Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey

İbrahim Kök, Feyza Yıldırım Okay, Özgecan Muyanlı, and Suat Özdemir

***Abstract*—Artificial intelligence (AI) and machine learning (ML) are widely employed to make the solutions more accurate and autonomous in many smart and intelligent applications in the Internet of Things (IoT). In these IoT applications, the performance and accuracy of AI/ML models are the main concerns; however, the transparency, interpretability, and responsibility of the models' decisions are often neglected. Moreover, in AI/ML-supported next-generation IoT applications, there is a need for more reliable, transparent, and explainable systems. In particular, regardless of whether the decisions are simple or complex, how the decision is made, which features affect the decision, and their adoption and interpretation by people or experts are crucial issues. Also, people typically perceive unpredictable or opaque AI outcomes with skepticism, which reduces the adoption and proliferation of IoT applications. To that end, explainable AI (XAI) has emerged as a promising research topic that allows ante-hoc and post-hoc functioning and stages of black-box models to be transparent, understandable, and interpretable. In this article, we provide an in-depth and systematic review of recent studies that use XAI models in the scope of the IoT domain. We classify the studies according to their methodology and application areas. Additionally, we highlight the challenges and open issues and provide promising future directions to lead the researchers in future investigations.**

***Index Terms*—Explainability, explainable artificial intelligence (XAI), Internet of Things (IoT), interpretability, interpretable machine learning (IML).**

I. Introduction

THE Internet of Things (IoT) is a prominent technology that connects smart things, allowing them to communicate with each other and provide better services to users [1]. By managing and controlling its underlying technologies, IoT transforms traditional electronic devices, such as sensors, actuators, RFID tags, cell phones, etc., into smart objects. Thus, it empowers objects to see, hear, think, and perform certain tasks by enabling them to synchronize and share information with each other [2]. In order to improve the quality of life, IoT offers numerous potentials and opportunities in different application areas, such as smart cities, smart buildings, smart agriculture, healthcare, finance, and military [3].

IoT allows the generation of massive amounts of data that needs to be fine-grained analysis. While this raw data is meaningless, artificial intelligence (AI) systems make it possible to extract meaningful information and provide insightful decisions that affect human lives (in critical fields, such as healthcare or autonomous systems) [4]. In recent years, as in other fields, IoT applications have extensively used AI models to overcome the problems caused by the rapidly increasing amount of data and the number of devices [5]. In particular, many AI models are widely employed in autonomous network management, device management, service management, and analysis of massive IoT data, which require smart decision making with high precision and accuracy [6].

As the AI technology becomes more integrated into our daily lives, it becomes increasingly important to comprehend how and why decisions are made. However, as machine learning (ML) models continue to evolve and become more powerful, they also tend to become increasingly complex and less transparent. These powerful models are generally called "black-box" and suffer from opaqueness. In other words, they exclude the internal logic from their users or stakeholders [7]. Therefore, recently, the concept of explainable AI (XAI) has started to attract the attention of researchers to cope with the current challenges and to design more explainable and interpretable AI systems. XAI enables the shining of light on the opaqueness of the black-box models to reveal unseen/hidden information, such as feature importance and correlations between features. They will provide more detailed information about how, why, and when they make decisions about the inner workings of black-box models and provide transparency to their users [8]. Thus, users are able to evaluate not only the result but also the input factors affecting the result when making a decision. XAI techniques achieve both explainability and high accuracy when they are applied to powerful and complex models. The studies of [7] and [9] emphasize the need for explanations of human-related issues, not just in computer science but also in cognitive science, philosophy, and psychology. According to these studies, reaching the outcome without any interpretation may result in intentional or unintentional discrimination or trust issues. XAI mitigates these problems by providing verification, improvement, learning, and compliance of legislation [9].

In this study, we envision that the integration of XAI approaches into the IoT domain will reveal serious research potential in terms of the transparency, explainability, and

Manuscript received 10 October 2022; revised 23 February 2023; accepted 12 June 2023. Date of publication 20 June 2023; date of current version 8 August 2023. *(Corresponding author: İbrahim Kök.)*

İbrahim Kök is with the Department of Computer Engineering, Pamukkale University, 20160 Denizli, Turkey (e-mail: ikok@pau.edu.tr).

Feyza Yıldırım Okay is with the Department of Computer Engineering, Gazi University, 06560 Ankara, Turkey (e-mail: feyzaokay@gazi.edu.tr).

Özgecan Muyanlı and Suat Özdemir are with the Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey (e-mail: ozgecanmuyanli@gmail.com; ozdemir@cs.hacettepe.edu.tr).

Digital Object Identifier 10.1109/JIOT.2023.3287678

2327-4662 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

interpretability of the AI and ML models in IoT. However, we see that there is not enough research effort in the literature on this subject. To date, there are several survey papers focused on XAI from a general perspective. For example, the studies of [10], [11], and [12] explain XAI and focus on XAI concepts, terminology, taxonomy, and challenges, whereas the study of [13] examines XAI for time-series data. However, to the best of our knowledge, there is no existing survey paper investigating or addressing the use of XAI techniques in the IoT domain. To fulfill this gap, in this article, we provide a comprehensive review of current studies on XAI in the IoT domain.

A. Motivation: The Role of XAI in IoT

In the context of the IoT, XAI is increasingly important because many IoT devices and systems make decisions that can have a significant impact on people's lives and the world around us. XAI provides a means to understand, interpret, and validate the decisions made by AI systems, which is particularly important for IoT devices and systems because they often operate autonomously and in real time.

IoT involves the deployment of many devices and systems in complex and ever-changing environments, making it difficult to understand the reasoning behind their decisions. XAI addresses this challenge by providing tools and methods that help us understand the inner workings of AI systems and validate the decisions they make.

For resource-constrained IoT devices, AI-based solutions are expensive because they typically require a significant amount of data and computational power. AI approaches often require: 1) hyperparameter optimization; 2) fine-tuning; 3) massive data sets; 4) robust computational capabilities; and 5) continuous data training. Therefore, it is important to apply AI algorithms in IoT in a resource-aware manner. XAI algorithms help identify the most important features in decision making and thus can help reduce the amount of data and computational requirements in AI-based IoT systems. By knowing which features play a role in decision making, IoT system designers can design efficient data collection and processing.

In the IoT domain, another important aspect of XAI is its ability to build trust between people and the AI systems they interact with. IoT devices and systems are becoming increasingly ubiquitous and are being used in a growing number of applications, from smart homes to healthcare. As people become more dependent on these systems, it is increasingly important to ensure that they are transparent, reliable, and trustworthy. XAI can help build that trust by providing a means for people to understand how the systems are making decisions and to validate their accuracy.

Here, considering the goals of XAI, the following additional reasons motivate us to adopt XAI approaches in the field of IoT [14], [15], [16], as shown in Fig. 1.

1) *Data Collection:* IoT devices generate vast amounts of data that can be used as input for XAI algorithms. By analyzing this data, XAI systems can make predictions, identify patterns, and make decisions based on the information.

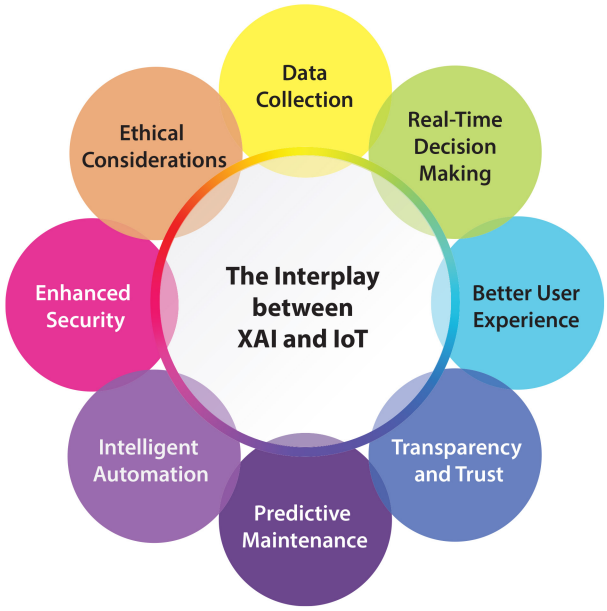

Fig. 1. Interplay between XAI and IoT.

2) *Real-Time Decision Making:* XAI algorithms can use data from IoT devices to make real-time decisions, such as controlling the temperature in a smart home or optimizing a supply chain. These systems can provide the transparency and accountability needed to build trust in AI. By incorporating XAI into IoT systems, it is possible to improve the quality of decisions that are made based on the data generated by IoT devices. XAI algorithms can analyze the data and provide insights into how decisions are made, which can be especially useful in complex or high-stakes situations.
3) *Better User Experience:* XAI can be used to create more intuitive and user-friendly interfaces for IoT devices. For example, XAI algorithms can be used to interpret the data generated by IoT devices and present it to users in a way that is easy to understand and interact with.
4) *Transparency and Trust:* XAI helps to increase transparency and accountability in IoT systems by providing explanations for the decisions that are made. This can help to build trust in the systems, especially in applications where the decisions made by the systems have a significant impact on people's lives.
5) *Predictive Maintenance:* XAI can be used in combination with IoT to support predictive maintenance in industries, such as manufacturing, transportation, and healthcare. By analyzing data from connected devices, XAI algorithms can identify potential problems before they occur, enabling proactive maintenance and reducing downtime.
6) *Intelligent Automation:* XAI can be used in conjunction with IoT to create intelligent automation systems that are capable of making decisions based on data collected from connected devices. These systems can be used to control various aspects of an IoT network, such as optimizing energy consumption or managing resources.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

7) *Enhanced Security:* XAI algorithms can be used to monitor IoT devices for potential security threats and to detect and respond to anomalies in real time. This can help to improve the security of IoT systems.
8) *Ethical Considerations:* As XAI and IoT are used in more critical applications, such as healthcare and finance, it is important to ensure that these systems are transparent and accountable and that their outputs align with ethical considerations. The combination of XAI and IoT can help to address these concerns by providing insights into how and why decisions are being made.

The interplay between XAI and IoT involves using the data generated by connected devices to support decision making while ensuring that the decisions are transparent and understandable to humans. This can help to create more intelligent, efficient, and trustworthy IoT systems.

To gain a clearer perspective on the requirements for XAI and IoT applications, several scenarios are presented. These scenarios can be derived for almost any IoT application. Here, are some illustrative examples to highlight the role of XAI in the IoT domain.

1) Smart Home: In a smart home, various devices, such as sensors, cameras, and thermostats collect data and make decisions to automate various tasks. For example, a smart thermostat can use XAI to determine the optimal temperature based on home occupancy, time of day, and weather conditions [17]. XAI can provide explanations for the decisions made by the smart thermostat. It can show why the temperature is set to a particular value, based on the data collected from the sensors and cameras. Furthermore, XAI can also help to ensure the transparency and trustworthiness of the smart home system. By offering justifications for the decisions made by the system, XAI can foster trust between the homeowner and the technology and guarantee that the technology is being utilized in a responsible and ethical manner.

2) Healthcare: In a healthcare environment, IoT devices, such as wearable devices and smart medical equipment, collect data, and make decisions to support patient care. For example, a wearable device equipped with XAI can monitor a patient's vital signs and make recommendations for treatment based on the collected data [18]. XAI can furnish explanations for the choices made by wearable devices. It can demonstrate the reason behind a specific treatment recommendation based on the patient's vital signs and other relevant information.

3) Air Quality: In air quality forecasting, when a significant number of IoT devices are deployed over a large area, resource constraints may result in the failure of these devices over time due to the processing and collection of all gases. This can lead to communication disruptions between devices or missing data collection. For example, sensors equipped with XAI can help identify which gases are crucial for the specific problem and application requirements. As a result, XAI can aid in the effective redistribution of IoT devices to detect only these critical gases, leading to a more resource-efficient outcome, including a longer lifetime for IoT devices and cost savings for developers.

4) Industrial Control Systems: In these systems, IoT devices, such as sensors and actuators, collect data, and make decisions to control various industrial processes. For example, a machine equipped with XAI can monitor production processes, detect anomalies, and make recommendations for optimizing the process [19]. XAI can provide explanations for the decisions made by the machine. It can show why a particular process is optimized based on the data collected from the sensors and actuators.

B. Research Methodology

After identifying the motivation for the study, a research methodology is established. It then describes the procedures used to select eligible papers that are consistent with the identified motivation.

1) *Literature Search Phase:* The first step is to define subject-specific search strings. These search terms include "XAI," "interpretable ML (IML)," "XAI in IoT," and "IML in IoT." These keywords are searched on the following digital libraries: ScienceDirect [20], IEEE Xplore [21], and Springer [22] to retrieve the related papers.
2) *Paper Selection Phase:* We used the following criteria to determine the papers to be excluded from this study: a) papers without peer review; b) white papers; c) papers without a robust evaluation section; d) papers published before 2008; and e) papers not directly related to IoT.
3) *Paper Classification Phase:* Due to page limitations, we selected 45 papers that focus on XAI in IoT and satisfy our selection criteria. These papers are classified in eight different IoT domains as follows: 2 articles in Autonomous Systems and Robotics, 4 in Energy Management, 5 in Environment, 4 in Finance, 7 in Healthcare, 7 in Industrial, 13 in Security and Privacy, and 3 in Smart Agriculture.

C. Contribution

As it is explained in the previous section XAI in IoT will play an important role in the future. To the best of our knowledge, there is not any previous work that comprehensively summarizes XAI in IoT. The main contributions of this article are as follows.

1) This study represents the first comprehensive and systematic survey paper that addresses and discusses the interplay between XAI and IoT, filling a critical gap in the literature by providing an original and novel contribution that has not been previously explored.
2) We present a thorough and rigorous comparison of recent studies that have investigated the application of XAI techniques in the IoT domain.
3) We highlight both the current and future challenges of XAI in IoT, while also outlining potential research directions to address these challenges and facilitate the development of improved and efficient XAI solutions.

The remainder of this article is organized as follows. In Section II, we present the terminology, taxonomy, and methodology of XAI. In Section III, we review current studies addressing XAI by considering IoT application areas. Section IV outlines challenges, open issues, and

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

future research directions. Finally, Section V concludes this article.

II. Explainable Artificial Intelligence

A. Terminology and Definitions

It is difficult to give a precise definition to the concept of explanation, which refers to the process of providing a reason or justification for an event, phenomenon, decision, or action. Explanations can be used to help understand the causes and consequences of a particular situation, or to provide insight into how things work or why they happen. They can be based on knowledge, data, intuition, or a combination of these. The goal of an explanation is to increase understanding and provide a basis for making informed decisions or taking appropriate action. Although some systems are inherently explainable, an explanation helps make a system more understandable. In particular, determining what is or is not a good explanation is a controversial issue in [23].

The term XAI has only recently gained popularity and widespread use, with roots dating back to the late 2010s. Growing concerns about the ethical implications of AI, particularly the lack of transparency and accountability in complex ML models, have fueled the need for XAI. Researchers and practitioners in the AI community have begun to focus on developing AI systems that are not only accurate and effective, but also transparent and interpretable [24].

The U.S. Defense Advanced Research Projects Agency (DARPA) played a key role in advancing XAI research by launching the XAI program in 2016 with the goal of developing AI systems that can provide clear, verifiable, and trustworthy explanations of their decision-making processes. This program brought together experts from various fields, including computer science, psychology, philosophy, and ethics, to explore new approaches to XAI [25].

XAI refers to AI systems that are transparent and interpretable, providing insight into their inner workings and decision-making processes. One example of XAI is in medical diagnosis. Imagine an AI system designed to diagnose skin lesions as either benign or malignant. The system uses a convolutional neural network (CNN) trained on a large data set of images of skin lesions to make predictions [26]. Typically, a CNN is considered a black box model because it is difficult to understand how the model makes predictions based on the input data. With XAI, however, the inner workings of the model can be explained. The model can be visualized to show which parts of the input image the model is paying attention to when making its prediction. This information can help validate the accuracy of the model and understand why it made a particular prediction.

Another example is financial risk assessment. AI systems such as gradient boosting decision trees (GBDT) are used to predict the risk of default for loan applicants. GBDT trains on historical credit data to make its predictions. On the other hand, XAI provides explanations about which characteristics, such as income, credit score, and loan amount have the most impact on the model's prediction [27].

In addition, XAI can provide a quantifiable measure of the model's confidence in its prediction. This information can be used to flag cases where the model is uncertain or to fine-tune the model to improve its performance. In this way, XAI can help build trust in AI systems by making them more transparent and interpretable, and by providing insight into their inner workings and decision-making processes.

In literature, XAI and IML are similar concepts that are frequently used interchangeably [25]. Both concepts emphasize the importance of explainability and interpretability, respectively. However, there are other similar terms used for the same purpose in the literature with subtle differences. To clarify the similarities and differences among these terms, we will provide brief definitions as follows [7], [12], [28], [29].

1) *Explainability:* It is the ability to have explanatory details and reasons a model provides to clarify its functioning and facilitate its understanding.
2) *Interpretability:* It is the ability to extend a model or its predictions understandable by humans. It is expressed through transparency.
3) *Understandability/Intelligibility:* It is the ability to make a human grasp its function about how it works without having to explain its underlying structure or algorithmic ways.
4) *Comprehensibility:* It is the ability to express learned information to make it understandable to humans.
5) *Transparency:* It is the opposite of being opaque for a black-box model. A system or a model becomes transparent when it is understandable by humans.
6) *Faithfulness:* It is the ability to be consistent in choosing the truly relevant features.
7) *Informativeness:* It is the ability of a strategy for explainability to offer end-users meaningful information.
8) *Explicitness:* It is the ability of a method to deliver immediate and clear explanations.

B. XAI Taxonomy

The taxonomy of XAI can be classified from a variety of perspectives [12], [30], as shown in Fig. 2. It is worth pointing out that there may be overlaps when a method is categorized under this taxonomy. That is, a method can be classified into one or more categories. A method can be classified as post-hoc, model-agnostic, or local, for example. Accordingly, it is more accurate to examine each method separately under its own subtaxonomy classification. Furthermore, each figure should be considered as an example method in the related group. Different figures can be used to represent the taxonomy.

1) Ante-Hoc Versus Post-Hoc: An explanation can be provided for a model in the pretraining, in-training, or post-training phases. XAI can be applied externally in pretraining or post-training phases, or the model itself is intrinsically interpretable during the training which is also called transparent.

1) Ante-hoc methods involve interpreting externally before the training phase or internally during the training phase. At the end of the training phase, the model becomes already explainable. Transparent methods, such as Decision Tree and Sparse Linear Regression generally

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

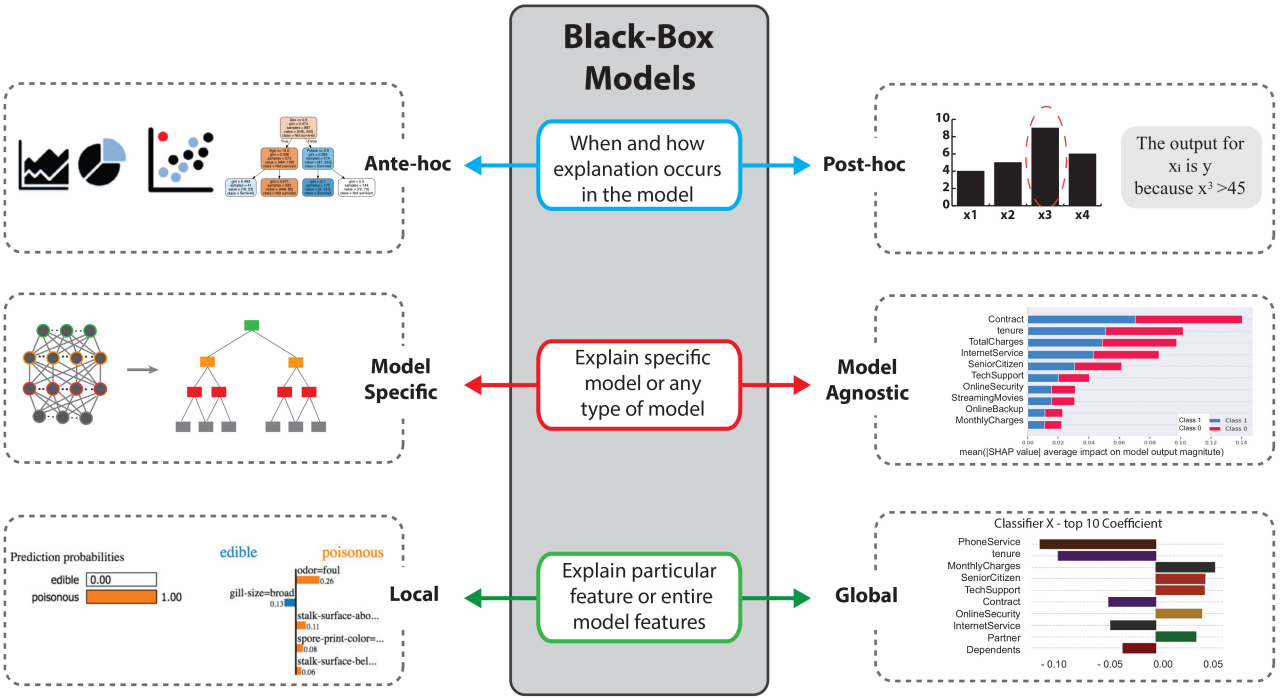

Fig. 2. Taxonomy of XAI in different perspectives.

give intrinsic explainability about the inner process of the structure.

2) Post-hoc methods, on the other hand, are applied externally after the training process of decision systems. In addition, post-hoc methods can use to supplement ante-hoc methods for providing additional information.

2) Model-Specific Versus Model-Agnostic: XAI can be grouped based on applying to specific model classes, which are model-specific and model-agnostic explanations.

1) Model-specific methods are typically designed for one type of model such as Deep Neural Network which is a well-known example of black-box models with a superior prediction performance despite its complex and opaque structure. The disadvantage of model-specific explanations is that there is a limitation in determining a model when the need for a particular type of explanation.

2) Model-agnostic methods can be applied to any type of model without limiting the model classes. It separates the explanation and the model class. Therefore, the explanations become independent of the model type.

3) Local Versus Global: According to the scope of the model, explanations of the decision models can be performed locally or globally.

1) Local explanation describes why and how certain predictions can be generated on a local level. It concentrates on ensuring interpretability by examining single or multiple instances. It is applied especially when predictions are linearly dependent on some features rather than complex dependence on all features.

2) The goal of global explanation is to characterize the model in its entirety. It seeks a global understanding of which features are more significant and what kinds of relationships are possible between them.

C. XAI Methodologies

With the XAI concept gaining popularity in AI, different XAI methods have begun to be developed. In literature, there are different XAI methods to help explain and interpret black-box models. These methods adopt different approaches and provide different interpretations [31]. Here, we classify them into five different groups as seen in Fig. 3.

1) Visual Explanation: Visual explanations cover a set of methods to examine the relationships between input and output or among input attributes, which allows users to understand the contributions of each input to the output. In case the feature set is small, interpreting these correlations is easier for users. Both local and global explanations are supported visually. However, as the feature set becomes large, visual explanations can fail to visualize correlations properly, causing users to misinterpret them. Partial dependence plot (PDP) [32] provides global explanations and reveals the dependency between the input features and output. Individual conditional explanation (ICE) [33] plot is a more recent method that is similar to PDP. PDP calculates the mean over the marginal distribution, whereas ICE keeps the entire distribution. Accumulated local effects (ALEs) [34] plots take the averages of changes in predictions and accumulate them on local grids. Also, ceteris paribus (CP) [35] plots and Breakdown plots [36] are employed to visualize the influence of features on the model prediction for a specific data instance.

2) Feature-Based Explanation: Feature-based approaches, like visual explanation methods, aim to assess the contribution of features to the model prediction. Furthermore, they consider some factors like type, robustness, and comprehensibility. Explanations can be local or global. Similarly, it can be model-specific or model agnostic. The Shapley

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

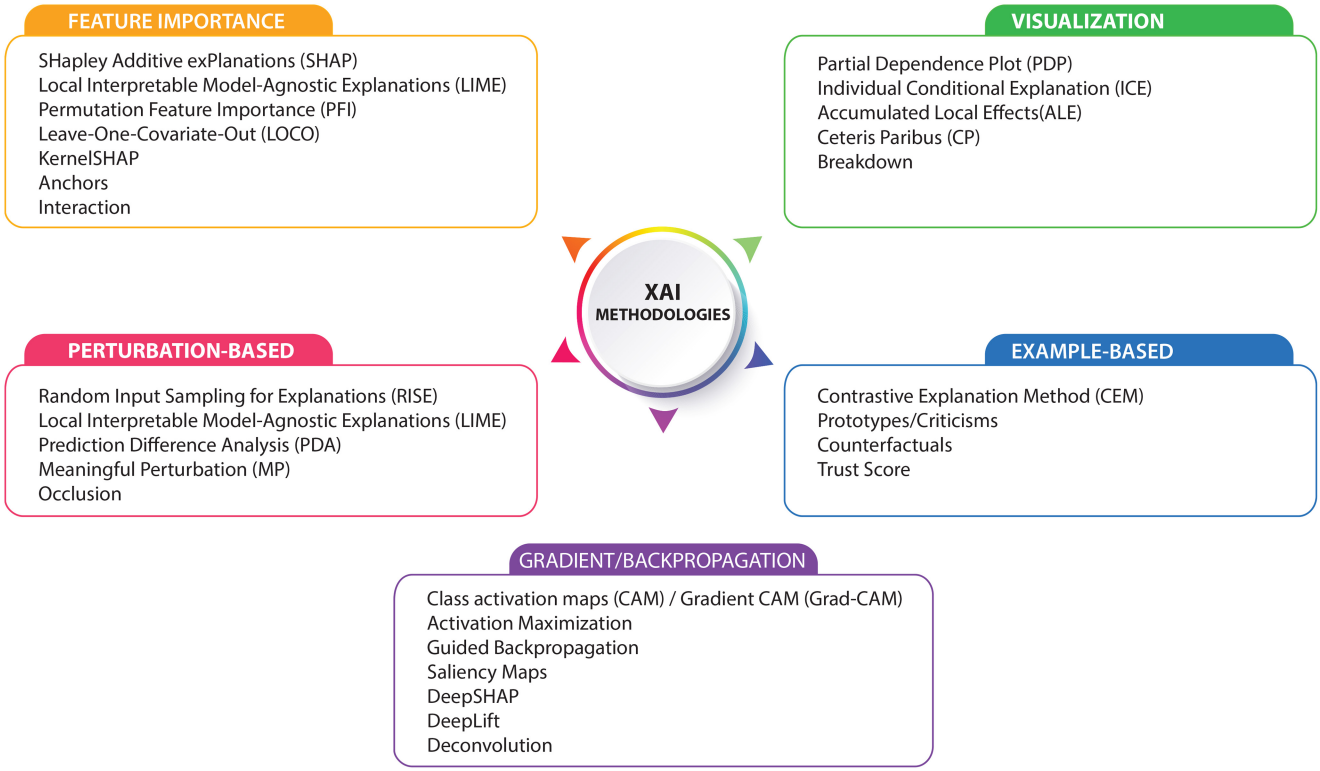

Fig. 3. XAI methodologies in IoT domain.

value is a game-theoretic approach to determine the feature importance of the model. Then, Shapley additive explanations (SHAPs) [37] method uses Shapley values and it is proposed for local and global explanations. Also, KernelSHAP [37] is proposed to overcome the conditional expectations issues of SHAP and approximates the calculation of SHAP. Also, local and global surrogate models attempt to explain the prediction of the model. While local surrogate models like local interpretable model-agnostic explanations (LIMEs) [38] focus on explaining the individual data instances, global surrogate models focus on the entire model. Anchors [39] have also the ability of local explanations by extracting a set of if-then rules. Feature Interaction [40] is used for identifying the interaction effect of the pair of feature–output or feature–feature. Permutation feature importance (PFI) [41] is a global explanation approach based on the idea of shuffling the values of unimportant features does not increase the prediction error. Also, leave-one-covariate-out (LOCO) [42] is another method that involves dropping each variable one at a time, retraining the model, and comparing the following model error to a baseline model that consists of all features.

3) Example-Based Explanation: Example-based explanations consist of model-agnostic methods that aim to explain the global or local behavior of a model or its underlying data distribution over certain data instances. Different from visual explanation or feature importance, they assist users to build compact models of the decision model. Counterfactuals [43], as one of the popular methods using example-based explanations, adopt the thinking of what could be done differently in order to get a different outcome. Accordingly, counterfactuals focus on locally explaining the change in the outcome due to changing details of input. The contrastive explanation method (CEM) [44] is another method that seeks for explanations why predictions differ from one another. Prototypes/Criticisms [45] is another strategy that attempts to identify samples in data that are highly representative or not, respectively. In addition, KNN [46] and Trust Score [47] can be also applied for local predictions.

4) Perturbation-Based Explanation: Perturbation-based methods explain a black-box model by iteratively probing with different variations of the inputs. Annotating, blurring, scrolling, and masking are some examples of distortions. It can be done at the feature level by substituting zero or random counterfactual samples for particular features, or by grouping a collection of pixels (superpixels) [48]. Occlusion [49] is a simple local method by perturbing an instance in which the input features of an instance are systematically replaced with a constant value, usually zero. Also, random input sampling for explanations (RISEs) [50] probes a model with randomly masked portions of the input instance as a generalized kind of occlusion. As a classifier explanation method, prediction difference analysis (PDA) [51] assigns a significance value to each feature concerning each class and calculates the importance of a feature by observing how prediction changes when the value of a feature is uncertain. Occlusion is a local approach for perturbing an instance in which the input features of an instance are systematically substituted with a constant, generally zero. Meaningful perturbation (MP) [52] is a local explanation approach for neural classifier

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

predictions based on a framework of meta-predictors. These meta-predictors have been trained to predict whether input features are present or not. Their prediction error is a metric for how accurate the explanation is.

5) Gradient/Backpropagation: Instead of perturbation-based methods that focus on variations of inputs, gradient methods focus on the information flow. They use the information flow during backpropagation to determine the relationship between input features and output. Gradient-based methods often use heatmaps of neurons or feature attributions to provide visual explanations. Activation Maximization [53] is one of the examples of gradient methods that use also visual explanations. Saliency Maps [54] and DeepLift [55] can be applied for local explanations. Also, DeepSHAP [56] exploits compositional architecture to provide computational efficiency by extending KernelSHAP. Also, Deconvolution [57] and Guided Backpropagation [58] are generally used for explaining CNN-based black-box models. Class activation maps (CAMs) [59] is a model-specific explanation model used for CNN. Then, CAM is generalized with a new method called Gradient-weighed CAM (GradCAM) [60] to support for more CNN architectures.

III. XAI in IoT Application Domains

In this section, we aim to provide a comprehensive review of XAI studies in IoT application domains. For this purpose, we review the studies presented in the literature and summarize them in Table I and Table II in accordance with the XAI taxonomy and methodology. We then match the application domain and methodology and present them in Table III.

A. Autonomous Systems and Robotics

In the autonomous and robotics domain, AI-powered robots can perform tasks in critical and dangerous environments that humans cannot perform, or typical applications, such as welding, ironing, painting, equipment placing/receiving, and palletizing, all of which are performed with high strength, speed, and accuracy [14]. On the other hand, XAI algorithms make robots' reasoning explainable to humans, providing a better understanding of decisions and increasing the level of trust in robots. Wang et al. [61] developed a new mechanism for robots to automatically generate explanations of their decisions based on partially observable Markov decision problems (POMDPs). They measured the performance of this mechanism in an agent-based test environment that simulates the tasks of a human–robot team. Experimental results show that robot explanations can improve task performance and improve trust and transparency. For robot explainability, the authors created a four-level natural language template, including no explanation, explanation of two sensor readings, explanation of three sensor readings, and confidence-level explanation. In this way, robots are enabled to produce text explanations based on postdoc techniques related to their decision processes.

Iyer et al. [62] proposed an explainable and object-sensitive deep reinforcement learning (DRL) model for the object recognition and classification problem. In the proposed model, the authors used a method that focuses on "object saliency maps" to provide human-intelligible visualization of DRLs states and actions. Kampik et al. [63] used text-based explainability approaches to explain the human-like actions of autonomous agents in human–robot interaction scenarios. The authors demonstrate the impact of the developed approaches on human participants through a study based on human–robot ultimatum games.

Guo [64] proposed the double dueling deep Q-learning neural network (DDDQN) network for the Quality of Service (QoS), Quality-of-Experience (QoE), and energy optimization of the UAVs in the UAV-supported 5G network. In the study, partial explainability is provided by extracting the state weights of the proposed DDDQN network.

B. Energy Management

AI-powered energy decision systems used in smart cities, smart grids, and smart home applications provide decision-makers with critical predictions about energy use and emissions. However, decision-makers want to know and believe the parameters that affect the model outputs in addition to the predicted results. Kim and Cho [65] proposed an energy demand prediction model for smart environments. The authors use an explainable autoencoder-based deep learning model to predict energy demands in various environment states. The proposed model produces consumption estimates of 15, 30, and 45 min according to defined states in the environment. The developed model has been tested on a home electrical energy consumption data set containing five-year data. The model predicts energy demands by taking the eight features (date, global active power, global reactive power, global intensity, voltage, submetering 1/2/3) in the environment as inputs. The authors provide more explainable information by visualizing monthly electricity demands with the t-SNA algorithm. Experimental results show that the proposed model achieves the best performance compared to other ML models (Decision Trees, MLP, LSTM, etc.). Sirmacek and Riverio [66] proposed two new algorithms (ML-based, computer vision-based) to predict real-time occupancy in smart office spaces by using very low-resolution heat sensors data. The authors used the well-known approach SHAPs to extract the contribution of features to the classification in the first algorithm and to reveal the contribution of each local pixel in the second algorithm. They showed that proposed algorithms can be used in many application areas related to the automation and efficient use of offices, spaces, and buildings. In another study, Zhang et al. [67] proposed an interpretable thermal comfort system to control and manage the level of comfort in smart buildings. In the proposed system, the authors used the SHAP method to explain and interpret logistic regression and decision tree models. Amiri et al. [68] proposed an artificial neural network-based transportation energy model to predict urban transportation energy. They used LIME for model transparency and explainability.

C. Environmental Monitoring

Environmental applications in IoT focus on many issues and problems in areas, such as smart cities, air monitoring,

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF XAI STUDIES BY IOT APPLICATION DOMAIN

IoT Domain	Reference	Year	ML/DL Model(s)	XAI Model Name	XAI Taxonomy		
					Ante-hoc/ Post-Hoc	Model Specific / Model Agnostic	Local / Global
Autonomous Systems and Robotics	Iyer et al. [62]	2018	DRLN	Saliency Map	Post-hoc	Agnostic	Local
	Guo et al. [64]	2020	DRL	-	Ante-hoc	Specific	Local
Energy Management	Kim et al. [65]	2019	Autoencoder	t-SNE	Post-hoc	Agnostic	-
	Sirmacek et al. [66]	2020	CatBoost	SHAP	Ante-hoc	Agnostic	Local
	Zhang et al. [67]	2020	LR, DT	SHAP	Post -hoc	Agnostic	Local
	Amiri et al. [68]	2021	NN	LIME	Post-hoc	Agnostic	Local
Environment	Kalamaras et al. [69]	2019	ARIMA, RF	SHAP	Post-hoc	Agnostic	Local
	Diallo et al. [70]	2020	CNN	Integrated Gradients	Post-hoc	Agnostic	Local
	Graham et al. [71]	2020	LSTM, XGBoost	SHAP	Post-hoc	Agnostic	Local
	Ryo et al. [72]	2021	SDM	LIME	Post-hoc	Agnostic	Local
	Qi et al. [73]	2022	XGBoost	SHAP	Post-hoc	Agnostic	Local
Finance	Sachan et al. [74]	2019	-	Belief-Rule-Base	-	-	-
	Bussman et al. [75]	2020	XGBoost	Shapley Values	Post-hoc	Agnostic	Global
	Gramegna et al. [76]	2020	XGBoost	SHAP	Post-hoc	Agnostic	Local
	Gite et al. [77]	2021	LSTM-CNN	LIME	Post-hoc	Agnostic	Local
Healthcare	Chittajallu et al. [78]	2019	ResNet50	XAI-CBIR	Post-hoc	Agnostic	-
	Hossain et al. [79]	2020	ResNet50, Deep tree, Inception v3	LIME, GradCAM	Post-hoc	Agnostic	-
	Monroe et al. [80]	2020	CNN	HihO	Post-hoc	Agnostic	Global
	Pnevmatikakis et al. [81]	2020	RF, DNN	SHAP	Post-hoc	Agnostic	Local
	Hatwell et al. [82]	2020	AdaBoost	Ada-WHIPS	Post-hoc	Agnostic	Local
	Dave et al. [83]	2020	XGBoost	LIME, SHAP	Post-hoc	Agnostic	Local, Global
	Gozzi et al. [84]	2022	CNN	Grad-CAM, SHAP	Post-hoc	Specific	Global
Industrial	Rehse et al. [85]	2019	DNN	-	Post-hoc	Agnostic	Local, Global
	Chen and Lee [86]	2020	CNN	Grad-CAM	Post-hoc	Agnostic	Local
	Sun et al. [87]	2020	CNN	CAM	Post-hoc	Agnostic	Local
	Serradilla et al. [88]	2020	RF	LIME, ELI5	Post-hoc	Agnostic	Local, Global
	Senoner et al. [89]	2021	DT	SHAP	Post-hoc	Agnostic	Local
	Mehdiyev et al. [90]	2021	DNN	Surrogate Decision Trees	Post-hoc	Agnostic	Local
	Brito et al. [91]	2022	kNN, CBLOF	SHAP	Post-hoc	Agnostic	Local
Security and Privacy	Wang et al. [92]	2016	One-vs-all classifier Multiclass classifier	SHAP	Post-hoc	Agnostic	Local, Global
	Parker et al. [93]	2019	SAE	DEMISe	Ante-hoc	Agnostic	Local
	Saharkhizan et al. [94]	2020	LSTM	DT	Post-hoc	Specific	Global
	Khan et al. [95]	2021	CNN, AE-LSTM	LIME	Post-hoc	Agnostic	Local
	Sarhan et al. [96]	2021	DFF, RF	SHAP	Post-hoc	Agnostic	Local
	Mahbooba et al. [97]	2021	DT	-	Ante-hoc	Specific	Global
	Nascita et al. [98]	2021	BiGRU	DeepSHAP	Post-hoc	Agnostic	Global
	Zolanvari et al. [15]	2021	ANN	TRUST	Ante-hoc	Agnostic	Local
	Rabah et al. [99]	2021	SVM, MLP, kNN, DT, RF, ET	LIME, Permutation	Post-hoc	Agnostic	Local, Global
	Gorzalczany and Rudzinski [100]	2021	Multiobjective Evolutionary Optimization Algorithm	Fuzzy Rule Base	Post-hoc	Specific	-
	Houda et al. [101]	2022	DNN	RuleFit, LIME, SHAP	Post-hoc	Agnostic	Local, Global
	Houng et al. [102]	2022	VAE, SVDD	SHAP	Post-hoc	Agnostic	Local, Global
	Le et al. [103]	2022	DT, RF	SHAP	Post-hoc	Agnostic	Local, Global
Smart Agriculture	Kundu et al. [104]	2021	Custom-Net	Grad-CAM	Post-hoc	Agnostic	Local
	Viana et al. [105]	2021	RF	LIME, PDP	Post-hoc	Agnostic	Local, Global
	Garrido et al. [106]	2022	ANN	DT	Ante-hoc	Specific	Global

water management, climate monitoring, etc. In solving the addressed problems, mostly AI models with high complexity and low interpretability are used. Therefore, interpretable and more transparent models are needed in these areas that affect human and environmental health. In this context, there are some current efforts to solve the existing problems. Barrett-Powell et al. [107] proposed a situational understanding explorer (SUE) platform for coalition situational understanding research that highlights capabilities in XAI for event processing in a dense urban terrain setting. Kalamaras et al. [69] proposed a new visual analytics platform that includes architecture and components created for air pollution monitoring within the scope of the AI4IoT project. In this context, there are two component designs specifically designed to provide an explanation of AI models. The first is the Annotated Line Chart, which visualizes the parameters of an ARIMA prediction model, the second is the SHAP Chart provides insight into the most important features used for Random Forest regression. Graham et al. [71] developed an environmental bio-sensing platform called Dynomics to evaluate patterns in transcriptional data at a genome scale by using XGBoost, LSTM classifiers, and XAI algorithms. They used the SHAP algorithm to gain insight into the features used by the predictive algorithms trained on transcriptional data. Thakker et al. [108] proposed a flood monitoring application using semantic Web technologies for the flood problem

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

TABLE II
SUMMARY OF XAI STUDIES ACCORDING TO THEIR METHODOLOGIES

Explanation Type/ Method	Reference	Problem Type	ML/DL Model(s)	Data Type		Dataset	Data Provider
				Input Data	Output Data		
Example-based	Zolanvari et al. [15]	Classification	ANN	Numerical	Numerical	WUSTL-IIoT	-
Feature Importance	Guo et al. [64]	-	DRL	Pictorial	Numerical	Base station data from London, Geo-tagged tweets	-
	Sirmacek et al. [66]	Classification	CatBoost	Pictorial	Numerical	Heat sensor data	-
	Kalamaras et al. [69]	Regression	ARIMA, RF	Time-series	Numerical	Pollution, weather, and traffic data	Norwegian Environment Agency Norwegian Meteorological Institute Norwegian Road Authorities
	Ryo et al. [72]	Regression	SDM	Numerical	Numerical	-	Zenodo Digital Repository
	Graham et al. [71]	Classification	LSTM, XGBoost	Numerical	Numerical	Dynomics data	University of California San Diego Biodynamics Laboratory
	Bussman et al. [75]	Classification	XGBoost	Numerical	Numerical	Credit scoring data	European External Credit Assessment Institution
	Pnevmatikakis et al. [81]	Classification	RF, DNN	Categorical	Numerical	RWD	Samsung Health
	Rehse et al. [85]	Classification	DNN	Numerical & Categorical	Textual	Sensor and production data	The DFKI-Smart-Lego-Factory
	Gramegna et al. [76]	Classification	XGBoost	Numerical & Categorical	Numerical	-	-
	Senoner et al. [89]	Regression	DT	Numerical	Numerical	Transistor chip production data	HitachiABB
	Serradilla et al. [88]	Regression	RF	Time-series	Numerical	-	-
	Brito et al. [91]	Classification	kNN and CBLOF	Numerical	Numerical	Bearing dataset, Gearbox dataset, Mechanical fault dataset	-
	Sarhan et al. [96]	Classification	DFF, RF	Numerical	Numerical	CSE-CIC-IDS2018, BoT-IoT, ToN-IoT	University of New Brunswick, Intelligent Security Group UNSW Canberra, Australian Defence Force Academy
	Amiri et al. [68]	Classification	NN	Numerical & Categorical	Numerical	Household Travel Survey (HTS) data	-
	Le et al. [103]	Classification	DT, RF	Numerical & Categorical	Visual	NF-BoT-IoT-v2, NF-ToN-IoT-v2, IoTDS20	-
	Kundu et al. [104]	Classification	Custom-Net	Pictorial	Visual	Imagery and Parametric data	Indian Council of Agricultural Research, All India Coordinated Research Project
	Viana et al. [105]	Classification	RF	Numerical	Visual	Agricultural Parcel-data	-
	Wang et al. [92]	Classification	One/Multiclass classifiers	Numerical & Categorical	Numerical	NSL-KDD dataset	Canadian Institute for Cybersecurity
	Zhang et al. [67]	Regression	LR and DT	Numerical	Numerical	-	-
	Qi et al. [73]	Classification	XGBoost	Numerical & Categorical	Numerical	Driver Behavior, Weather and Congestion data	AutoNavi Navigation Company
	Parker et al. [93]	Classification	SAE	Numerical	Numerical	AWID-CLS dataset	University of the Aegean
	Rabah et al. [99]	Classification	SVM, MLP, kNN DT, RF, ET	Numerical	Numerical	Real Traffic Data	UCI Machine Learning Repository
	Gorzalczany and Rudzinski [100]	Classification	Multiobjective Evolutionary Optimization Algorithm	Numerical	Numerical	MQTT-IoT-IDS dataset	IEEE Dataport
	Houda et al. [101]	Classification	RF	Numerical	Numerical	NSL-KDD and UNSW-NB15 dataset	Canadian Institute for Cybersecurity
	Huong et al. [102]	Classification	VAE, SVDD	Numerical	Numerical	NSL-KDD dataset	Canadian Institute for Cybersecurity
	Dave et al. [83]	Classification	XGBoost	Numerical	Numerical	Heart Disease Dataset	UCI Machine Learning Repository
Feature Importance and Perturbation	Gite et al. [77]	Classification	LSTM-CNN	Textual	Visual	News Headlines dataset, Yahoo Finance dataset	Pulse, Yahoo
	Khan et al. [95]	Classification	CNN, AE-LSTM	Time-series	Numerical	Real-world gas pipeline system data	-
Feature Importance and Gradient/Backpropagation	Gozzi et al. [84]	Classification	CNN	Pictorial	Visual	EMG data	-
Gradient/Backpropagation	Iyer et al. [62]	Classification	DRLN	Pictorial	Visual	MS. Pacmann game screenshots	-
	Diallo et al. [70]	-	CNN	Time-series	Numerical	Space Suttle Marotta valve dataset	-
	Chen and Leeok [86]	Classification	CNN	Pictorial	Visual	Bearing dataset	Case Western Reserve University (CWRU)
	Sun et al. [87]	Classification	CNN	Pictorial	Visual	Base-excited cantilever beam dataset and Water pump dataset	-
	Nascita et al. [98]	Classification	BiGRU	Numerical	Numerical	MIRAGE-2019	-
	Saharkhizan et al. [94]	Classification	LSTM	Time-series	Numerical	-	-
Perturbation and Visualization	Hossain et al. [79]	Classification	ResNet50, Deep tree, Inception v3	Pictorial	Visual	-	-
Rule Extraction	Hatwell et al. [82]	Classification	AdaBoost	Numerical & Categorical	Rules	Breast cancer, Cardiotocography, Diabetic retinopathy, Cleveland heart, Mental health survey 14/16, Thyroid, Hospital readmission, Understanding society	UCI Machine Learning Repository
	Mehdiyev et al. [90]	Classification	DNN	Numerical & Categorical	Numerical	Real-life process log data	Volvo IT Belgium
	Garrido et al. [106]	Regression	ANN	Numerical	Numerical	The climate data	-
Transparent	Sachan et al. [74]	Classification	-	Numerical	Textual	Credit data	Credit bureau server
	Mahbooba et al. [97]	Classification	DT	Numerical & Categorical	Rules	KDD benchmark dataset	The UCI KDD Archive Information and CS University of California, Irvine
Visualization	Kim et al. [65]	Regression	Autoencoder	Numerical	Numerical	Household electric power consumption	UCI Machine Learning Repository
	Chittajallu et al. [78]	Classification	ResNet50	Pictorial	Visual	Cholec80	-
	Monroe et al. [80]	Classification	CNN	Pictorial	Visual	PPMI RD	-

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

TABLE III
XAI METHODOLOGY MAPPING BY IOT APPLICATION DOMAIN

	Autonomous Systems and Robotics	Energy Management	Environment	Finance	Healthcare	Industrial	Security and Privacy	Smart Agriculture
Transparent/Rule Extraction				[74]	[82]	[90]	[97]	[106]
Example-based							[15]	
Feature Importance	[64]	[66], [68], [67]	[69], [71], [72] [73]	[75], [76], [77]	[81], [83], [84]	[85], [86], [87], [88], [91], [89]	[95], [96], [103] [92], [101], [102] [93], [99], [100]	[104], [105]
Gradient/Backpropagation	[62]		[70]		[84]		[94], [98]	
Perturbation				[77]	[79]		[95]	
Visualization		[65]			[78], [79], [80]			

in smart cities. In this application, the authors created an explainable hybrid image classification model by combining CNN-based DL models and semantic techniques. CNN was used to determine the object coverage proportion in the drainage and gully images obtained from critical geographic areas, whereas semantic techniques have been used to define the relationship between the coverage level and the objects. In this work, the authors preferred rule-based explainability in image classification. Diallo et al. [70] proposed a CNN algorithm that aims to reduce the problem of adaptation space and can be used in an on-campus smart environment monitoring platform. The authors aim to design a reliable system using XAI in the learning and prediction process of the proposed algorithm. In another smart city application, Qi et al. [73] proposed an IML framework based on the XGboost algorithm to predict and analyze the traffic order levels on urban expressways. The authors used the SHAP method to interpret the XGboost algorithm results and to evaluate the relationships between the factors affecting the result and the traffic order. Ryo et al. [72] performed an XAI-based animal species distribution model (SDM) analysis over time in terms of ecology, biogeography, and conservation biology. In the work, the authors demonstrated that XAI can be used to improve the interpretability of SDMs by performing the distribution analysis of African elephants with the LIME model.

D. Financial System

AI models are rapidly changing the way the financial system works, providing cost savings as well as operational efficiency in areas, such as asset management, investment advisory, risk forecasting, and lending and customer service [109]. However, due to the nature of the financial domain, AI decisions in these applications also contain risks that can be costly due to their consequences. For this reason, XAI should be considered among the priority issues in the financial field. Sachan et al. [74] developed a belief-rule-based (BRB) explainable decision support system to automate the process of lending loans. The authors aimed to reveal the chain of events that clarified the decision process by adding a structure containing factual and heuristic rules to the traditional IF-THEN rule-based system. Bussmann et al. [75] proposed an XAI model that can be used to measure the risks arising in loan purchases. The proposed model aimed to design a system that can explain the credit score of borrowers and predict their future behavior. The explainability of the proposed model is provided by using the TreeSHAP method, which provides explanations based on Shapley values. Gramegna and Giudici [76] proposed a technological insurance model for the insurance industry that allows an understanding of the purchase and cancelation behavior of customers. The authors used the XGBoost algorithm for the extraction of the behavior patterns and SHAP for the model agnostic interpretability. Gite et al. [77] proposed a stock price prediction moded based on LSTM and CNN. With the proposed model, it is aimed that investors can learn how and when stock prices fall or rise and make decisions accordingly. In this study, the interpretation and explainability of the model outputs were carried out using LIME.

E. Healthcare

Decision making in the healthcare domain affects people directly, and it is very difficult to compensate for their negative consequences. Therefore, AI models used in this field should not only perform well but also be reliable, transparent, interpretable, and explainable. Especially in IoT applications that provide monitoring, diagnosis, and health advice, this need should be addressed as a priority. There are many studies for this purpose. For example, Chittajallu et al. [78] proposed a human-assisted XAI system called XAI-CBIR that enables content-based image retrieval for use in surgical education. In XAI-CBIR, the CNN-based DL model lists similar pictures by extracting the semantic descriptors of the image in the query video. It iteratively trains itself based on relevant feedback from users. The developed system provides the explainability of the pictures similar to the picture in the query by creating a visual saliency map. Hossain et al. [79] proposed a three-tiered (stakeholder layer, edge layer, and cloud layer) smart healthcare framework that uses 5G networking to combat COVID-19-like pandemics. The framework is capable of detecting COVID-19 using chest X-ray or CT scan images, as well as features, such as social distancing, masks, and body temperature control. The authors used ResNet50, Deep Tree, and Inception v3 models in the edge layer. Interactive explainability is provided by using local interpretable model-agnostic (LIMA) on the knowledge graphs produced based on the learning parameters of these DL models. Dave et al. [83] focused on the usability of feature- and example-based XAI techniques on the heart disease data set to ensure the reliability of AI systems used in the healthcare domain. The authors showed that black box model behaviors can be explained using feature-based XAI techniques SHAP and LIME, and example-based techniques Anchors, Counterfactuals, Integrated gradients, CEM, and KernelSHAP. Hatwell et al. [82] developed a

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

new adaptive weighted high-importance path particles (Ada-WHIPS) model to make the AdaBoost model, which uses computer-assisted diagnostics in healthcare, more explainable. Ada-WHISP uses a new formulation to explain the classification of AdaBoost models with simple classification rules. Pnevmatikakis et al. [81] developed a risk assessment system for professionals in the health insurance sector. The proposed system also includes a virtual coaching system that provides the prediction of people's lifestyles and the production of applicable lifestyle recommendations. In this system, RandomForest and DNN algorithms are used to predict the lifestyle of individuals. SHAP was used for the explainability of the prediction results. Monroe et al. [80] proposed a CNN-based hierarchical occlusion (HihO) model that rapidly increases the interpretability of statistical findings in medical imaging workflows in IoT healthcare applications. The authors compared the developed method with GradCAM and (Parkinson Progression Markers Initiative) RISE methods on Parkinson's progression markers initiative (PPMI) data set. The proposed model has been shown to render 20 and 200 times faster than GradCAM and RISE models, respectively. Gozzi et al. [84] focused on the explainability of IA models used to classify hand movements based on EMG signals. The authors specifically investigated the effect of XAI models to provide improvements in the life of amputees using myo-controlled prostheses. The authors used SVM, LDA, XRT, and CNN algorithms in the classification process of hand gestures, and gradCAM and SHAP in the XAI process.

F. Industrial Domain

With Industry 4.0, AI systems in IIoT enable machines to perform tasks, such as self-monitoring, interpretation, diagnosis, and analysis autonomously in production lines and processes of manufacturing, logistics, and related industries. In IIoT, XAI methods can enable the adoption of new technologies and digital transformation and better product quality controls. Oyekanlu [110] developed LSTM-RNN-based explainable deep learning models for the prediction of time series data based on energy and electricity consumption in Industrial IoT systems. Later, he designed an IIoT system based on edge, fog, and cloud layers for the applicability of these models. Distributed osmotic computing approach is used to show how low-cost hardware can be used in the designed system. In the proposed IIoT system, data preprocessing and data quality enhancement in edge devices, and developed LSTM and RNN models are used in the fog layer. Christou et al. [111] focused on estimating the remaining useful life (RULs) of machines on production lines using the Qarma family of algorithms, which provides rule-based explainability for industry IoT applications. The authors predict the RULs of the drilling machine and the Quality Management in the Wheel Production Line in the Automotive industry based on rule-based explainability. Rehse et al. [85] focused on real-time process management in the smart lego factory for AI-based Industry 4.0 applications. In this context, they developed an RNN-based deep learning model that performs process predictions. In order to make the process outcome estimates more understandable to workers and visitors, they designed an interface that offers global and local explanations from poct-hoc explainability techniques.

In Industrial IoT, unforeseen failures can cause disruption of production processes, jeopardize employee safety and increase operational costs. Therefore, it is critical to monitor the health status of the mechanical equipment and to diagnose the failure conditions. Sun et al. [87] proposed a CNN-based DL model for machine health monitoring and automatic fault diagnosis. The authors have integrated a layer called CAMs into the proposed model, which provides a visual explanation. In this way, it provides a visual explanation of the image by localizing the damaged part with CAM without placing any sensors on the machines. Chen and Lee [86] focused on the development of explicable CNNs for classification in vibration signals analysis. The authors used gradient class activation mapping (Grad-CAM), which generates heat maps by calculating the weights of each feature map according to the classification scores, for CNN explainability. The authors also verified the model explainability with NN, ANFIS, and decision trees.

Serradilla et al. [88] developed an RF algorithm for estimating the RUL of industrial machines. ELI5 and LIME techniques were used for the local and global interpretability of the created model. Senoner et al. [89] designed a DT-based decision model to improve process quality in the manufacturing process. The designed model was tested on the transistor chip production line. The SHAP model was used in the analysis of the relationship between the parameters in the production and the quality of the production process. Mehdiyev and Fettke [90] developed a conceptual framework for predictive process monitoring that provides approaches to guide researchers and practitioners. In this context, they proposed a new post-doc explanation approach called Surrogate Decision Trees, which will enable the results of models, such as DNN, CNN, LSTM, and GAN to be understood. Brito et al. [91] developed an approach for fault detection and detection in rotating machines based on many unsupervised anomaly detections and clustering algorithms, such as KNN, SVM, histogram-based outlier score (HBOS), isolation forest (IF), and local outlier factor (LOF). The explainability of the models used in this approach is provided by SHAP and local depth-based feature importance for the IF (Local-DIFFI). The proposed approach has been tested on the bearing, chance, and mechanical failure data sets.

G. Security and Privacy

Security and privacy are of paramount importance in all applications of IoT. AI and ML-based algorithms/models are widely used for intrusion detection, anomalous traffic detection, authentication, and malware detection in IoT security systems. However, these systems need XAI techniques for interpreting model decisions, reasoning, and providing trust management.

Liu et al. [112] proposed a DNN-based framework to quickly and reliably detect time-dependent anomalous events in IoT data. Voronoi diagrams were used to explain the DNN classification model. The framework is tested on both real

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

aviation communication systems and simulation data, and its effectiveness has been demonstrated. Mahbooba et al. [97] considered the explainability of ML classifiers to improve trust management in intrusion detection systems. The study focused on the explainability method based on rule inference. In the study, they compared the accuracy and explainability of classifiers, such as decision trees, random forest, and SVM on the widely used KDD benchmark data set. The results show that the decision tree classifier has both higher accuracy and produces more interpretable rules. For the same research problem, Sarhan et al. [96] proposed an IML-based IDS to ensure security in IoT networks. The proposed system is capable of detecting different types of attacks in various network environments and has a generalizable structure. The explainability and interpretability of the ML algorithms in the proposed system were carried out through the SHAP method. In another study of the same scope, Le et al. [103] aimed to increase the intrusion detection performance of IDS systems working with ML algorithms in IoT-based security system data sets. The authors used DT and RF algorithms and used SHAP to explain and interpret the classification decisions of these algorithms.

Gorzałczany and Rudziński [100] concentrated on developing an interpretable IoT intrusion detection system using fuzzy rule-based classifiers. Specifically, the authors aimed to optimize the accuracy-interpretability tradeoff of IDSs using the well-known multiobjective evolutionary optimization approach. Parker et al. [93] proposed an interpretable intrusion detection technique called DEMISe that uses stacked autoencoder (SAE)-based feature extraction models for IoT networks. Rabah et al. [99] proposed an ML framework for botnet attack detection in IoT. In the proposed framework, kNN, SVM, MLP, and tree-based algorithms were used to detect. The authors used LIME and PFI in interpreting the results of the algorithm, thus providing support to cybersecurity experts.

Zolanvari et al. [15] developed a new XAI model called transparency based on statistical theory (TRUST) that statistically explains model outputs in AI-based systems. The authors used factor analysis in the transformation of the model inputs, and multimodal Gaussian distribution in determining the model output probabilities. The developed TRUST model was compared with the LIME model and it was emphasized that it was more successful than LIME in terms of speed and accuracy. Khan et al. [95] proposed a framework based on CNN and LSTM models for the detection of cyber threats in IIoT networks. In the proposed framework, complex attacks are examined on time series data with the sliding window (SW) technique with fixed length and classified with DL models. The authors used the LIME technique for DL models explainability. Similarly, Saharkhizan et al. [94] designed multiple LSTM models for cyber-attack detection in IoT networks. then authors used a decision tree to unify and explain the outputs of LSTM models. Nascita et al. [98] designed an architecture called MIMETIC-ENHANCED, which can perform traffic classification of mobile IoT devices based on explainability analysis. In this architecture, the authors used the bidirectional GRU model for traffic classification within the architecture and the DeepSHAP method for the explainability of the BiGRU model.

Wang et al. [92] focused on the explainability of machine algorithms used in IDSs that provide IoT network security. They proposed a framework for local and global explanations of IDS decisions based on the SHAP technique. The proposed framework aimed to increase the transparency of IDS forecasts and to assist cybersecurity employees in better understanding IDS decisions. In another study on the security domain, El Houda et al. [101] addressed the explainability of deep learning-based IDSs. In this study, the authors realized the explainability of the DNN-based IDS system using LIME, SHAP, and RuleFit methods. The authors showed that the developed framework guarantees the interpretability and transparency of its decisions against well-known IoT attacks. Huong et al. [102] proposed a distributed architecture called FedeX, which aims to detect anomalies occurring in industry control systems. They developed a hybrid federated learning model (FedVAE-SVDD) based on variational autoencoder (VAE) and support vector data description (SVDD) within the architecture. In FedeX, anomalies are detected on the edge computing infrastructure and in real time. The authors use/have used the SHAP model to visualize and explain the anomaly predictions of the FedVAE-SVDD model.

H. Smart Agriculture

Smart agriculture opened the door to sustainable and responsible agriculture by making better management decisions with AI-powered decision support systems. However, these systems face user adoption issues. For this reason, it is critical to make AI systems used in agriculture interpretable and understandable by the user. In this context, there are some studies in the literature. Tsakiridis et al. [113] proposed a decision support system called Vital that fully automates the irrigation of open fields. In the proposed system, the authors use the XAI approach to show that Vital can help conserve water and resources by making more precise decisions in irrigation management. Gandhi et al. [114] proposed a framework that allows for field water system mechanization, irrigation control, and precision farming based on a fuzzy logic approach. Authors aim to obtain the exact and the best atmosphere where a crop can easily cultivate and can give maximum yield. In this context, the proposed model is tested with different crops (barley, cotton, millets, groundnut, etc.) and soil types (clayey, red, sandy, etc.). For this purpose, the proposed Mamdani-type rule-based system takes temperature, humidity, soil moisture, and water nutrient sensor values, then it makes smart decisions. Kenny et al. [115] proposed a case-based reasoning system called PBI-CBR that predicts grass growth for dairy farmers. The authors aimed to provide users with high-accuracy decisions and post-doc explainability by using the same district and farm data in the development of the system.

Kundu et al. [104] developed an IoT-based automated data collection and classification framework for plant disease detection. Custom-Net, which is created from state-of-the-art models, such as ResNet, Inception, and VGG, was used for disease detection. The authors used GradCAM for model

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

explainability. Viana et al. [105] proposed a framework using explanatory ML to effectively plan and manage the use of wheat, maize, and olive land. In this framework, the effects of features, such as slope, soil type, and drainage density as well as socio-economic conditions were investigated. The authors used the RF algorithm as an ML model and PFI, PDPs, and LIME for explainability. Garrido et al. [106] focused on developing multivariate and explainable ML models to predict evaporative water loss in irrigated agriculture. For this purpose, the ANN model, which takes climate variables, such as soil, temperature, solar heating, and pressure as input, has been proposed. Model explainability was performed based on rule inference based on DT.

IV. Challenges, Open Issues, and Future Research Directions

There is ample evidence that humans have over-trusted AI systems in the past, and they still have a long way to go before they can fully trust them today [116]. Apart from the benefits and potential that XAI brings to the realm of complex decision systems, there are some drawbacks as well [12], [48], [117] which are listed below.

1) Within the scope of XAI, some key concepts are contradictory or imprecise. It has nonstandardized terminology. Everyone agrees, for example, that algorithmic explanations should be faithful. However, it is unclear whether the loyalty should be to the target model or to the data generation process. These conceptual dilemmas cause misunderstanding and pointless debate [117].
2) The bulk of XAI methods does not measure expected error rates. This makes it difficult to subject algorithmic explanations to severe tests, as required by any scientific hypothesis [117].
3) The absence of a quantitative measure of the completeness and accuracy of interpretable systems. When assessing the quality of interpretability of a model, it should not vary according to the field knowledge of the observer. It is necessary to determine the most appropriate objective measurement metrics [48].
4) When ML results are more explicitly stated in terms of how they behave for a specific problem, malicious people can use this information for their own benefit. An attacker, for example, can understand what information an ML algorithm utilizes to arrive at a specific result via interpretability. With this knowledge, it attempts to manipulate the ML algorithm by seeking to find minor changes in the inputs that will result in a different output result [12].
5) Some commonly used XAI models are vulnerable to adversarial attacks and this raises concerns about whether should we trust the XAI models if it is manipulated [118].
6) There is a lack of proper structure to combine multiple XAI methods with the aim of generating more complete explanations.
7) Explanations can be illogical for the nonexperts. An explanation may neglect too many features, which goes beyond the purpose of models. It explains what are important features, but does not explain what is the exact relationships between features and why they are important. The explanations are required for looking into the domain experts or some techniques to find out the correlations between the features and what they mean for the overall result.
8) Although XAI-based models can increase model confidence and transparency, these models do not guarantee that the system is trained on accurate or unbiased data sets, resulting in weaknesses in the training process, design, model, and target function. This is a lack of confidentiality which poses a security concern [14].

In light of the above challenging issues, some future directions can be listed as follows to be a guide for the related users.

1) For complete and clear explanations, model-specific methods should be used for improved fidelity since they are capable of looking into deeper the model architecture and specialized for an explanation of the inner structure about how it reaches the decision. Also, combining methods can overcome the partial explanations and provide a complete explanation for the model.
2) For easier explanations, automation of XAI models can be provided as ML does recently, performing certain tasks, such as feature selection and hyperparameter tuning. In addition, advances in technology will allow for improvements in libraries and packages, which enable a high level of explanations.
3) Trusting the results of XAI becomes a growing issue as the applications of IoT become wider. Therefore, improving the security and resiliency of XAI models can be an emerging topic to deal with. The issues, such as security, accountability, fairness, and ethics are discussed in the context of Responsible AI by the authors of [12].
4) IoT network usually poses a distributed architecture where IoT data is collected and processed by fog or cloud servers. They collaborate with AI systems. By exploiting federated learning, fog servers have the potential to explain black-box models locally and generate local decisions, while cloud servers explain models globally and generate aggregated explanations.
5) IoT has various application areas, such as wearable systems or nano-based IoT. XAI may use application-specific interfaces so that the model can understand the details of applications and generates granular explanations.

V. Conclusion

In this article, we present a detailed survey of XAI models in AI-based IoT applications. Our study provides researchers with a comprehensive perspective on the usage areas of XAI methods in the IoT domain. In this context, we first explain the usage requirements and potential benefits of XAI methods in IoT. Then, we comprehensively examine XAI studies in IoT by application areas. In addition, we summarize all the

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

studies with a broad perspective in accordance with the XAI terminology and taxonomy. We finally present innovative ideas for potential future studies by focusing on future direction and open issues.

REFERENCES

[1] C. C. Sobin, "A survey on architecture, protocols and challenges in IoT," *Wireless Pers. Commun.*, vol. 112, no. 3, pp. 1383–1429, 2020.

[2] S. H. Shah and I. Yaqoob, "A survey: Internet of Things (IoT) technologies, applications and challenges," in *Proc. IEEE Smart Energy Grid Eng. (SEGE)*, 2016, pp. 381–385.

[3] A. H. Ngu, M. Gutierrez, V. Metsis, S. Nepal, and Q. Z. Sheng, "IoT middleware: A survey on issues and enabling technologies," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 1–20, Feb. 2017.

[4] S. C. Mukhopadhyay, S. K. S. Tyagi, N. K. Suryadevara, V. Piuri, F. Scotti, and S. Zeadally, "Artificial intelligence-based sensors for next generation IoT applications: A review," *IEEE Sensors J.*, vol. 21, no. 22, pp. 24920–24932, Nov. 2021.

[5] A. Ghosh, D. Chakraborty, and A. Law, "Artificial intelligence in Internet of Things," *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 208–218, 2018.

[6] A. Kishor and C. Chakraborty, "Artificial intelligence and Internet of Things based healthcare 4.0 monitoring system," *Wireless Pers. Commun.*, vol. 127, pp. 1615–1631, Jul. 2021.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[8] S. Ali et al., "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Inf. Fusion*, to be published.

[9] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.

[10] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, 2018, pp. 80–89.

[11] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[12] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[13] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (XAI) on time-series data: A survey," 2021, *arXiv:2104.00950*.

[14] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to eXplainable artificial intelligence in industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.

[15] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "TRUST XAI: Model-agnostic explanations for AI with a case study on IIoT security," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 2967–2978, Feb. 2023.

[16] A. R. Javed, W. Ahmed, S. Pandya, P. K. R. Maddikunta, M. Alazab, and T. R. Gadekallu, "A survey of explainable artificial intelligence for smart cities," *Electronics*, vol. 12, no. 4, p. 1020, 2023.

[17] D. Das et al., "Explainable activity recognition for smart home systems," 2021, *arXiv:2105.09787*.

[18] C. C. Yang, "Explainable artificial intelligence for predictive modeling in healthcare," *J. Healthcare Inform. Res.*, vol. 6, no. 2, pp. 228–239, 2022.

[19] N. X. Hoang, N. V. Hoang, N. H. Du, T. T. Huong, K. P. Tran, and N. V. Hoang, "Explainable anomaly detection for industrial control system cybersecurity," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 1183–1188, 2022.

[20] "Sciencedirect." Accessed: Jan. 30, 2022. [Online]. Available: https://www.sciencedirect.com/

[21] "IEEE Xplore." Accessed: Jan. 30, 2022. [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp

[22] "Providing researchers with access to millions of scientific documents from journals, books, series, protocols, reference works and proceedings." Accessed: Jan. 30, 2022. [Online]. Available: https://link.springer.com/

[23] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdiscip. Rev. Data Min. Knowl. Disc.*, vol. 11, no. 1, 2021, Art. no. e1391.

[24] R. W. Andrews, J. M. Lilly, D. Srivastava, and K. M. Feigh, "The role of shared mental models in human-AI teams: A theoretical review," *Theor. Issues Ergonom. Sci.*, vol. 24, no. 2, pp. 129–175, 2023.

[25] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[26] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, "A deep learning approach based on explainable artificial intelligence for skin lesion classification," *IEEE Access*, vol. 10, pp. 113715–113725, 2022.

[27] J. J. Ohana, S. Ohana, E. Benhamou, D. Saltiel, and B. Guez, "Explainable AI (XAI) models applied to the multi-agent environment of financial markets," in *Proc. 3rd Int. Workshop, EXTRAAMAS Explainable Transparent AI Multi-Agent Syst.*, 2021, pp. 189–207.

[28] M. van den Berg and O. Kuiper. "XAI in the financial sector: A conceptual framework for explainable AI (XAI)." 2020. [Online]. Available: https://www.hu.nl/-/media/hu/documenten/onderzoek/projecten/

[29] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," 2020, *arXiv:2006.00093*.

[30] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[31] U. Kamath and J. Liu, *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Cham, Switzerland: Springer, 2021.

[32] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[33] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Stat.*, vol. 24, no. 1, pp. 44–65, 2015.

[34] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. Royal Stat. Soc. B, Stat. Methodol.*, vol. 82, no. 4, pp. 1059–1086, 2020.

[35] M. Kuźba, E. Baranowska, and P. Biecek, "pyCeterisParibus: Explaining machine learning models with ceteris paribus profiles in Python," *J. Open Source Softw.*, vol. 4, no. 37, p. 1389, 2019.

[36] M. J. Pontiveros, G. A. Solano, C. A. Tee, and M. L. Tee, "Explainable machine learning applied to single-nucleotide polymorphisms for systemic lupus erythematosus prediction," in *Proc. 11th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, 2020, pp. 1–8.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[38] O. Ribeiro, L. Gomes, and Z. Vale, "IoT-based human fall detection system," *Electronics*, vol. 11, no. 4, p. 592, 2022.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–9.

[40] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 916–954, 2008.

[41] F. Galkin, A. Aliper, E. Putin, I. Kuznetsov, V. N. Gladyshev, and A. Zhavoronkov, "Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects," BioRxiv. 2018. [Online]. Available: https://doi.org/10.1101/507780

[42] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018.

[43] D. Lewis, *Counterfactuals*. Hoboken, NJ, USA: Wiley, 2013.

[44] A. Dhurandhar et al., "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–22.

[45] P. Stock and M. Cisse, "ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 498–512.

[46] J. Chakraborty, K. Peng, and T. Menzies, "Making fair ML software using trustworthy explanation," in *Proc. 35th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, 2020, pp. 1229–1233.

[47] J. Druce, M. Harradon, and J. Tittle, "Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems," 2021, *arXiv:2106.03775*.

[48] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

[49] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," 2016, *arXiv:1612.08220*.

[50] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.

[51] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," 2017, *arXiv:1702.04595*.

[52] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3429–3437.

[53] D. Erhan, A. Courville, and Y. Bengio, "Understanding representations learned in deep architectures," Dept. dInformatique Recherche Opérationnelle, Université de Montréal/DIRO, Rep. 1355, 2010.

[54] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Workshop Int. Conf. Learn. Represent.*, 2014, pp. 1–8.

[55] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[56] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining models by propagating Shapley values of local components," in *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability*. Cham, Switzerland: Springer, 2021, pp. 261–270.

[57] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[58] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[59] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[61] N. Wang, D. V. Pynadath, and S. G. Hill, "The impact of POMDP-generated explanations on trust and performance in human-robot teams," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2016, pp. 997–1005.

[62] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, "Transparency and explanation in deep reinforcement learning neural networks," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2018, pp. 144–150.

[63] T. Kampik, J. C. Nieves, and H. Lindgren, "Explaining sympathetic actions of rational agents," in *Proc. Int. Workshop Explainable, Transparent Auton. Agents Multi-Agent Syst.*, 2019, pp. 59–76.

[64] W. Guo, "Partially explainable big data driven deep reinforcement learning for green 5G UAV," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–7.

[65] J.-Y. Kim and S.-B. Cho, "Electric energy consumption prediction by deep learning with state explainable autoencoder," *Energies*, vol. 12, no. 4, p. 739, 2019.

[66] B. Sirmacek and M. Riveiro, "Occupancy prediction using low-cost and low-resolution heat sensors for smart offices," *Sensors*, vol. 20, no. 19, p. 5497, 2020.

[67] W. Zhang, Y. Wen, K. J. Tseng, and G. Jin, "Demystifying thermal comfort in smart buildings: An interpretable machine learning approach," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8021–8031, May 2021.

[68] S. S. Amiri, S. Mottahedi, E. R. Lee, and S. Hoque, "Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption," *Comput., Environ. Urban Syst.*, vol. 88, Jul. 2021, Art. no. 101647.

[69] I. Kalamaras et al., "Visual analytics for exploring air quality data in an AI-enhanced IoT environment," in *Proc. 11th Int. Conf. Manage. Digit. Ecosyst.*, 2019, pp. 103–110.

[70] A. B. Diallo, H. Nakagawa, and T. Tsuchiya, "An explainable deep learning approach for adaptation space reduction," in *Proc. IEEE Int. Conf. Auton. Comput. Self-Org. Syst. Compan. (ACSOS-C)*, 2020, pp. 230–231.

[71] G. Graham et al., "Genome-scale transcriptional dynamics and environmental biosensing," *Proc. Nat. Acad. Sci.*, vol. 117, no. 6, pp. 3301–3306, 2020.

[72] M. Ryo, B. Angelov, S. Mammola, J. M. Kass, B. M. Benito, and F. Hartig, "Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models," *Ecography*, vol. 44, no. 2, pp. 199–205, 2021.

[73] H. Qi, Y. Yao, X. Zhao, J. Guo, Y. Zhang, and C. Bi, "Applying an interpretable machine learning framework to the traffic safety order analysis of expressway exits based on aggregate driving behavior data," *Physica A, Stat. Mech. Appl.*, vol. 597, Jul. 2022, Art. no. 127277.

[74] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, and Y. Li, "An explainable AI decision-support-system to automate loan underwriting," *Expert Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113100.

[75] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Comput. Econ.*, vol. 57, no. 1, pp. 203–216, 2021.

[76] A. Gramegna and P. Giudici, "Why to buy insurance? An explainable artificial intelligence approach," *Risks*, vol. 8, no. 4, p. 137, 2020.

[77] S. Gite, H. Khatavkar, K. Kotecha, S. Srivastava, P. Maheshwari, and N. Pandey, "Explainable stock prices prediction from financial news articles using sentiment analysis," *PeerJ Comput. Sci.*, vol. 7, p. e340, Jan. 2021.

[78] D. R. Chittajallu et al., "XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, 2019, pp. 66–69.

[79] M. S. Hossain, G. Muhammad, and N. Guizani, "Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics," *IEEE Netw.*, vol. 34, no. 4, pp. 126–132, Jul./Aug. 2020.

[80] W. S. Monroe, F. M. Skidmore, D. G. Odaibo, and M. M. Tanik, "HihO: Accelerating artificial intelligence interpretability for medical imaging in IoT applications using hierarchical occlusion," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 6027–6038, 2021.

[81] A. Pnevmatikakis, S. Kanavos, G. Matikas, K. Kostopoulou, A. Cesario, and S. Kyriazakos, "Risk assessment for personalized health insurance based on real-world data," *Risks*, vol. 9, no. 3, p. 46, 2021.

[82] J. Hatwell, M. M. Gaber, and R. M. A. Azad, "Ada-WHIPS: Explaining AdaBoost classification with applications in the health sciences," *BMC Med. Inform. Decis. Making*, vol. 20, no. 1, pp. 1–25, 2020.

[83] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable AI meets healthcare: A study on heart disease dataset," 2020, *arXiv:2011.03195*.

[84] N. Gozzi, L. Malandri, F. Mercorio, and A. Pedrocchi, "XAI for myo-controlled prosthesis: Explaining EMG data for hand gesture classification," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108053.

[85] J.-R. Rehse, N. Mehdiyev, and P. Fettke, "Towards explainable process predictions for industry 4.0 in the DFKI-smart-lego-factory," *KI-Künstliche Intelligenz*, vol. 33, no. 2, pp. 181–187, 2019.

[86] H.-Y. Chen and C.-H. Lee, "Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis," *IEEE Access*, vol. 8, pp. 134246–134256, 2020.

[87] K. H. Sun, H. Huh, B. A. Tama, S. Y. Lee, J. H. Jung, and S. Lee, "Vision-based fault diagnostics using explainable deep learning with class activation maps," *IEEE Access*, vol. 8, pp. 129169–129179, 2020.

[88] O. Serradilla, E. Zugasti, C. Cernuda, A. Aranburu, J. R. de Okariz, and U. Zurutuza, "Interpreting remaining useful life estimations combining explainable artificial intelligence and domain knowledge in industrial machinery," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, 2020, pp. 1–8.

[89] J. Senoner, T. Netland, and S. Feuerriegel, "Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing," *Manage. Sci.*, vol. 68, no. 8, pp. 5704–5723, 2021.

[90] N. Mehdiyev and P. Fettke, "Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring," in *Interpretable Artificial Intelligence: A Perspective of Granular Computing. Studies in Computational Intelligence*. Cham, Switzerland: Springer, 2021, pp. 1–28.

[91] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mech. Syst. Signal Process.*, vol. 163, Jan. 2022, Art. no. 108105.

[92] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.

[93] L. R. Parker, P. D. Yoo, T. A. Asyhari, L. Chermak, Y. Jhi, and K. Taha, "DEMISe: Interpretable deep extraction and mutual information selection techniques for IoT intrusion detection," in *Proc. 14th Int. Conf. Avail., Rel. Security*, 2019, pp. 1–10.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.

[94] M. Saharkhizan, A. Azmoodeh, A. Dehghantanha, K.-K. R. Choo, and R. M. Parizi, "An ensemble of deep recurrent neural networks for detecting IoT cyber attacks using network traffic," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8852–8859, Sep. 2020.

[95] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial IoT networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11604–11613, Jul. 2022.

[96] M. Sarhan, S. Layeghy, and M. Portmann, "An explainable machine learning-based network intrusion detection system for enabling generalisability in securing IoT networks," 2021, *arXiv:2104.07183*.

[97] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, Jan. 2021, Art. no. 6634811.

[98] A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, and A. Pescapé, "XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4225–4246, Dec. 2021.

[99] N. B. Rabah, B. Le Grand, and M. K. Pinheiro, "IoT botnet detection using black-box machine learning models: The trade-off between performance and interpretability," in *Proc. IEEE 30th Int. Conf. Enabling Technol. Infrastruct. Collaborat. Enterprises (WETICE)*, 2021, pp. 101–106.

[100] M. B. Gorzałczany and F. Rudziński, "Intrusion detection in Internet of Things with MQTT protocol—An accurate and interpretable genetic-fuzzy rule-based solution," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 24843–24855, Dec. 2022.

[101] Z. A. El Houda, B. Brik, and L. Khoukhi, "Why should I trust your IDS? An explainable deep learning framework for intrusion detection systems in Internet of Things networks," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1164–1176, 2022.

[102] T. T. Huong et al., "Federated learning-based explainable anomaly detection for industrial control systems," *IEEE Access*, vol. 10, pp. 53854–53872, 2022.

[103] T.-T.-H. Le, H. Kim, H. Kang, and H. Kim, "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method," *Sensors*, vol. 22, no. 3, p. 1154, 2022.

[104] N. Kundu et al., "IoT and interpretable machine learning based framework for disease prediction in pearl millet," *Sensors*, vol. 21, no. 16, p. 5386, 2021.

[105] C. M. Viana, M. Santos, D. Freire, P. Abrantes, and J. Rocha, "Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach," *Ecol. Indicat.*, vol. 131, Nov. 2021, Art. no. 108200.

[106] M. C. Garrido, J. M. Cadenas, A. Bueno-Crespo, R. Martínez-España, J. G. Giménez, and J. M. Cecilia, "Evaporation forecasting through interpretable data analysis techniques," *Electronics*, vol. 11, no. 4, p. 536, 2022.

[107] K. Barrett-Powell et al., "An experimentation platform for explainable coalition situational understanding," 2020, *arXiv:2010.14388*.

[108] D. Thakker, B. K. Mishra, A. Abdullatif, S. Mazumdar, and S. Simpson, "Explainable artificial intelligence for developing smart cities solutions," *Smart Cities*, vol. 3, no. 4, pp. 1353–1382, 2020.

[109] J. Danielsson, R. Macrae, and A. Uthemann, "Artificial intelligence and systemic risk," *J. Bank. Financ.*, vol. 140, Jul. 2022, Art. no. 106290.

[110] E. Oyekanlu, "Distributed osmotic computing approach to implementation of explainable predictive deep learning at industrial IoT network edges with real-time adaptive wavelet graphs," in *Proc. IEEE First Int. Conf. Artif. Intell. Knowl. Eng. (AIKE)*, 2018, pp. 179–188.

[111] I. T. Christou, N. Kefalakis, A. Zalonis, and J. Soldatos, "Predictive and explainable machine learning for Industrial Internet of Things applications," in *Proc. 16th Int. Conf. Distrib. Comput. Sens. Syst. (DCOSS)*, 2020, pp. 213–218.

[112] Y. Liu, J. Wang, J. Li, S. Niu, and H. Song, "Zero-bias deep learning enabled quick and reliable abnormality detection in IoT," 2021, *arXiv:2105.15098*.

[113] N. L. Tsakiridis et al., "Versatile Internet of Things for agriculture: An explainable AI approach," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, 2020, pp. 180–191.

[114] R. Gandhi, S. Bhardwaj, B. Sehgal, and D. Gupta, "An explainable AI approach for agriculture using IoT," in *Proc. Int. Conf. Innov. Comput. Commun.*, 2021, pp. 1–7.

[115] E. M. Kenny et al., "Bayesian case-exclusion and explainable AI (XAI) for sustainable farming," in *Proc. 29th Int. Joint Conf. Artif. Intell. 17th Pacific Rim Int. Conf. Artif. Intell. (IJCAI-PRICAI)*, 2021, pp. 1–5.

[116] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[117] D. S. Watson, "Conceptual challenges for interpretable machine learning," *Synthese*, vol. 200, no. 2, p. 65, 2022.

[118] G. Fidel, R. Bitton, and A. Shabtai, "When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2020, pp. 1–8.

İbrahim Kök received the M.Sc. and Ph.D. degrees in computer science from Gazi University, Ankara, Turkey, in 2015 and 2020, respectively.

He is currently an Assistant Professor with the Department of Computer Engineering, Pamukkale University, Denizli, Turkey. His current research interests include Internet of Things (IoT), deep learning, AI-enabled IoT, and data analytics.

Feyza Yıldırım Okay received the M.Sc. and Ph.D. degrees in computer engineering from the Graduate School of Natural and Applied Sciences, Gazi University, Ankara, Turkey, in 2013 and 2019, respectively.

She is currently working as a Research Assistant with Gazi University. Her research interests include Internet of Things, fog computing, software-defined networking, and network security.

Özgecan Muyanlı received the B.S. degree in computer engineering from Gazi University, Ankara, Turkey, in 2020. She is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Hacettepe University, Ankara.

She is working as a Software Engineer with TUSAŞ. Her research interests include artificial intelligence, data analytics, big data, and Internet of Things.

Suat Özdemir received the M.Sc. degree in computer science from Syracuse University, Syracuse, NY, USA, in August 2001, and the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in December 2006.

He is with the Department of Computer Engineering, Hacettepe University, Ankara, Turkey. His current research interests include Internet of Things, data analytics, artificial intelligence, and network security.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on February 05,2024 at 12:27:52 UTC from IEEE Xplore. Restrictions apply.