risk assessment model and found that the technique increased the transparency and trustworthiness of the model. Different local explanation techniques have strengths and weaknesses in explaining complex ML and DL models in industrial applications. Several studies have compared the performance of different techniques, such as LIME, SHAP, and Anchors, on various tasks and datasets [73,139,140]. These studies have shown that the effectiveness of these techniques may depend on the type of model, dataset, and task and that no single technique is universally superior.

XAI has been proposed as a way to address the lack of transparency and limitations of AI's usage in critical fields [14,143]. The goal of XAI, according to [144], is to provide techniques that let end users comprehend, believe in, and control the rising number of AI systems. For typical learning models with high-dimensional datasets, interpretability and performance accuracy frequently trade off. Models with high architectural complexity, such as DL and random forests, often provide better performance accuracy, but are called "black-box models" because they lack transparency and explainability. In contrast, white-box or glass-box models, such as linear, graphical, and decision trees, provide transparent and understandable results. Still, their performance may be subpar [145]. However, recent research by [146] challenges the assumption that relatively simple ML models are easy to interpret, as interpretability depends on the context of use, users involved, and complexity. Therefore, Schoonderwoerd et al. [147] argue for a more human-centric approach to XAI, which prioritizes user needs over the explainability of the ML model. As a result, there may be situations where a high-performance model is preferred over an explainable one, or vice versa [148].

3.3.2. Recommendations

Section 3.3.1 discusses the challenges of interpreting complex models and the limitations of local explanations in capturing global model behavior, which can hinder the adoption and trust in AI models. To address these challenges and improve the interpretability and transparency of industrial AI applications, we suggest the following remedial measures based on our analysis and knowledge:

- *Data quality assurance*: Ensuring high-quality datasets is essential to minimize the impact of challenges associated with local explanation techniques. By performing data cleaning, normalization, and preprocessing, we can ensure that the datasets are high quality and minimize the risks of producing unreliable and inaccurate explanations [149].
- *Model validation*: Thorough model validation and testing should be carried out to ensure that the model is accurate and reliable [145,146]. Our study suggests involving end-users in the interpretability process, prioritizing the evaluation of the model's interpretability according to the context of use, and providing clear explanations of the model's limitations and assumptions to enhance transparency and trust in the model [150].

- *Appropriate choice of explanation techniques*: The efficacy of local explanation strategies varies according to model type, dataset, and task. Thus, choosing the most appropriate technique for the application is crucial. Researchers can also develop hybrid models that combine black-box and white-box models' strengths to achieve high performance and interpretability [148].
- *Human-in-the-loop*: A human-in-the-loop approach can improve the quality of local explanations and enhance trust in the model. By including human experts in the decision-making process, we can ensure that the local explanations are relevant and accurate for the intended use case [151].

Different local explanation techniques have strengths and weaknesses in explaining complex ML and DL models in industrial applications. It is, therefore, essential to select the appropriate technique, ensure data quality, validate models, and include human experts in the decision-making process to minimize the impact of the challenges associated with local explanation techniques. To support our arguments, we draw on the literature, which has shown that the effectiveness of local explanation techniques may vary depending on the model type, dataset, and task [152]. Furthermore, recent research has challenged the assumption that relatively simple ML models are easy to interpret, as interpretability depends on the context of use, the users involved, and complexity [153]. Therefore, our study emphasizes the need for a more human-centric approach to XAI, prioritizing user needs over the explainability of the ML model.

In conclusion, our paper provides a roadmap for using local explanation techniques in industrial AI applications, addressing the challenges associated with their use and enhancing the models' transparency, interpretability, and trust. However, we acknowledge that our framework is not exhaustive and that further research is needed to develop more robust and effective approaches to XAI in industrial settings.

4. Discussion of Gaps and Limitations in the Current Literature

Local explanation techniques have gained significant attention due to their potential to increase transparency and interpretability in complex ML and DL models used in industrial applications. While these techniques have shown promise, several gaps and limitations must be addressed.

4.1. Discussion of Survey Results and Gaps

Section 3 presents our survey results, allowing us to answer the four main research questions outlined in Section 2. Analyzing the data collected gave us insights into using local explanation techniques in industrial applications and their distribution, benefits, limitations, and effectiveness.

Q1: What local explanation techniques are used in industrial applications? SHAP, LIME, and Grad-CAM variations are among the most extensively employed local explanation approaches in industrial applications, according to the survey data presented in Section 3.1.1. SHAP assigns an importance score to each feature based on the Shapley value from cooperative game theory. LIME is a model-agnostic explanation method that learns an interpretable model locally around the prediction to explain the predictions of any black-box model. Grad-CAM variations in deep neural networks use gradient information to depict and explain the features contributing to a certain prediction. In numerous industrial areas, these strategies have proven beneficial in delivering local explanations and boosting transparency and trust in AI models.

Q2: How widespread are practical industrial applications of local explanation techniques? The practical industrial applications of local explanation techniques have gained significant traction in recent years. According to the survey results in Section 3.1.2, local explanation techniques are widely used in various fields, including healthcare, energy and building management, and industrial engineering. Among these fields, healthcare is one of the most prominent sectors in adopting local explanation techniques. These techniques are used extensively for clinical decision making, disease prediction, and patient

monitoring. In healthcare, local explanation techniques play a vital role in helping medical professionals to understand the reasoning behind the predictions of medical diagnosis models, improving the accuracy of diagnosis, and building trust in the models. Energy and building management is another sector that has embraced local explanation techniques to optimize energy consumption and improve building performance. These techniques are used to interpret the predictions of energy consumption models, identify the factors affecting energy consumption, and recommend energy-efficient practices. Similarly, industrial engineering uses local explanation techniques for predictive maintenance and quality control. These techniques can help to identify potential machine failures, predict maintenance needs, and improve product quality. The survey results demonstrate that local explanation techniques are increasingly prevalent in various industrial sectors. Healthcare, energy and building management, and industrial engineering are among the leading fields in adopting these techniques.

Q3: What are the benefits and limitations of local explanation techniques for industrial applications? The ability to explain specific predictions of complex models, identify the essential components that lead to the prediction, find biases and errors, and establish trust in the models are all advantages of local explanation techniques. Local explanations are also useful in increasing AI systems' transparency, accountability, and fairness. On the other hand, as outlined in Section 3.3.1, one of the major limitations of local explanation techniques is the lack of standardization in evaluating and comparing them. Although numerous studies have proposed various techniques, there is currently no agreed-upon standard for evaluating their performance, which makes it difficult to compare different methods. This limitation poses a challenge in determining the most effective technique for specific applications. To address this issue, there is a need for standard datasets, metrics, and benchmarks, to enable a fair and accurate comparison of different techniques in different applications [14]. Another important gap is the limited understanding of the robustness and generalizability of local explanation techniques across different datasets and applications. While many studies have evaluated the effectiveness of these techniques on specific tasks and datasets, it is unclear whether the findings can be generalized to other tasks and datasets. The effectiveness of these techniques may vary depending on the type of data and application, making it challenging to determine the most appropriate technique to use in different scenarios. Therefore, there is a need for a more comprehensive and diverse evaluation of these techniques across different applications and datasets [154]. Additionally, local explanation techniques may have ethical and social implications that must be addressed. For example, these techniques may reveal sensitive information about individuals or groups, and there is a risk of unintended consequences if these explanations are not carefully designed and implemented. Therefore, it is important to consider these techniques' ethical and social implications and to develop guidelines for their responsible use [155]. Finally, after analyzing our survey results and statistical distributions, it is evident that the development of local explanation techniques is imbalanced across different industrial application fields. In summary, local explanation techniques can become more useful and effective in real-world settings by addressing these limitations and gaps in the current literature.

Q4: How to build effective local explanation techniques in practical settings? The effectiveness of local explanation techniques in practical settings can vary depending on the specific use case and the data quality and models, as mentioned in Section 3.3.2. While these techniques can provide valuable insights into the behavior of complex models, their effectiveness may be limited in some cases. For example, if the model is too complex or the dataset is too noisy, the local explanations may not be reliable or accurate. However, studies have shown that when combined with other interpretability methods, such as global explanations or feature importance analysis, local explanations can provide a more comprehensive understanding of the model behavior and improve the trustworthiness of the models. Additionally, the human-centered approach is crucial in building effective local explanations in industrial applications. By involving domain experts and end-users in

the explanation process, the explanations can be tailored to meet the stakeholders' specific needs and preferences, increasing their effectiveness and acceptance.

4.2. Analysis of Potential Biases and Limitations of the Study

One of the potential biases and limitations of this study is the focus on a limited number of local explanation techniques and industrial applications. While we have discussed some popular techniques and applications, this study has not covered many other techniques and applications. Moreover, the effectiveness of these techniques may vary depending on the specific task, dataset, and model, and our findings may not be generalizable to other contexts.

Another limitation is the reliance on the existing literature and case studies, which may be subject to publication bias and other forms of bias. Moreover, many of the studies that we have reviewed have used different evaluation metrics and benchmarks, which may make it difficult to compare their findings. Future studies should consider using standard metrics and benchmarks to enable better comparison and evaluation of different local explanation techniques.

5. Conclusions and Future Directions

In conclusion, this literature review has provided insights into the industrial applications of local explanation techniques for ML models. We have discussed the benefits and limitations of these techniques and highlighted the challenges in their practical use.

Our findings imply that local explanation strategies can help to improve the interpretability and transparency of black-box models in various industrial settings, including banking, healthcare, and manufacturing. However, limitations and roadblocks remain, such as computational costs and trade-offs between model accuracy and interpretability. Consistent evaluation measures, datasets, and benchmarks are required to address these issues to enable fair and reliable comparisons of diverse methodologies. Furthermore, the human-centered design of local explanations should be considered, as the ultimate goal is to enable human users to understand and trust AI systems.

To advance the field, we recommend that future research focuses on developing more effective and efficient local explanation techniques that can be applied to large-scale datasets and complex models. Additionally, ethical considerations must be addressed to ensure the responsible use of these techniques and minimize potential harm to stakeholders.

It is crucial to highlight limitations to this literature evaluation, such as the study scope and inherent biases in the selection criteria and methods. Nonetheless, this review contributes significantly to the field and can be a resource for researchers and practitioners interested in XAI and industrial applications.

To summarize, local explanation techniques can potentially increase the transparency and interpretability of complex ML and DL models in industrial settings. Further research, however, is required to address the challenges and limitations of these techniques and ensure their responsible use.

Author Contributions: Conceptualization, T.-T.-H.L.; methodology, T.-T.-H.L. and A.T.P.; software, T.-T.-H.L. and Y.E.O.; validation, T.-T.-H.L. and H.K. (Howon Kim); formal analysis, T.-T.-H.L., Y.E.O. and A.T.P.; investigation, H.K. (Howon Kim); resources, T.-T.-H.L. and A.T.P.; data curation, T.-T.-H.L., Y.E.O. and H.K. (Hyoeun Kang); writing—original draft preparation, T.-T.-H.L., Y.E.O., A.T.P. and H.K. (Hyoeun Kang); writing—review and editing, T.-T.-H.L. and H.K. (Howon Kim); visualization, T.-T.-H.L. and A.T.P.; supervision, T.-T.-H.L. and H.K. (Howon Kim); project administration, H.K. (Howon Kim); funding acquisition, H.K. (Howon Kim). All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01797) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-00903, Development of Physical Channel Vulnerability-based Attacks and its Countermeasures for Reliable On-Device Deep Learning Accelerator Design, 50%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
XAI	Explainable artificial intelligence
ML	Machine learning
DL	Deep learning
DQN	Deep Q-network
VGG	Visual geometry group
SHAP	Shapley additive explanations
LIME	Local interpretable model-agnostic explanations
Grad-CAM	Gradient-weighted class activation mapping
RF	Random forest
DHS	District heating systems
XAI-FDD	Explainable artificial intelligence-fault detection and diagnosis
DNN	Deep neural networks
DFNN	Dynamic fuzzy neural networks
LightGBM	Light gradient-boosting machine
LSTM	Long short-term memory
DDPG	Deep deterministic policy gradient
A2C	Advantage actor–critic
LRP	Layer-wise relevance propagation
AF	Atrial fibrillation
ECG	Electrocardiogram
FOG	Freezing of gait
PD	Parkinson's disease
EEG	Electroencephalogram
Deep CCXR	Deep COVID-19 CXR detection
CAM	Class activation mapping
RFC	Random forest classifier
CNN	Convolutional neural network
FDSs	Food delivery services
RAMs	Regression activation maps
SSPs	Siamese network and multi-view representation to forecast plant small secreted peptides
AIDCC	Automatic and intelligent data collector and classifier
FCM	Fuzzy cognitive maps
OAK4XAI	Model towards out-of-box explainable artificial intelligence
AgriComO	Agriculture computer ontology

References

1. Alex, D.T.; Hao, Y.; Armin, H.A.; Arun, D.; Lide, D.; Paul, R. Patient Facial Emotion Recognition and Sentiment Analysis Using Secure Cloud With Hardware Acceleration. In *Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications*; University of Texas at San Antonio: San Antonio, TX, USA, 2018; pp. 61–89.
2. Lee, S.M.; Seo, J.B.; Yun, J.; Cho, Y.-H.; Vogel-Claussen, J.; Schiebler, M.L.; Gefter, W.B.; Van Beek, E.J.; Goo, J.M.; Lee, K.S.; et al. Deep Learning Applications in Chest Radiography and Computed Tomography. *J. Thorac. Imaging* **2019**, *34*, 75–85. [CrossRef]
3. Chen, R.; Yang, L.; Goodison, S.; Sun, Y. Deep-learning Approach to Identifying Cancer Subtypes Using High-dimensional Genomic Data. *Bioinformatics* **2020**, *36*, 1476–1483. [CrossRef] [PubMed]
4. Byanjankar, A.; Heikkila, M.; Mezei, J. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. In Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 7–10 December 2015; pp. 719–725. [CrossRef]
5. Chen, Y.-Q.; Zhang, J.; Ng, W.W.Y. Loan Default Prediction Using Diversified Sensitivity Undersampling. In Proceedings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 15–18 July 2018; pp. 240–245. [CrossRef]
6. Zhang, Z.; Neill, D.B. Identifying Significant Predictive Bias in Classifiers. *arXiv* **2016**, arXiv:1611.08292. Available online: http://arxiv.org/abs/1611.08292 (accessed on 20 February 2023).
7. Hester, N.; Gray, K. For Black men, Being Tall Increases Threat Stereotyping and Police Stops. *Proc. Nat. Acad. Sci. USA* **2018**, *115*, 2711–2715. [CrossRef]
8. Parra, G.D.L.T.; Rad, P.; Choo, K.-K.R.; Beebe, N. Detecting Internet of Things Attacks Using Distributed Deep Learning. *J. Netw. Comput. Appl.* **2020**, *163*, 102662. [CrossRef]
9. Chacon, H.; Silva, S.; Rad, P. Deep Learning Poison Data Attack Detection. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; pp. 971–978. [CrossRef]
10. Dam, H.K.; Tran, T.; Ghose, A. Explainable Software Analytics. In Proceedings of the ICSE-NIER '18: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, Gothenburg Sweden, 27 May–3 June 2018; pp. 53–56. [CrossRef]
11. Scott, A.C.; Clancey, W.J.; Davis, R.; Shortliffe, E.H. *Explanation Capabilities of Production-Based Consultation Systems*; Technical Report; Stanford University: Stanford, CA, USA, 1977.
12. Swartout, W.R. Explaining and Justifying Expert Consulting Programs. In *Computer-Assisted Medical Decision Making. Computers and Medicine*; Reggia, J.A., Tuhrim, S., Eds.; Springer, New York, NY, USA, 1985. [CrossRef]
13. Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harv. J. Law Technol.* **2018** , *31*, 842–861. Available online: https://ssrn.com/abstract=3063289 (accessed on 20 February 2023). [CrossRef]
14. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
15. Omeiza, D.; Webb, H.; Jirotka, M.; Kunze, L. Explanations in Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 10142–10162. [CrossRef]
16. Wang, S.; Atif Qureshi, M.; Miralles-Pechuán, L.; Reddy Gadekallu, T.; Liyanage, M. Explainable AI for B5G/6G: Technical Aspects, Use Cases, and Research Challenges. *arXiv* **2021**, arXiv:2112.04698. https://doi.org/arxiv-2112.04698.
17. 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. https://doi.org/10.48550/arXiv.2112.11561.
18. Senevirathna, T.; Salazar, Z.; La V.H.; Marchal, S.; Siniarski, B.; Liyanage, M.; Wang, S. A Survey on XAI for Beyond 5G Security: Technical Aspects, Use Cases, Challenges and Research Directions. *arXiv* **2022**, arXiv:2204.12822. https://doi.org/10.48550/arXiv.2204.12822.
19. Sakai, T.; Nagai, T. Explainable Autonomous Robots: A Survey and Perspective. *Adv. Robot.* **2022**, *36*, 219–238. [CrossRef]
20. Emaminejad, N.; Akhavian, R. Trustworthy AI and Robotics: Implications for the AEC Industry. *Autom. Constr.* **2022**, *139*, 104298. [CrossRef]
21. Alimonda, N.; Guidotto, L.; Malandri, L.; Mercorio, F.; Mezzanzanica, M.; Tosi, G. A Survey on XAI for Cyber Physical Systems in Medicine. In Proceedings of the 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), Rome, Italy, 26–28 October 2022; pp. 265–270. [CrossRef]
22. Machlev, R.; Heistrene, L.; Perl, M.; Levy, K.Y.; Belikov, J.; Mannor, S.; Levron, Y. Explainable Artificial Intelligence (XAI) Techniques for Energy and Power Systems: Review, Challenges and Opportunities. *Energy AI* **2022**, *9*, 100169. [CrossRef]
23. Zhang, Z.; Al Hamadi, H.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* **2022** , *10*, 93104–93139. [CrossRef]
24. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
25. Sheu, R.-K.; Pardeshi, M.S. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors* **2022**, *22*, 8068. [CrossRef]
26. Owens, E.; Sheehan, B.; Mullins, M.; Cunneen, M.; Ressel, J.; Castignani, G. Explainable Artificial Intelligence (XAI) in Insurance. *Risks* **2022**, *10*, 230. [CrossRef]

27. 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. Ind. Inform.* **2022**, *18*, 5031–5042. [CrossRef]
28. Di Martino, F.; Delmastro, F. Explainable AI for Clinical and Remote Health Applications: A Survey on Tabular and Time Series Data. *Artif. Intell. Rev.* **2022**, *56*, 5261–5315. [CrossRef]
29. Weber, P.; Carl, K.V.; Hinz, O. Applications of Explainable Artificial Intelligence in Finance—A Systematic Review of Finance, Information Systems, and Computer Science literature. *Manag. Rev. Q.* **2023** , 1–41. [CrossRef]
30. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **2023**, *23*, 634. [CrossRef] [PubMed]
31. Nazir, S.; Dickson, D.M.; Akram, M.U. Survey of Explainable Artificial Intelligence Techniques for Biomedical Imaging with Deep Neural Networks. *Comput. Biol. Med.* **2023**, *156*, 106668. [CrossRef]
32. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *ArXiv* **2020**, arXiv:2006.11371.
33. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]
34. Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum, S. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Appl. Sci.* **2022**, *12*, 1353. [CrossRef]
35. Kok, I.; Okay, F.Y.; Muyanli, O.; Ozdemir, S. Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey. *arXiv* **2022**, arXiv:2206.04800.
36. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Chapter 6. Available online: https://christophm.github.io/interpretable-ml-book (accessed on 23 February 2023).
37. Zhang, K.; Xu, P.; Zhang, J. Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control. In Proceedings of the 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), Wuhan, China, 30 October–1 November 2020; pp. 711–716. [CrossRef]
38. Renda, A.; Ducange, P.; Marcelloni, F.; Sabella, D.; Filippou, M.C.; Nardini, G.; Stea, G.; Virdis, A.; Micheli, D.; Rapone, D.; et al. Federated Learning of Explainable AI Models in 6G Systems: Towards Secure and Automated Vehicle Networking. *Information* **2022**, *13*, 395. [CrossRef]
39. Sequeira, P.; Gervasio, M. Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations. *arXiv* **2019**, arXiv:1912.09007.
40. He, L.; Aouf, N.; Song, B. Explainable Deep Reinforcement Learning for UAV Autonomous Path Planning. *Aerosp. Sci. Technol.* **2021**, *118*, 107052. [CrossRef]
41. Zhang, Z.; Tian, R.; Sherony, R.; Domeyer, J.; Ding, Z. Attention-Based Interrelation Modeling for Explainable Automated Driving. In *IEEE Transactions on Intelligent Vehicles*; IEEE: Piscataway, NJ, USA, 2022. [CrossRef]
42. Cui, Z.; Li, M.; Huang, Y.; Wang, Y.; Chen, H. An Interpretation Framework for Autonomous Vehicles Decision-making via SHAP and RF. In Proceedings of the 2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI), Nanjing, China, 28–30 October 2022; pp. 1–7. [CrossRef]
43. Nahata, R.; Omeiza, D.; Howard, R.; Kunze, L. Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 223–230. [CrossRef]
44. Kim, M.; Jun, J.-A.; Song, Y.; Pyo, C.S. Explanation for Building Energy Prediction. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence, Jeju, Republic of Korea, 21–23 October 2020; pp. 1168–1170. [CrossRef]
45. Arjunan, P.; Poolla, K.; Miller, C. Energystar++: Towards More Accurate and Explanatory Building Energy Benchmarking. *Appl. Energy* **2020**, *276*, 115413. [CrossRef]
46. Movahedi, A.; Derrible, S. Interrelated Patterns of Electricity, Gas, and Water Consumption in Large-scale Buildings. *Engrxiv* **2020**, 1–22. [CrossRef]
47. Kuzlu, M.; Cali, U.; Sharma, V.; Güler, Ö. Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools. *IEEE Access* **2020**, *8*, 187814–187823. [CrossRef]
48. Chakraborty, D.; Alam, A.; Chaudhuri, S.; Başağaoğlu, H.; Sulbaran, T.; Langar S. Scenario-based Prediction of Climate Change Impacts on Building Cooling Energy Consumption with Explainable Artificial Intelligence. *Appl. Energy* **2021**, *291*, 116807. [CrossRef]
49. Golizadeh, A.Y.; Aslansefat, K.; Zhao, X.; Sadati, S.; Badiei, A.; Xiao, X.; Shittu, S.; Fan, Y.; Ma, X. Hourly Performance Forecast of a Dew point Cooler Using Explainable Artificial Intelligence and Evolutionary Optimisations by 2050. *Appl. Energy* **2021**, *281*, 116062. [CrossRef]
50. Lu, Y.; Murzakhanov, I.; Chatzivasileiadis, S. Neural Network Interpretability for Forecasting of Aggregated Renewable Generation. In Proceedings of the 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aachen, Germany, 25–28 October 2021; pp. 282–288. [CrossRef]
51. Gao, Y.; Ruan, Y. Interpretable Deep Learning Model for Building Energy Consumption Prediction Based on Attention Mechanism. *Energy Build* **2021**, *252*, 111379. [CrossRef]
52. Zdravković, M.; Ćirić, I.; Ignjatović, M. Towards Explainable AI-assisted Operations in District Heating Systems. *IIFAC-PapersOnLine* **2021**, *54*, 390–395. [CrossRef]

53. Moraliyage, H.; Dahanayake, S.; De Silva, D.; Mills, N.; Rathnayaka, P.; Nguyen, S.; Alahakoon, D.; Jennings, A. A Robust Artificial Intelligence Approach with Explainability for Measurement and Verification of Energy Efficient Infrastructure for Net Zero Carbon Emissions. *Sensors* **2022**, *22*, 9503. [CrossRef]
54. Arjunan, P.; Poolla, K.; Miller, C. BEEM: Data-driven Building Energy Benchmarking for Singapore. *Energy Build* **2022**, *260*, 111869. [CrossRef]
55. Geyer, P.; Singh, M.M.; Chen, X. Explainable AI for Engineering Design: A Unified Approach of Systems Engineering and Component-based Deep Learning. *arXiv* **2022**, arXiv:2108.13836. https://doi.org/10.48550/arXiv.2108.13836.
56. Grzeszczyk, T.A.; Grzeszczyk, M.K. Justifying Short-term Load Forecasts Obtained with the Use of Neural Models. *Energies* **2022**, *15*, 1852. [CrossRef]
57. Li, M.; Wang, Y. Power Load Forecasting and Interpretable Models based on GS_XGBoost and SHAP. *J. Phys. Conf. Ser.* **2022**, *2195*, 012028. [CrossRef]
58. Moon, J.; Park, S.; Rho, S.; Hwang, E. Interpretable Short-term Electrical Load Forecasting Scheme Using Cubist. *Comput. Intell Neurosci.* **2022**, *2022*, 1–20. [CrossRef] [PubMed]
59. Wenninger, S.; Kaymakci, C.; Wiethe, C. Explainable Long-term Building Energy Consumption Prediction using Qlattice. *Appl. Energy* **2022**, 308, 118300. [CrossRef]
60. Zdravković, M.; Ćirić, I.; Ignjatović, M. Explainable Heat Demand Forecasting for the Novel Control Strategies of District Heating Systems. *Annu. Rev. Control* **2022**, *53*, 405–413. [CrossRef]
61. Srinivasan, S.; Arjunan, P.; Jin, B.; Sangiovanni-Vincentelli, A.L.; Sultan, Z.; Poolla, K. Explainable AI for Chiller Fault-detection Systems: Gaining Human Trust. *Computer* **2021**, *54*, 60–68. [CrossRef]
62. Wastensteiner, J.; Weiss, T.M.; Haag, F.; Hopf, K. Explainable AI for Tailored Electricity Consumption Feedback–an Experimental Evaluation of Visualizations. *arXiv* **2022**, arXiv:2208.11408. https://doi.org/10.48550/arXiv.2208.11408.
63. Sim, T.; Choi, S.; Kim, Y.; Youn, S.H.; Jang, D.-J.; Lee, S.; Chun, C.-J. eXplainable AI (XAI)-Based Input Variable Selection Methodology for Forecasting Energy Consumption. *Electronics* **2022**, *11*, 2947. [CrossRef]
64. Graham, G.; Csicsery, N.; Stasiowski, E.; Thouvenin, G.; Mather, W.H.; Ferry, M.; Cookson, S.; Hasty, J. Genome-scale Transcriptional Dynamics and Environmental Biosensing. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 3301–3306. [CrossRef] [PubMed]
65. Gao, S.; Wang, Y. Explainable Deep Learning Powered Building Risk Assessment Model for Proactive Hurricane Response. *Risk Anal.* **2022** , 1–13. [CrossRef]
66. Masahiro, R.; Boyan, A.; Stefano, M.; Jamie, M.K.; Blas, M. Benito, F.H. Explainable Artificial Intelligence Enhances the Ecological Interpretability of Black-box Species Distribution Models. *Ecography* **2020**, *44*, 199–205. [CrossRef]
67. Dikshit, A.; Pradhan, B. Interpretable and Explainable AI (XAI) Model for Spatial Drought Prediction. *Sci. Total Environ.* **2021**, *801*, 149797. [CrossRef]
68. Kim, M.; Kim, D.; Jin, D.; Kim, G. Application of Explainable Artificial Intelligence (XAI) in Urban Growth Modeling: A Case Study of Seoul Metropolitan Area, Korea. *Land* **2023**, *12*, 420. [CrossRef]
69. Gramegna, A.; Giudici, P. Why to Buy Insurance? An Explainable Artificial Intelligence Approach. *Risks* **2020**, *8*, 137. [CrossRef]
70. Benhamou, E.; Ohana, J.-J.; Saltiel, D. Guez, B.; Ohana, S. Explainable AI (XAI) Models Applied to Planning in Financial Markets. Université Paris-Dauphine Research Paper No. 3862437. 2021. Available online: https://ssrn.com/abstract=3862437 (accessed on 2 February 2023). [CrossRef]
71. Gite, S.; Khatavkar, H.; Kotecha, K.; Srivastava, S.; Maheshwari, P.; Pandey N. Explainable Stock Prices Prediction from Financial News Articles using Sentiment Analysis. *PeerJ. Comput. Sci.* **2021**, *7*, e340. [CrossRef] [PubMed]
72. Babaei, G.; Giudici, P. Which SME is Worth an Investment? An Explainable Machine Learning Approach. 2021. Available online: http://dx.doi.org/10.2139/ssrn.3810618 (accessed on 2 February 2023).
73. de Lange, P.E.; Melsom, B.; Vennerod, C.B.; Westgaard, S. Explainable AI for Credit Assessment in Banks. *J. Risk Financ. Manag.* **2022**, *15*, 556. [CrossRef]
74. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable AI in Fintech Risk Management. *Front. Artif. Intell.* **2020**, *3*, 26. [CrossRef]
75. Kumar, S.; Vishal, M.; Ravi, V. Explainable Reinforcement Learning on Financial Stock Trading using SHAP. *arXiv* **2022**, arXiv:2208.08790. https://doi.org/10.48550/arXiv.2208.08790.
76. Pawar, U.; O'Shea, D.; Rea, S.; O'Reilly, R. Incorporating Explainable Artificial Intelligence (XAI) to Aid the Understanding of Machine Learning in the Healthcare Domain. In Proceedings of the The 28th Irish Conference on Artificial Intelligence and Cognitive ScienceAt: Technological University Dublin, Dublin, Ireland, 7–8 December 2020; Volume 2771, pp. 169–180.
77. Dissanayake, T.; Fernando, T.; Denman, S.; Sridharan, S.; Ghaemmaghami, H.; Fookes, C. A Robust Interpretable Deep Learning Classifier for Heart Anomaly Detection without Segmentation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2162–2171. [CrossRef]
78. Cecilia, P.; Alan, P.; Dino, P. Doctor XAI: An Ontology-based Approach to Black-box Sequential Data Classification Explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Association for Computing Machinery, New York, NY, USA, 27–30 January 2020; pp. 629–639. [CrossRef]
79. Naik, H.; Goradia, P.; Desai, V.; Desai, Y.; Iyyanki, M. Explainable Artificial Intelligence (XAI) for Population Health Management—An Appraisal. *Eur. J. Electr. Eng. Comput. Sci.* **2021**, *5*, 64–76. [CrossRef]
80. Beebe-Wang, N.; Okeson, A.; Althoff, T.; Lee, S.-I. Efficient and Explainable Risk Assessments for Imminent dementia in an Aging Cohort Study. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2409–2420. [CrossRef]

81. Kim, S.-H.; Jeon, E.-T.; Yu, S.; Oh, K.; Kim, C.K.; Song, T.-J.; Kim, Y.-J.; Heo, S.H.; Park, K.-Y.; Kim, J.-M.; et al. Interpretable Machine Learning for Early Neurological Deterioration Prediction in Atrial Fibrillation-related Stroke. *Sci. Rep.* **2021**, *11*, 20610. https://doi.org10.1038/s41598-021-99920-7. [CrossRef] [PubMed]
82. Rashed-Al-Mahfuz, M.; Haque, A.; Azad, A.; Alyami, S.A.; Quinn, J.M.; Moni, M.A. Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic kidney disease (ckd) for use in low-cost diagnostic screening. *IEEE J. Transl. Eng. Health Med.* **2021**, *9*, 4900511. [CrossRef] [PubMed]
83. Zhang, Y.; Yang, D.; Liu, Z.; Chen, C.; Ge, M.; Li, X.; Luo, T.; Wu, Z.; Shi, C.; Wang, B.; et al. An Explainable Supervised Machine Learning Predictor of Acute Kidney Injury After Adult Deceased Donor Liver Transplantation. *J. Transl. Med.* **2021**, *19*, 1–15 . [CrossRef]
84. Mousavi, S.; Afghah, F.; Acharya, U.R. HAN-ECG: An Interpretable Atrial Fibrillation Detection Model Using Hierarchical Attention Networks. *Comput. Biol. Med.* **2020**, *127*, 104057. [CrossRef]
85. Ivaturi, P.; Gadaleta, M.; Pandey, A.C.; Pazzani, M.; Steinhubl, S.R.; Quer, G. A Comprehensive Explanation Framework for Biomedical Time Series Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2398–2408. [CrossRef] [PubMed]
86. Shashikumar, S.P.; Josef, C.S.; Sharma, A.; Nemati, S. DeepAISE an Interpretable and Recurrent Neural Survival Model for Early Prediction of Sepsis. *Artif. Intell. Med.* **2021**, *113*, 102036. [CrossRef] [PubMed]
87. Filtjens, B.; Ginis, P.; Nieuwboer, A.; Afzal, M.R.; Spildooren, J.; Vanrumste, B.; Slaets, P. Modelling and Identification of Characteristic Kinematic Features Preceding Freezing of Gait with Convolutional Neural Networks and Layer-wise Relevance Propagation. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 341. [CrossRef]
88. Dutt, M.; Redhu, S.; Goodwin, M.; Omlin, C.W. SleepXAI: An Explainable Deep Learning Approach for Multi-class Sleep Stage Identification. *Appl. Intell.* **2022** , 1–14. [CrossRef]
89. Brunese, L.; Mercaldo, F.; Reginelli, A.; Santone, A. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Comput. Methods Programs Biomed.* **2020**, *196*, 105608. [CrossRef]
90. Yang, G.; Ye, Q.; Xia, J. Unbox the Black-box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Minireview, two Showcases and Beyond. *Inf. Fusion* **2022**, *77*, 29–52. [CrossRef]
91. Singh, A.; Balaji, J.J.; Rasheed, M.A.; Jayakumar, V.; Raman, R.; Lakshminarayanan, V. Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. *Clin. Ophthalmol.* **2021**, *15*, 2573–2581. [CrossRef] [PubMed]
92. Xu, F.; Jiang, L.; He, W.; Huang, G.; Hong, Y.; Tang, F.; Lv, J.; Lin, Y.; Qin, Y.; Lan, R.; et al. The Clinical Value of Explainable Deep Learning for Diagnosing Fungal Keratitis Using in Vivo Confocal Microscopy Images. *Front. Med.* **2021**, *8*, 797616. [CrossRef] [PubMed]
93. Chetoui, M.; Akhloufi, M.A.; Yousefi, B.; Bouattane, E.M. Explainable COVID-19 Detection on Chest X-rays Using an End-to-end Deep Convolutional Neural Network Architecture. *Big Data Cogn. Comput.* **2021**, *5*, 73. bdcc5040073. [CrossRef]
94. Barata, C.; Celebi, M.E.; Marques, J.S. Explainable Skin Lesion Diagnosis Using Taxonomies. *Pattern Recognit.* **2021**, *110*, 107413. [CrossRef]
95. Singh, R.K.; Pandey, R.; Babu, R.N. COVIDScreen: Explainable Deep Learning Framework for Differential Diagnosis of COVID-19 using Chest Xrays. *Neural. Comput. Appl.* **2021**, *33*, 8871–8892. [CrossRef]
96. Shi, W.; Tong, L.; Zhu, Y.; Wang, M.D. COVID-19 Automatic Diagnosis with Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2376–2387. 3074893. [CrossRef]
97. Figueroa, K.C.; Song, B.; Sunny, S.; Li, S.; Gurushanth, K.; Mendonca, P.; Mukhia, N.; Patrick, S.; Gurudath, S.; Raghavan, S.; et al. Interpretable Deep Learning Approach for Oral Cancer Classification using Guided Attention Inference Network. *J. Biomed. Opt.* **2022**, *27*, 015001. [CrossRef]
98. Malhotra, A.; Mittal, S.; Majumdar, P.; Chhabra, S.; Thakral, K.; Vatsa, M.; Singh, R.; Chaudhury, S.; Pudrod, A.; Agrawal, A.; et al. Multi-task Driven Explainable Diagnosis of COVID-19 using Chest X-ray Images. *Pattern Recognit.* **2022**, *122*, 108243. [CrossRef]
99. Civit-Masot, J.; Bañuls-Beaterio, A.; Domínguez-Morales, M.; Rivas-Pérez, M.; Muñoz-Saavedra, L.; Corral, J.M.R. Non-small Cell Lung Cancer diagnosis aid with Histopathological Images using Explainable Deep Learning Techniques. *Comput. Methods Programs Biomed.* **2022**, *226*, 107108. [CrossRef]
100. Kim, D.; Chung, J.; Choi, J.; Succi, M.D.; Conklin, J.; Longo, M.G.F.; Ackman, J.B.; Little, B.P.; Petranovic, M.; Kalra, M.K.; et al. Accurate Auto-labeling of Chest X-ray Images based on Quantitative Similarity to an Explainable AI Model. *Nat. Commun.* **2022**, *13*, 1867. [CrossRef]
101. Aldhahi, W.; Sull, S. Uncertain-CAM: Uncertainty-Based Ensemble Machine Voting for Improved COVID-19 CXR Classification and Explainability. *Diagnostics* **2023**, *13*, 441. [CrossRef]
102. Mercaldo, F.; Belfiore, M.P.; Reginelli, A.; Brunese, L.; Santone, A. Coronavirus COVID-19 Detection by Means of Explainable Deep Learning. *Sci. Rep.* **2023**, *13*, 462. [CrossRef] [PubMed]
103. Oztekin, F.; Katar, O.; Sadak, F.; Yildirim, M.; Cakar, H.; Aydogan, M.; Ozpolat, Z.; Talo Yildirim, T.; Yildirim, O.; Faust, O.; et al. An Explainable Deep Learning Model to Prediction Dental Caries Using Panoramic Radiograph Images. *Diagnostics* **2023**, *13*, 226. [CrossRef] [PubMed]
104. Naz, Z.; Khan, M.U.G.; Saba, T.; Rehman, A.; Nobanee, H.; Bahaj, S.A. An Explainable AI-Enabled Framework for Interpreting Pulmonary Diseases from Chest Radiographs.*Cancers* **2023**, *15*, 314. [CrossRef]
105. Mukhtorov, D.; Rakhmonova, M.; Muksimova, S.; Cho, Y.-I. Endoscopic Image Classification Based on Explainable Deep Learning. *Sensors* **2023**, *23*, 3176. [CrossRef] [PubMed]

106. Grezmak, J.; Zhang, J.; Wang, P.; Loparo, K.A.; Gao, R.X. Interpretable Convolutional Neural Network Through Layer-wise Relevance Propagation for Machine Fault Diagnosis. *IEEE Sen. J.* **2020**, *20*, 3172–3181. [CrossRef]
107. Serradilla, O.; Zugasti, E.; Cernuda, C.; Aranburu, A.; de Okariz, J. R.; Zurutuza, U. Interpreting Remaining Useful Life Estimations Combining Explainable Artificial Intelligence and Domain Knowledge in Industrial Machinery. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
108. Oh, C.; Jeong, J. VODCA: Verification of Diagnosis Using CAM-Based Approach for Explainable Process Monitoring. *Sensors* **2020**, *20*, 6858. [CrossRef]
109. Abid, F.B.; Sallem, M.; Braham, A. Robust Interpretable Deep Learning for Intelligent Fault Diagnosis of Induction Motors. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 3506–3515. [CrossRef]
110. Hong, C.W.; Lee, C.; Lee, K.; Ko, M.-S.; Kim, D.E.; Hur, K. Remaining Useful Life Prognosis for Turbofan Engine Using Explainable Deep Neural Networks with Dimensionality Reduction. *Sensors* **2020**, *20*, 6626. [CrossRef]
111. Kim, M.S.; Yun, J.P.; Park, P. An Explainable Convolutional Neural Network for Fault Diagnosis in Linear Motion Guide. *IEEE Trans. Ind. Inform.* 2021, 17, 4036–4045. [CrossRef]
112. Darian, M.O.; Gilbert-Rainer, G. Stable and Explainable Deep Learning Damage Prediction for Prismatic Cantilever Steel Beam. *Comput. Ind.* **2021**, *125*, 103359. [CrossRef]
113. Liu, C.; Qin, C.; Shi, X.; Wang, Z.; Zhang, G.; Han, Y. TScatNet: An Interpretable Cross-Domain Intelligent Diagnosis Model with Antinoise and Few-Shot Learning Capability. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3506110. [CrossRef]
114. Brito, L.C.; Susto, G.A.; Brito, J.N.; Duarte, M.A. An Explainable Artificial Intelligence Approach for Unsupervised Fault Detection and Diagnosis in Rotating Machinery. *Mech. Syst. Signal Process.* **2022**, *163*, 108105. [CrossRef]
115. Li, T.; Zhao, Z.; Sun, C.; Cheng, L.; Chen, X.; Yan, R.; Gao, R.X. WaveletKernelNet: An Interpretable Deep Neural Network for Industrial Intelligent Diagnosis. in *IEEE Trans. Syst. Man. Cybern. Syst.* **2022**, *52*, 2302–2312. [CrossRef]
116. Brusa, E.; Cibrario, L.; Delprete, C.; Di Maggio, L.G. Explainable AI for Machine Fault Diagnosis: Understanding Features' Contribution in Machine Learning Models for Industrial Condition Monitoring. *Appl. Sci.* **2023**, *13*, 2038. [CrossRef]
117. Chen, H.-Y.; Lee, C.-H. Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis. *IEEE Access* **2020**, *8*, 134246–134256. [CrossRef]
118. Sun, K.H.; Huh, H.; Tama, B.A.; Lee, S.Y.; Jung, J.H.; Lee, S. Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps. *IEEE Access* **2020**, *8*, 129169–129179. [CrossRef]
119. Wang, M.; Zheng, K.; Yang, Y.; Wang, X. An Explainable Machine Learning Framework for Intrusion Detection Systems. *IEEE Access* **2020**, *8*, 73127–73141. [CrossRef]
120. Alenezi, R.; Ludwig, S.A. Explainability of Cybersecurity Threats Data Using SHAP. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; pp. 1–10. [CrossRef]
121. Roshan, K.; Zafar, A. Utilizing XAI Technique to Improve Autoencoder based Model for Computer Network Anomaly Detection with Shapley Additive explanation (SHAP). *arXiv* **2021**, arXiv:2112.08442.
122. Karn, R.R.; Kudva, P.; Huang, H.; Suneja, S.; Elfadel, I.M. Cryptomining Detection in Container Clouds Using System Calls and Explainable Machine Learning. *IEEE Trans. Parallel Distrib. Syst.* **2021**, *32*, 674–691. [CrossRef]
123. Le, T.-T.-H.; Kim, H.; Kang, H.; Kim, H. Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method. *Sensors* **2022**, *22*, 1154. [CrossRef]
124. El Houda, Z.A.; Brik, B.; Senouci, S.-M. A Novel IoT-Based Explainable Deep Learning Framework for Intrusion Detection Systems. *IEEE Internet Things Mag.* **2022**, *5*, 20–23. [CrossRef]
125. Oseni, A.; Moustafa, N.; Creech, G.; Sohrabi, N.; Strelzoff, A.; Tari, Z.; Linkov, I. An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 1000–1014. [CrossRef]
126. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security. *IEEE Internet Things J.* **2023**, *10*, 2967–2978. [CrossRef]
127. Viana, C.M.; Santos, M.; Freire, D.; Abrantes, P.; Rocha, J. Evaluation of the factors Explaining the Use of Agricultural Land: A Machine Learning and Model-Agnostic Approach. *Ecol. Indic.* **2021**, *131*, 108200. [CrossRef]
128. Ryo, M. Explainable Artificial Intelligence and Interpretable Machine Learning for Agricultural Data Analysis. *Artif. Intell. Agric.* **2022**, *6*, 257–265. [CrossRef]
129. Adak, A.; Pradhan, B.; Shukla, N.; Alamri, A. Unboxing Deep Learning Model of Food Delivery Service Reviews Using Explainable Artificial Intelligence (XAI) Technique. *Foods* **2022**, *11*, 2019. [CrossRef] [PubMed]
130. Cartolano, A.; Cuzzocrea, A.; Pilato, G.; Grasso, G.M. Explainable AI at Work! What Can It Do for Smart Agriculture? In Proceedings of the 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM), Naples, Italy, 5–7 December 2022; pp. 87–93. [CrossRef]
131. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Guanter, L. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 024019. [CrossRef]
132. Kawakura, S.; Hirafuji, M.; Ninomiya, S.; Shibasaki, R. Analyses of Diverse Agricultural Worker Data with Explainable Artificial Intelligence: XAI based on SHAP, LIME, and LightGBM. *Eur. J. Agric. Food Sci.* **2022**, *4*, 11–19.. [CrossRef]
133. Li, Z.; Jin, J.; Wang, Y.; Long, W.; Ding, Y.; Hu, H.; Wei, L. ExamPle: Explainable Deep Learning Framework for the Prediction of Plant Small Secreted Peptides. *Bioinformatics* **2023**, *39*, btad108. [CrossRef]

134. Kundu, N.; Rani, G.; Dhaka, V.S.; Gupta, K.; Nayak, S.C.; Verma, S.; Ijaz, M.F.; Woźniak, M. IoT and Interpretable Machine Learning Based Framework for Disease Prediction in Pearl Millet. *Sensors* **2021**, *21*, 5386. [CrossRef]
135. Kawakura, S.; Hirafuji, M.; Ninomiya, S.; Shibasaki, R. Visual Analysis of Agricultural Workers using Explainable Artificial Intelligence (XAI) on Class Activation Map (CAM) with Characteristic Point Data Output from OpenCV-based Analysis. *Eur. J. Artif. Intell. Mach. Learn.* **2022** , *2*, 1–8. [CrossRef]
136. Apostolopoulos, I.D.; Athanasoula, I.; Tzani, M.; Groumpos, P.P. An Explainable Deep Learning Framework for Detecting and Localising Smoke and Fire Incidents: Evaluation of Grad-CAM++ and LIME. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 1124–1135. [CrossRef]
137. Ngo, Q.H.; Kechadi, T.; Le-Khac, N.A. OAK4XAI: Model Towards Out-of-Box eXplainable Artificial Intelligence for Digital Agriculture. In *Artificial Intelligence XXXIX: 42nd SGAI International Conference on Artificial Intelligence, AI 2022, Cambridge, UK, 13–15 December 2022*; Springer International Publishing: Cham, Switzerland, 2022; pp. 238–251.
138. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You? In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'16, San Francisco, CA, USA, 13–17 August 2016; ACM Press: New York, NY, USA, 2016; pp. 1135–1144.
139. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017 ; pp. 4765–4774.
140. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2018 ; Volume 32, ISBN 978-1-5108-6096-4. [CrossRef]
141. 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 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
142. Arras, L.; Horn, F.; Montavon, G.; Müller, K. R.; Samek, W. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLoS ONE* **2017**, *12*, e0181142. [CrossRef]
143. Arrieta, A.B.; Díaz-Rodríguez, N.; Ser, J.D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R. et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
144. David, G. *Broad Agency Announcement Explainable Artificial Intelligence (XAI)*; Technical report; Defense Advanced Research Projects Agency Information Innovation Office, Arlington, VA, USA, 2016; pp. 22203–2114.
145. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
146. Gunning, D.; Vorm, E.; Wang, Y.; Turek, M. DARPA's Explainable AI (XAI) Program: A Retrospective. *Authorea* **2021**, *2*, e61. [CrossRef]
147. Schoonderwoerd, T.A.; Jorritsma, W.; Neerincx, M.A.; Van Den Bosch, K. Human-centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *Int. J. Hum.-Comput. Stud.* **2021**, *154*, 102684. [CrossRef]
148. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
149. Burkart, N.; Huber, M.F. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* **2021**, *70*, 245–317. [CrossRef]
150. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
151. Koh, P.W.; Liang, P. Understanding Black-box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1885–1894.
152. Goyal, A.; He, K.; Bengio, Y. Understanding and improving deep learning techniques for image recognition. *arXiv* **2021**, arXiv:2104.08821. https://doi.org/10.48550/arXiv.1907.06119.
153. Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A.M. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In *Machine Learning and Knowledge Extraction. CD-MAKE*; Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11015._1. [CrossRef]
154. Hooker, G.; Erhan, D.; Kindermans, P.J. A Benchmark for Interpretability Methods in Deep Neural Networks. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019. [CrossRef]
155. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a Right to Explanation of Automated Decision-making Does not Exist in the General Data Protection Regulation. *Int. Data Priv. Law* **2018**, *7*, 76–99. [CrossRef]