# 9

# Explainability and Interpretability Concepts for Edge AI Systems

**Ovidiu Vermesan[1], Vincenzo Piuri[2], Fabio Scotti[2], Angelo Genovese[2], Ruggero Donida Labati[2], and Pasquale Coscia[2]**

[1]SINTEF AS, Norway
[2]Università degli Studi di Milano, Italy

## Abstract

The increased complexity of artificial intelligence (AI), machine learning (ML) and deep learning (DL) methods, models, and training data to satisfy industrial application needs has emphasised the need for AI model providing explainability and interpretability. Model Explainability aims to communicate the reasoning of AI/ML/DL technology to end users, while model interpretability focuses on in-powering model transparency so that users will understand precisely why and how a model generates its results.

Edge AI, which combines AI, Internet of Things (IoT) and edge computing to enable real-time collection, processing, analytics, and decision-making, introduces new challenges to acheiving explainable and interpretable methods. This is due to the compromises among performance, constrained resources, model complexity, power consumption, and the lack of benchmarking and standardisation in edge environments.

This chapter presents the state of play of AI explainability and interpretability methods and techniques, discussing different benchmarking approaches and highlighting the state-of-the-art development directions.

**Keywords:** edge AI, AI explainability, AI interpretability, explainable AI, XAI, trustworthy edge AI.

## 9.1 Introduction

Explainability and interpretability are terms used to describe how understandable edge artificial intelligence (AI), machine learning (ML), and deep learning (DL) models provide insight into their decision-making, as their complexity and opacity otherwise make it challenging to comprehend their behaviour. This is required to get confidence that edge AI models are dependable (e.g., reliable, resilient, secure, safe), trustworthy, and adhere to ethical principles appropriate to context, while ensuring that they are minimised. It is necessary to distinguish between explainability and interpretability to help developers and users in determining an AI/ML approach meets particular use cases.

Explainability is the ability to explain the decision-making process in terms that are understandable to the end user. An explainable model provides a clear and intuitive explanation of the decisions made, enabling users to understand why the model has produced a particular result; it focuses on why an algorithm has made a specific decision and how that decision can be justified. It requires a straightforward and intuitive presentation of information using an ontology familiar to the user. It is particularly valuable and beneficial in the case of deep neural networks, where the models are difficult to interpret due to the convoluted structure and complex internal interactions.

Interpretability is the ability to understand the decision-making process of an edge AI model. An interpretable edge AI model provides clear information about the relationship between inputs and outputs. An interpretable algorithm can be explained clearly and understandably by a person. Interpretability is essential to ensure that users will trust AI models.

While there are methods to explain the behaviour of models that are not inherently interpretable, interpretability serves as a gold standard for model explainability in a direct and transparent manner.

Superior AI explainability and interpretability come at the expense of performance, as illustrated Figure 9.1 [7]. When datasets are large, and the data are related to images or text, neural networks can meet the customer's AI/ML objective with high performance. For cases where complex methods are required to maximise performance, data scientists may focus on model explainability rather then of interpretability [7].

A conceptual workflow for the design of AI models which includes both interpretability and explainability is illustrated in Figure 9.2.

Interpretability is mostly associated with model training, evaluation, and quality assurance, while the explainability is a consideration of the deployed AI model.

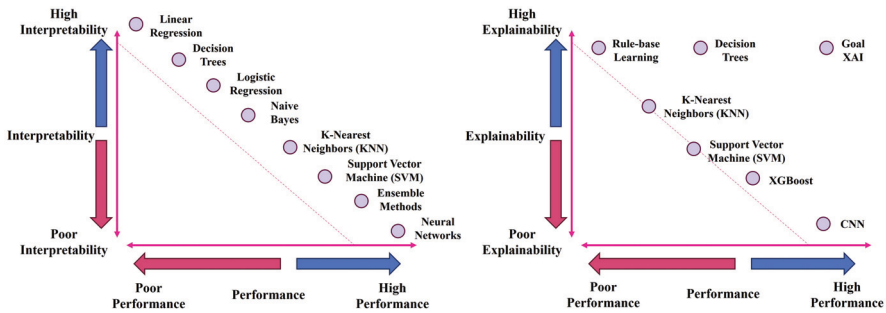


**Figure 9.1** AI Interpretability and Explainability vs Performance for Common ML Algorithms (Adapted from [7])

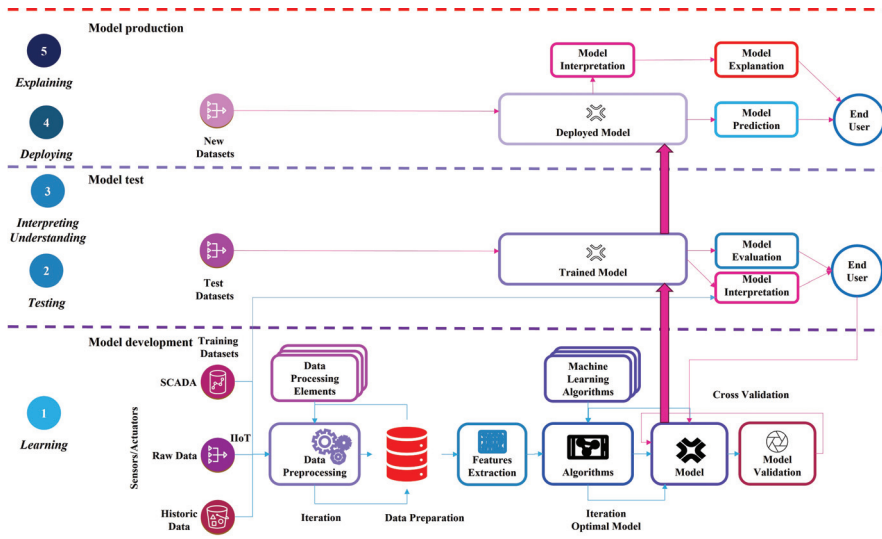


**Figure 9.2** Conceptual Workflow explainable and interpretable ML model development

The European Union's Artificial Intelligence Act (AIA) [3] addresses AI explainability and interpretability. The AIA is a comprehensive regulatory framework that promotes transparency, accountability, and the protection of individual rights in the face of AI's growing influence, aiming to ensure the ethical and responsible use of AI. A significant proportion of current AI-based software falls within the scope of the AIA.

The European Parliament has amended the AIA by introducing Article 28 b, aligned with the 2019 OECD AI Principles [11], which states that AI "should be robust, secure, and safe throughout its lifecycle so that it functions

appropriately and does not pose unreasonable safety risks." [12]. The new Article 28b features nine responsibilities for developers of foundation models. Of these nine obligations, the following three are the most relevant for AI designers;

**Risk identification** [Article 28b(2a)], which specifies that it is mandatory to identify and mitigate reasonably foreseeable risks (inaccuracy, discrimination, etc.) with the support of independent experts.

**Testing and evaluation** obliges AI providers to make adequate design choices to ensure that the foundation AI model achieves appropriate levels of performance, predictability, interpretability, corrigibility, safety, and cybersecurity. AI model functions are the building blocks for many downstream functions, so Article 28b(2c) aims to ensure that these meet the minimum standards and do not compromise systemic quality.

**Documentation** is an obligation for AI providers in the form of data sheets, model cards and intelligible use instructions. This is required to avoid that black box AI foundational models being deployed without knowing their processes or capabilities.

The documentation should include the following elements:

- A description of the data sources used in the development of the AI foundational model.
- An explanation of the capabilities and limitations of the foundational model, including reasonably foreseeable risks and the measures that have been taken to mitigate these, as well as the remaining unmitigated risks with an explanation of the motivation for which they could not be contained.
- A description of the training resources utilised by the foundation model, including the required computing power, the training time, and other relevant information related to the model's size, performance, and energy efficiency.
- A description of the model's performance based on public state-of-the-art industry benchmarking methods.
- A report and explanation of the results of relevant internal and external testing and optimisation of the model.

An overview of the responsibilities across the AI value chain according to the AIA is illustrated in Figure 9.3. The AIA provides a holistic approach to address the challenges posed by foundation models at different stages along the entire AI value chain. This approach considers that along the AI value

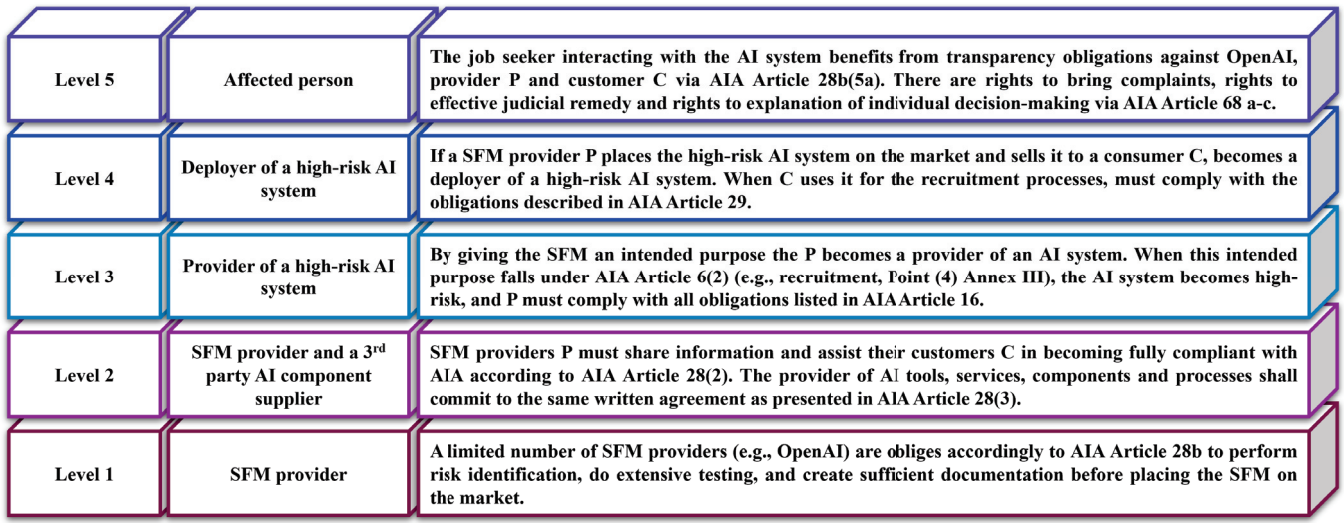


**Figure 9.3** Responsibilities Across the AI Value Chain

chain, multiple entities will supply tools, services and components, including data collection and pre-processing, model training, model retraining, model testing and evaluation, hardware/software integration. The complexity of the AI value chain requires transparency in a manner that permits traceability and explainability while making users aware that they are interacting with an AI system [3].

This chapter is organised as follows. Section 1 introduces the edge AI explainability and interpretability research area, including the proper definitions of the terms. Section 2 presents the goals of AI explainability and interpretability. Section 3 provides an overview of the state of the art of existing edge AI explainability and interpretability approaches, methods and techniques, and the actual advantages/disadvantages. Section 4 describes possible benchmarking techniques for edge AI explainability and interpretability to align with edge AI systems' trustworthiness requirements. Section 5 presents more detail on edge AI explainability and interpretability elements and specific issues. Section 6 describes the challenges, open issues, and future research directions for edge AI explainability and interpretability. Section 7 draws the conclusions.

## 9.2 AI Explainability and Interpretability Goals

Explainable and interpretable artificial intelligence enables trustworth predictive analytics, anomaly alerts, and decision-making. Data from edge devices can be analysed to predict maintenance for machines in industry and to optimise resource allocation in manufacturing. Effectively managing a distributed range of explainable systems to provide faithful computations on the data collected from edge devices is a fundamental challenge in deploying transparent edge-based AI applications. Creating effective solutions that can easily combine and accumulate decisions made by multiple models is still under development. It represents one of the key research areas to be investigated in the future [47]. Also aggregating explainability and interpretability in such composed systems represents a key challenge.

Over many years, researchers have primarily focused on enhancing model performance, relegating the intricate inner mechanisms that drive the output to a secondary analysis. Classical neural networks rely on millions of parameters (e.g., VGGNet has $\sim$138M parameters, and ResNet-152 has $\sim$60.3M parameters) [84]. Understanding the interconnections and communication pathways in these networks remains a challenging task. Furthermore, despite their remarkable performance, these models also exhibit

vulnerabilities; object detectors and classification models, for example, can be easily deceived with slight alterations to input signals using adversarial examples [44], or decisions could be based on entirely incorrect features. Gender biases and stereotypes also pose challenges for Natural Language Processing (NLP) [45].

An understanding of the underlying mechanisms driving AI-driven model results has emerged as an imperative. This understanding is also a fundamental goal for human progress and for enhancing current AI-based systems. With the advent of new methodologies and large datasets, various sectors, including finance, transportation, healthcare, and security, have adopted approaches that are not only comprehensible but also endowed with an appropriate level of trustworthiness and effective oversight. For example, medical diagnosis systems usually employ visual explanations to provide support for their decisions, increasing the classification confidence [42]. The financial sector also heavily relies on interpretable methods for extracting trends and seasonalities from historical time series data [46].

In scenarios involving the proliferation of edge devices within a system, strategies that guarantee reliability, transparency, interoperability and foundational defence against vulnerabilities and errors become imperative, particularly in critical domains. The reliability of the analytics platform becomes crucial in these application scenarios. Autonomous systems equipped with the ability to perceive, learn, and make decisions represent the fundamental trajectory of future AI-based systems. Their actions must satisfy specific requirements and be explained in critical contexts.

Domains where interpretable systems find application span a diverse spectrum, for example:

**Agriculture**: Systems adept at extracting high-level insights from satellite images and remote sensors provide invaluable farming decision support. The possibility to expound upon the derived information is pivotal for informed decision-making [38].

**Finance**: Insurance companies and banks rely on automated systems to profile clients. These systems are pivotal in evaluating loan eligibility, demanding a transparent rationale for granting or withholding loans. Clear justifications are imperative for accountability and audit [36].

**Industry and Autonomous Robots**: Deploying automated systems to prevent human injuries requires the ability to proactively prevent individuals from specific actions. These systems must operate in a manner that absolves companies of liability for any unintended or improper action [37], while allowing post event analysis of any interventions that were performed.

**Medical Diagnosis**: Classifying magnetic resonance imaging (MRI) scans or histopathological images necessitates the elucidation of outcomes and the identification of causative factors. This is crucial for ensuring accurate diagnoses and comprehensible justifications for medical conclusions and interventions [35, 42].

**Military and Security**: Territorial defence and soldier training could considerably benefit from support systems that explain actions. These systems can enhance the efficiency of achieving goals, ensuring that tactical manoeuvres and training regimens are effective and comprehensively rationalised [39].

**Recommendation Systems and Marketing**: Typical applications consist of profiling users to support marketing endeavours that augments corporate revenues and facilitates the targeted promotion of products. Transparency in explaining these attributes fosters customer engagement and strategic decision-making [40].

**Smart Cities**: Aspects such as lighting, energy management, and traffic control within smart buildings and urban infrastructures are very applicable to AI. As the number of interconnected devices increases, AI-based frameworks must explain decisions regarding different aspects of human life (e.g., water supply, waste management, governance, etc.). Addressing cybersecurity and privacy challenges with explainable and interpretable methods is crucial for smart city development [43].

In addition, the General Data Protection Regulation (GDPR) [41], which codifies regulations on information privacy in the European Union and the European Economic Area, imposes legal obligations upon developers to elucidate decisions that hold the potential for impact on individuals. Finally, systems that inspire user confidence by being unambiguous and explainable are much more likely to be positively received and well engaged with.

## 9.3 AI Explainability and Interpretability Methods and Techniques

Highly accurate models are favoured over those that offer superior explainability but diminished accuracy, given that the primary objective of a machine learning system centres on its performance. However, it is not uncommon for these systems to be viewed as opaque by human evaluators, and the interpretation of their decision-making processes is often relegated to a subsidiary investigation.

Interpretability can enhance multiple aspects of a machine learning model. It can rectify biases learned during training, ensure that only meaningful variables contribute to the output, and measure robustness against adversarial perturbations. Sectors such as healthcare, finance, and security necessitate a profound understanding of ML models to uphold equity, responsibility, and transparency principles.

AI explainability and interpretability primarily focus on two aspects of an ML system: data and model. As illustrated in Figure 9.4, exploratory data analysis and visualisation represent important tools for gaining insights from data.

Dimensionality reduction techniques, such as PCA, ICA, t-SNE, LDA, and autoencoders, are used in cases involving many variables. These techniques convert high-dimensional data into a lower-dimensional form while preserving or extracting their internal structures.

Several frameworks implement data exploration and explanation techniques to express each feature's relevance through graphs, heatmaps, and various plots. Contrastive analyses provide interpretations that study the impact of features in achieving a desired output rather than solely focusing on the outcome itself.

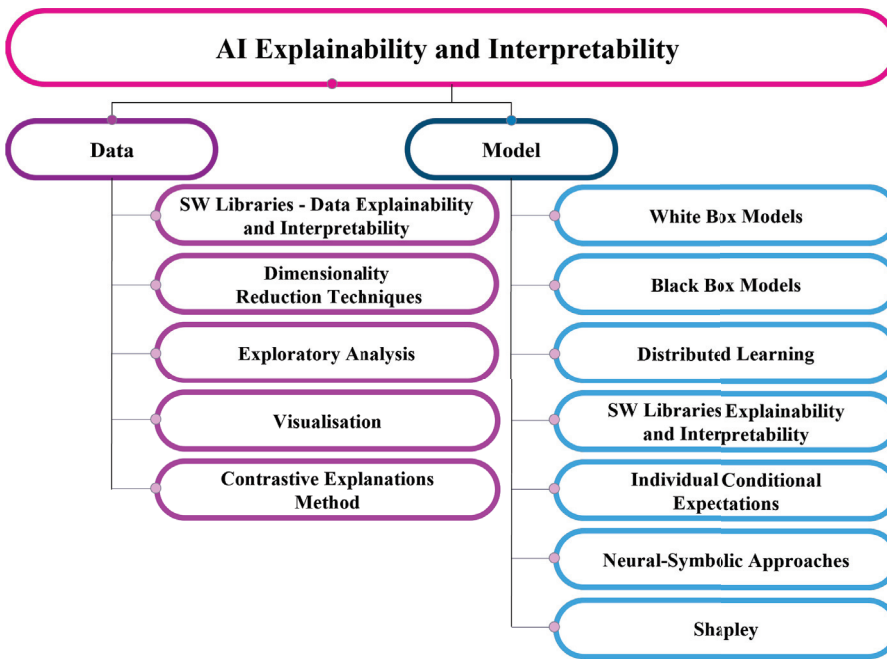


**Figure 9.4** Data and Model AI Explainability and Interpretability Classification

While data explainability provides insights into the collected data, model explainability and interpretability focuses on the techniques used to understand the models. Specifically, explainable and interpretable models are categorised into transparent surrogate models, as illustrated in Figure 9.5.

Models classified as transparent inherently offer comprehensive insight through their intrinsic design or explicit processes aligned with the input data. Logistic or linear regression, decision trees, k-nearest neighbours and rule-based methods are examples of transparent models. This characteristic is mainly owned by ante-hoc methods.

Ante-hoc techniques allow embedding explainability into a model from the beginning. Post-hoc techniques enable models to be trained normally, with explainability only included at testing time.

Generalised additive models (GAMs) [54], for example, represent one of the first classes of nonparametric interpretable models, where the impact

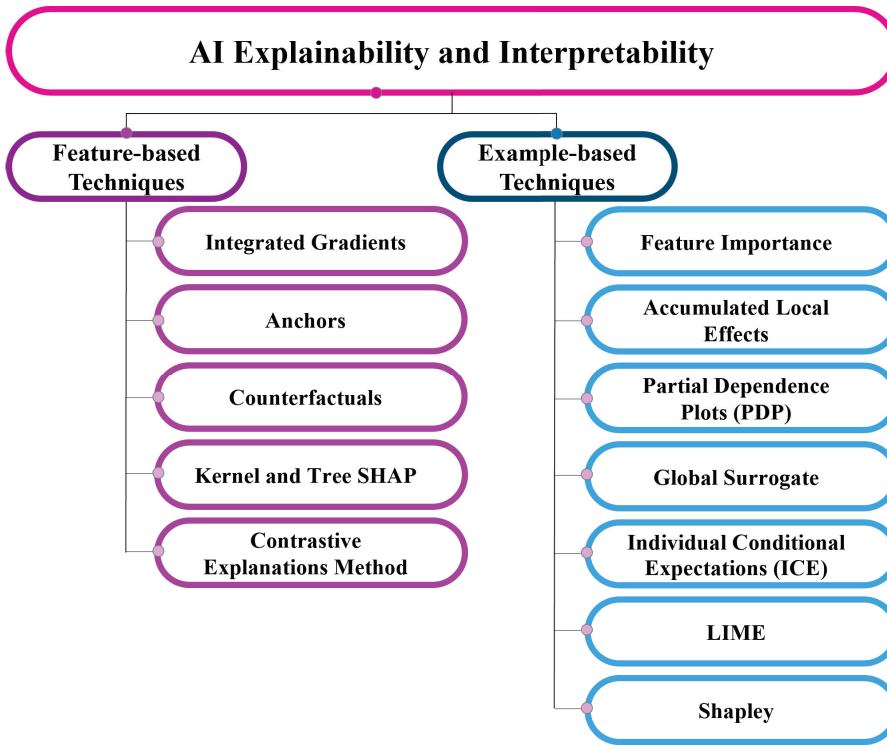


**Figure 9.5** AI Explainability and Interpretability Model Approach Classification

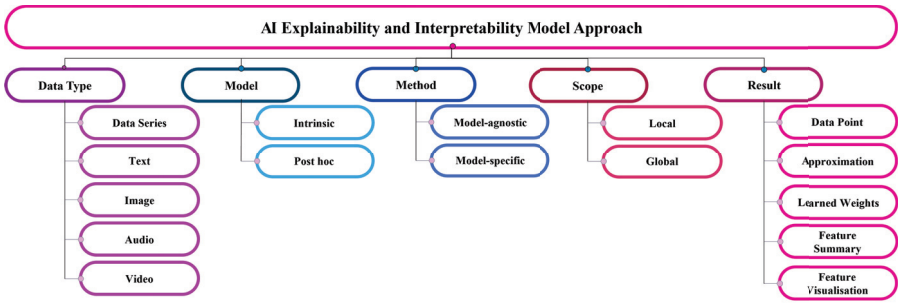


**Figure 9.6** AI Explainability and Interpretability Model-Agnostic Approach Classification

of the examined variables is captured through smooth linear (or nonlinear) functions. Being additive, the effect, or impact of each variable can be measured independently from the others. Decision trees follow a tree-based logic, where control statements switch between specific paths to uncover rules behind decisions.

While computationally cheaper to evaluate, transparent models may not fulfil the performance criteria of the task at hand. Surrogate models use approximation criteria to emulate the operative dynamics of the primary model by assimilating the input-output relationship and exploiting fidelity measures [50] to evaluate their performance.

These models present fewer challenges in interpretation. They are created post-hoc and offer more flexibility and usability compared to the models they are built on top of. Post-hoc explainability refers to models that are not inherently interpretable by design and represent a class that encompasses diverse means to increase the explainability.

Post-hoc techniques offer valuable approximations of the inner workings or information flow to produce understandable representations using graphs, rule sets, score maps, or natural language.

While model-specific techniques extract explainable representations tailored to a particular learning algorithm or the internal structure of a model, model-agnostic techniques utilise model inputs and predictions to replicate the learning mechanism and generate explanations, as illustrated in Figure 9.6 and Figure 9.7.

Among model-specific techniques, feature importance highlights the impact of each feature on the decision.

Condition-based explanation defines oriented questions to allow the model to provide possible explanations with a set of conditions.

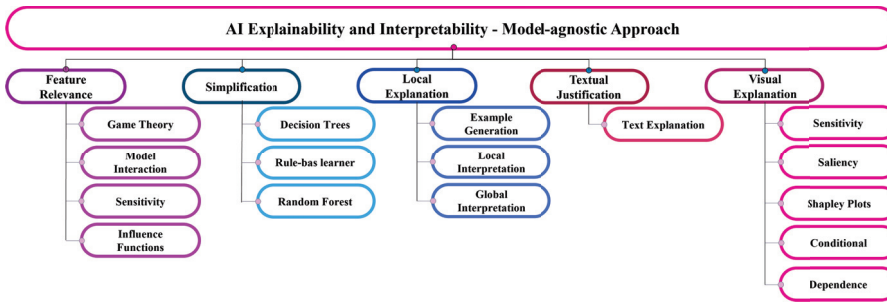


**Figure 9.7** AI Explainability and Interpretability Model-Specific Approach Classification

Knowledge distillation methods [70] or rule-based learners [71, 72] also strongly rely on the original model.

Model-specific post-hoc explainable techniques cannot be employed with arbitrary models. In this circumstance, model-agnostic techniques can be considered since they involve conducting pairwise analyses of model inputs and predictions, aiming to comprehend the learning mechanism and generate explanations. This class, which does not make any assumptions about the model, includes visualisation-based techniques [73, 74], knowledge extraction [75, 76], and influence methods [77, 78]. Knowledge extraction provides a comprehensible representation of the model. Influence methods, instead, investigate the importance or resilience of hidden units by recording signal variations within the model.

The way explanations are presented is also inextricably linked to the nature of the data under examination. For instance, saliency, or attention, maps are prevalent to explain decisions derived from visual data (popular saliency methods are GradCAM [60], DeepLIFT [61] and SmoothGrad [62]); conversely, for textual data, specific segments of text that contribute to the resultant output are typically highlighted. Moreover, a predetermined set of rules can be applied to highlight the relevance of attributes in influencing the prediction.

Visual explanations represent one of the most important classes of methods used for classification, detection, and recognition tasks. Their success can be ascribed to the immediate representation of the decisions, highlighting what region of the input images generated that specific response. The medical domain, for example, extensively relies on these approaches [69].

These methods are typically used for visually understanding convolutional neural networks (CNNs) [66, 67, 68]. Most visual explanation techniques use backpropagation-based approaches that compute partial derivatives concerning each input feature or intermediate deep neural network layers [47] [48].

Another key distinction of the explanation generation processes relies on type of extracted explanations, which are representative of instances (local) or are broadly applicable (global). In this regard, local methods investigate the output of the models for specific samples and refer to a dynamic explanation process.

In this context, Local Interpretable Model-agnostic Explanations (LIME) [55] builds a surrogate model around the sample, which is easy to explain. A trade-off between unfaithfulness and the complexity of the model allows non-experts to interpret decisions by weighing the most critical parameters. Despite there being no guarantee that the surrogate models inherit the same properties as the original model, it is model-agnostic and only requires small perturbations to the input data.

Model Agnostic Supervised Local Explanations (MAPLE) [59] is a supervised neighbourhood approach that combines local linear models and ensembles of decision trees. SHAP (SHapley Additive exPlanations) [56] is another technique, based on game theory, used to explain the predicted output by computing the contribution of each input feature to the prediction.

Shapley values could refer to individual feature values or groups of feature values. For instance, pixels can be grouped into super pixels to explain an image. This method can be used both locally and globally. Other examples are counterfactual explanations [57].

Random Forest Feature Importance [63], Quasi Regression [64] and Global Sensitivity Analysis (GSA) [65] are examples of global methods that measure the importance of the features that contributed to the prediction highlighting their overall influence.

In this context, Partial Dependence Plots (PDPs) represent a class of visualisation-based techniques that define a global method able to visualise the effect of the values of a specific feature by marginalising all the other features.

Along with t-SNE, PCA and Quasi Regression, in these techniques the explanation is directly inferred from the black box model, compared to surrogate models. These methods are categorised as illustrated in Figure 9.8.

Whilst numerous methods were developed to explain the results, criteria to assess the explainability of a model are a fundamental and active area of research since several properties, such as casualty, target's belief, or trustiness, cannot be easily formalised [57].

Complexity and sparsity represent two critical aspects of evaluating a model to define its interpretability. The Predictive, Descriptive, Relevant (PDR) framework [58] proposes three desiderata for evaluating and

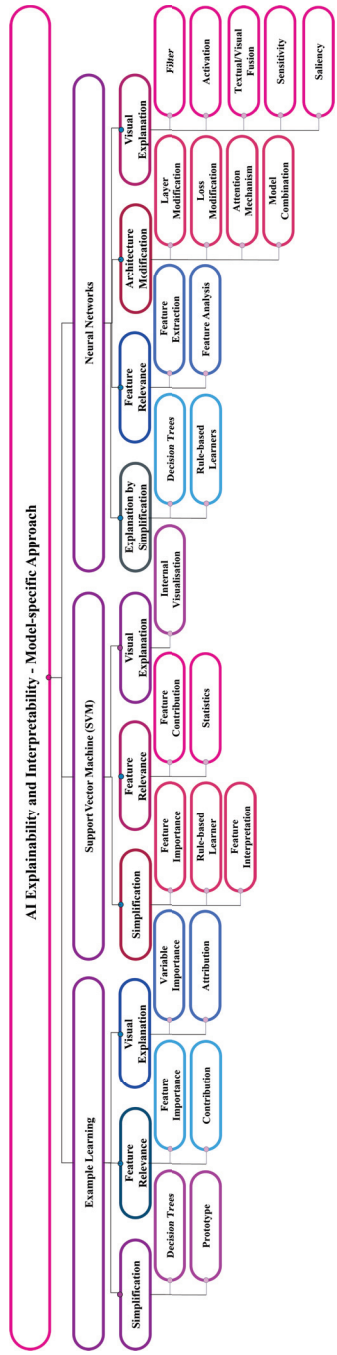


**Figure 9.8** Feature- and Example-based AI Explainability and Interpretability Techniques

constructing interpretations: predictive accuracy, descriptive accuracy, and relevancy.

## 9.4 Benchmarking

The effectiveness of interpretable and explainable AI (XAI) techniques is influenced by various factors, including the user, usage context, model type, data characteristics, and desired form of explanation. Several approaches have been introduced in the literature to analyse and measure such effectiveness, the performance, and impact of interpretable and explainable AI techniques in real-life applications. However, the definition of a standard set of measures for evaluationg the effectiveness of interpretable and explainable AI techniques is still an open research problem, and there has yet to be an agreement on standard benchmarking methods.

The lack of accords stems from the fact that a qualitative human-based evaluation of the explanation is often necessary to assess the explanation quality. Nevertheless, several research trends are oriented towards the definition of quantitative approaches, enabling an automatic measurement of interpretable and explainable AI techniques, and allowing us to effectively compare different techniques [28].

It is therefore possible to distinguish two kinds of approaches to evaluate the effectiveness of interpretable and explainable AI techniques: *i)* quantitative evaluation methods, which involve creating an objective metric or benchmark to measure explanations without human involvement and that offer the advantage of facilitating comparisons between different explanation methods; *ii)* qualitative evaluation methods, which involve humans in evaluating explanations and permit evaluating the beneficial effects of interpretable and explainable AI methods from the users' perspective.

Quantitative evaluation approaches can be classified according to different taxonomies in the literature. As an example, [28] classifies evaluation approaches according to the type of application (images classifiers generating heatmaps, and natural language processing techniques). Moreover, recent studies propose the use of synthetically generated data with known properties to quantitatively evaluate the performance of interpretable and explainable AI methods [34]. However, generating realistic synthetic data with specific properties known a priori can be challenging for real application contexts. Together with classifying quantitative evaluation approaches, some work in the literature also review the measures used to evaluate their effectiveness. For example, [30] describes the following figures of assessments:

- *Fidelity* seeks to assess the accuracy of function $f$ in emulating function $b$. Variations of fidelity exist, contingent upon the specific type of explainer being examined [31].
- *Stability* confirms that comparable instances yield consistent explanations. The assessment of stability can be accomplished using the Lipschitz constant [32].
- *Deletion* involves eliminating the features that were deemed important by the explanation method $f$, observing how the performance of $b$ deteriorates as a result. One of the deletion methods is Faithfulness [32], which seeks to confirm whether the relevance scores truly reflect significance: higher importance values are anticipated for attributes that substantially influence the ultimate prediction.
- *Insertion* employs a complementary approach to deletion. Typically, both insertion and deletion evaluations are customised for specific types of explainers: Feature Importance explainers for tabular data, Saliency Maps for image data, and Sentence Highlighting for text data.
- *Monotonicity* [33] can be viewed as a manifestation of an insertion approach. It assesses the impact of $b$ by systematically introducing each attribute in ascending order of importance. In this scenario, the anticipation is for the performance of the black-box model to progressively improve as more features are added, leading to monotonically increasing model performance.
- *Running time* is the computational time needed to provide interpretations or explanations. The running time of the technique used to explain the decisions made by the model in real time and cloud applications can be a critical factor. It is important for systems to provide interpretations or explanations in a timely manner.

Qualitative evaluation approaches can be classified according to whether they are designed to analyse explainable or interpretable AI methods. The qualitative analysis of explainable AI methods is mainly based on the statistical analysis of questionnaires submitted to human evaluation, which may be designed with different goals [29]:

- Evaluate the a priori goodness of explanations.
- Assess users' satisfaction with explanations.
- Uncover user's mental model of an AI system.
- Evaluate user's curiosity or need for explanations.
- Analyse the level of user's trust and reliance on the AI.
- Assess how the human-system work performs.

The qualitative analysis of interpretable AI methods is based on measures that can be systematised into three categories [30]:

- Functionally-grounded measures, which analyse the impact of the system in the considered application context.
- Application-grounded evaluation methods, which require evaluations performed by the set of human experts for which the system has been designed.
- Human-grounded measures, which assess interpretations using non-expert humans.

## 9.5 Edge AI Explainability and Interpretability

Integrating IoT, edge computing and AI can revolutionise how intelligent devices interact and enable a new era of innovative applications. By bringing computation, analytics, and connectivity closer to the data source, edge AI technologies reduce latency, enhance privacy, optimise bandwidth, and enable the online/offline operation.

Challenges such as limited computing resource, data quality and training, security and privacy, scalability, interoperability, ethical considerations, and explainability and interpretability must be addressed carefully. As these technology fields continue to advance, IoT, edge computing, and AI convergence are unlocking new opportunities, enabling intelligent decision-making and real-time insights at the edge.

AI at the edge extends ethical concerns about biased decision-making, algorithmic transparency, and accountability to that environment. As the number of intelligent edge devices increases, it is necessary to address ethical considerations and ensure that edge AI systems are fair, transparent, and accountable while edge AI models are explainable and interpretable. Compliance with legal regulations regarding data privacy, bias, and responsible AI usage is also crucial.

In the literature, there are only a limited number of studies on edge AI interpretability and explainability [80, 83]. Most of the work considers autonomous driving technologies [17], preventive healthcare applications [18, 80], and IoT [19].

Considering autonomous driving technologies, the study on edge AI interpretability and explainability regard different kinds of applications. There are methods for analysing images acquired from external cameras and Lidar sensors [20], and studies analysing the driver behaviour [21].

In preventive healthcare applications, interpretability and explainability techniques can detect possible health problems, as well as assist healthcare experts and family members in making critical healthcare decisions [22].

In the context of IoT devices, interpretability and explainability can be used to achieve heterogeneous goals according to the considered application scenario. For example, there are studies on edge AI interpretability and explainability for managing traffic [23], smart buildings [24], smart homes [25], environmental monitoring [26], and industrial control systems [27, 81, 82].

However, current studies on edge AI interpretability and explainability are limited to specific applications and do not propose a general approach for designing and developing interpretable and explainable AI technologies for the edge. This process is particularly challenging. In fact, developing edge AI solutions requires integrating edge AI hardware, software, AI stack building blocks techniques/methods/models and data addressed as a holistic edge AI design framework for the whole edge AI system.

Edge AI interpretability and explainability must apply to the edge AI model and data, as illustrated in Figure 9.4.

## 9.6 Challenges and Open Issues

Edge AI models are implemented and run on devices at the edge of a network, enabling real-time data processing and analysis. Edge processing is characterised by constrained computing, memory, power budget, and latency resources. Edge AI interpretability manages the extent to which a cause and effect can be observed within an edge AI system.

At the same time, explainability addresses how the internal mechanisms of an edge ML or DL system can be explained in human terms and representations. AI explainable and interpretable methods and techniques provide additional processing requirements and affect the overall performance of the AI-based systems implemented at the edge. This section presents several challenges, open issues, and future research directions that must be addressed for a successful edge AI deployment.

Edge AI model complexity vs interpretability and explainability is a challenge, considering AI decision-making must be transparent and understandable. Edge DL models are typically accurate but difficult to interpret. As a result, a trade-off between model complexity, interpretability, and explainability may be accepted. Complex models, such as edge deep neural networks (DNNs), capture convoluted patterns in data and provide prime performance. DNNs act as black boxes, making interpreting their behaviour or internal

decisions challenging. AI models, such as decision trees or linear regression, are more straightforward and interpretable but offer lower performance on complex tasks and are more difficult to create.

The open issue is how to find the optimal balance to develop AI models that are powerful and robust enough to provide accurate results and yet sufficiently simple to be understandable. In many cases, this requires hybrid approaches, developing new edge AI interpretability and explainability techniques and methods, or accepting unavoidable trade offs in either explainability/interpretability or performance. In summary, achieving interpretability and explainability comes at the expense of edge AI model deployment. Simpler models that are easy to interpret may not perform as well as their complex replicas. Balancing the demand for explanation and interpretation with the requirement for models offering high-level performance is challenging.

Edge AI deployment and the management of AI models on many edge devices can be challenging considering the integration of edge AI explainable and interpretable methods, as it could be difficult to ensure that models perform optimally across all devices. Resource-constrained edge devices can also make running complex updates or retraining models challenging. This can be a significant problem as it is essential to monitor the performance of edge AI models and their explainable or interpretable surrogate models (twins) and implement regular maintenance, upgrades, and updates to prevent model degradation.

A lack of expertise in the field of edge AI explainability and interpretability will limit the adoption and deployment of edge AI. This can comprise the technical aspects of edge AI explainability and interpretability, such as how to build and optimise efficient explainable and interpretable models for edge devices and understanding the broader ramifications of using edge AI, such as real-time processing, latency, and security concerns. A lack of expertise can make it difficult to effectively design edge AI explainable and interpretable models and utilise them in edge AI applications to meet customers' requirements. It can also make it challenging for edge AI model providers and users to adequately evaluate the potential risks and benefits of using edge AI, limiting their ability to make informed decisions about possible adoption and deployment of edge AI.

Developing and deploying edge AI is a time-consuming and costly process and implies a trade-off between explainable and interpretable features and performance. Difficulties are associated with integrating edge AI explainable and interpretable models with edge devices, especially the ones with limited resources. The complexity and time associated with deploying edge AI explainable and interpretable models is a challenge, especially

when dealing with large models, requiring extensive tuning and optimisation. Deploying, managing, and maintaining edge AI explainable and interpretable models on many edge devices is time-consuming and requires significant resources.

Updating and upgrading the edge AI explainable and interpretable models aligned with the improvements and advancements of edge AI models is essential to extend the lifetime of edge AI solutions. Adapting the features to the latest market advancements can be challenging, as edge AI solution providers must plan for incorporating the newest edge AI explainable and interpretable technology into their developments to stay competitive.

Edge AI explainability and interpretability is a relatively unexplored field with no standard definitions, established mature methods and techniques, best practices, or benchmarking methods. This can make it difficult for edge AI designers to know which approaches to adopt and how to measure their performance and efficiency. The choice of the approach depends on the specific edge AI model, its complexity, the intended solution, and the application's requirements. Combining different techniques may provide a more comprehensive interpretability and explainability solution for edge AI systems.

## 9.7 Conclusion

Explainable and interpretable AI models are applied to AI-based systems to complement them, facilitating the parallel use of data treatment, knowledge processing algorithms and analysable, and answerable implementations. This allows systems to simultaneously process relational and non-relational data from databases and sources that generate data in real-time, such as IoT sensors, and analyse the decision and outputs of the AI models.

The advancements in AI and edge AI require data analysis systems with AI algorithms and the parallel use of mathematical models for the creation of self-explanatory, self-answerable models that incorporate, for example, convolutional neural networks, deep symbolic learning, fuzzy logic, compartmental mathematical models, Bayesian networks, dynamic data assimilation models, and other models from the ML and DL domains.

The concepts of AI and edge AI explainability and interpretability are presented alongside emphasising that interpretability focuses on understanding the inner workings of the models. By contrast, explainability focuses on explaining the decisions made. As a result of the differences between the two concepts, interpretability requires more significant detailing than explainability.

The field of edge AI explainability and interpretability is evolving rapidly, and new approaches, methods and techniques are being developed to improve the explainability and interpretability of AI models and make them more transparent and more functional by improving visualisation methods, decomposition techniques, explanations based on examples, and ante-hoc and post-hoc approaches.

Edge AI involves deploying AI models on devices with inherent resource constraints, such as limited computing power, memory, and latency. Achieving a clear understanding of causality within these systems and making their internal workings and outputs comprehensible to humans often necessitates the use of hybrid approaches or the acceptance of trade-offs, with performance typically taking precedence.

The trade-offs are essential to edge AI explainability and interpretability as performance, energy consumption, complexity, and speed are constantly optimised against each other in resource-constrained edge devices. This is even more relevant considering the need for regular AI model updatability and upgradability.

Another essential consideration is that AI and edge AI models with advanced explainability or interpretability are mainly required in high-risk AI-based applications. Highly explainable/interpretable models can be used to assess AI-based systems by an independent third party and make another party accountable or liable while building trust between designers, developers, and users.

Currently, standardised definitions, mature methods, best practices, and benchmarking techniques are lacking in the field of edge AI explainability and interpretability. Nevertheless, there is an ongoing trend to explore comprehensive solutions that strike a balance between complexity, transparency, and the specific requirements of various applications. Addressing these challenges also requires the implementation of rigorous regulations and robust data quality validation. These efforts are becoming increasingly crucial as the networks of interconnected devices expand, adding complexity to the entire systems and emphasising the need for transparency.

This article attempts to classify and structure the existing concepts, offering the taxonomy needed to understand the multi-dimensionality of elements that must be considered, such as data (e.g., data type, data sets, and data use, encompassing – training, validation, testing, and inference, various AI model methods (e.g., model specific, model agnostic, etc.), extend (e.g., local, global) and the quality and behavioural properties (e.g., causality, transferability, fairness, informativeness, etc.).

In this context, edge AI explainability and interpretability solutions aim to ensure that AI models are transparent, accountable, and compliant with regulations, increasing user confidence and facilitating their adoption in various industries and applications.

## Acknowledgements

This research was conducted as part of the EdgeAI "Edge AI Technologies for Optimised Performance Embedded Processing" project, which has received funding from KDT JU under grant agreement No 101097300. The KDT JU receives support from the European Union's Horizon Europe research and innovation program and Austria, Belgium, France, Greece, Italy, Latvia, Luxembourg, Netherlands, and Norway.

## References

[1] A. Das and P. Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. Available at: https://doi.org/10.48550/arXiv.2006.11371

[2] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018, pp. 0210-0215. Available at: https://doi.org/10.23919/MIPRO.2018.8400040

[3] European Parliament. Artificial Intelligence Act. P9_TA(2023)0236. Online at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

[4] K. Gade, S.C.Geyik, K. Kenthapadi, V. Mithal and A. Taly. Explainable AI in Industry, KDD '19: Proceedings of the $25^{th}$ ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, July 2019, pp. 3203–3204. Available at: https://doi.org/10.1145/3292500.3332281

[5] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G-Z. Yang. "XAI—Explainable artificial intelligence." *Science robotics* 4, no. 37, 2019. Available at: https://openaccess.city.ac.uk/id/eprint/23405/8/

[6] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed. Explainable Artificial Intelligence Approaches: A Survey. Available at: https://doi.org/10.48550/arXiv.2101.09429

[7] J. King, B. Zhang, H. Mahboobi and S. Roy. "Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions". AWS Whitepaper. 2021. Available at: https://docs.aws.amazon.com/pdfs/whitepapers/latest/model-explainability-aws-ai-ml/model-explainability-aws-ai-ml.pdf?did=wp_card&trk=wp_card

[8] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*. 2021; 23(1):18. Available at: https://doi.org/10.3390/e23010018

[9] D. Minh, H.X. Wang, Y.F. Li, *et al.* Explainable artificial intelligence: a comprehensive review. *Artif Intell Rev* **55**, 3503–3568 (2022). Available at: https://doi.org/10.1007/s10462-021-10088-y

[10] M.Z. Naser, An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference, Automation in Construction, Volume 129, 2021, 103821, ISSN 0926-5805. Available at: https://doi.org/10.1016/j.autcon.2021.103821

[11] OECD AI Principles overview. Available at: https://oecd.ai/en/ai-principles

[12] OECD AI Principle 1.4. Robustness, security and safety. Available at: https://oecd.ai/en/dashboards/ai-principles/P8

[13] G. P. Reddy and Y. V. P. Kumar, "Explainable AI (XAI): Explained," *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania, 2023, pp. 1-6. Available at: https://doi.org/10.1109/eStream59056.2023.10134984

[14] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI, International Journal of Human-Computer Studies, Volume 146, 2021, 102551, ISSN 1071-5819. Available at: https://doi.org/10.1016/j.ijhcs.2020.102551

[15] V. Vishwarupe, P. M. Joshi, N. Mathias, S. Maheshwari, S. Mhaisalkar, and V. Pawar, Explainable AI and Interpretable Machine Learning: A Case Study in Perspective, Procedia Computer Science, Volume 204, 2022, pp. 869-876, ISSN 1877-0509. Available at: https://doi.org/10.1016/j.procs.2022.08.105

[16] F. Xu, H. Uszkoreit, Y. Du, W, Fan, D. Zhao, and J. Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Tang, J., Kan, MY., Zhao, D., Li, S., Zan, H. (eds) Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science, vol 11839. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-32236-6_51

[17] D. Omeiza, H. Webb, M. Jirotka and L. Kunze, Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10142-10162, Aug. 2022, Available at: https://doi.org/10.1109/TITS.2021.3122865

[18] F. Di Martino, F. Delmastro. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, vol. 56, pp. 5261–5315, 2023, Available at: https://doi.org/10.1007/s10462-022-10304-3

[19] İ. Kök, F. Y. Okay, Ö. Muyanlı and S. Özdemir, Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey. *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14764-14779, 15 Aug.15, 2023. Available at: https://doi.org/10.1109/JIOT.2023.3287678

[20] M. Abukmeil, A. Genovese, V. Piuri, F. Rundo and F. Scotti, "Towards Explainable Semantic Segmentation for Autonomous Driving Systems by Multi-Scale Variational Attention," *2021 IEEE International Conference on Autonomous Systems (ICAS)*, Montreal, QC, Canada, 2021, pp. 1-5, Available at: https://doi.org/10.1109/ICAS49788.2021.9551172

[21] M. P. S. Lorente, E. M. Lopez, L. A. Florez, A. L. Espino, J. A. I. Martínez, and A. S. de Miguel. Explaining Deep Learning-Based Driver Models, Applied *Sciences*, vol. 11, no. 8, p. 3321, Apr. 2021, Available at: https://doi.org/10.3390/app11083321

[22] R.-K. Sheu and M. S. Pardeshi, A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System, *Sensors*, vol. 22, no. 20, p. 8068, Oct. 2022, Available at: https://doi.org/10.3390/s22208068

[23] 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, Feb. 2023, Available at: https://doi.org/10.3390/electronics12041020

[24] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities, *Energy and AI*, vol. 9, 2022, Available at: https://doi.org/10.1016/j.egyai.2022.100169

[25] A. Dobrovolskis, E. Kazanavičius, and L. Kižauskienė, Building XAI-Based Agents for IoT Systems, Applied Sciences, vol. 13, no. 6, p. 4040, Mar. 2023, Available at https://doi.org/10.3390/app13064040

[26] I. Kalamaras, I. Xygonakis, K. Glykos, S. Akselsen, A. Munch-Ellingsen, H. T. Nguyen, A. J. Lepperod, K. Bach, K. Votis, D. Tzovaras. Visual analytics for exploring air quality data in an AI-enhanced IoT

environment. *Proceedings of the 11th International Conference on Management of Digital EcoSystems (MEDES '19)*. Association for Computing Machinery, New York, NY, USA, 103–110, 2020, Available at https://doi.org/10.1145/3297662.3365816

[27] T.-T.-H. Le, A. T. Prihatno, Y. E. Oktian, H. Kang, and H. Kim, Exploring Local Explanation of Practical Industrial AI Applications: A Systematic Literature Review. *Applied Sciences*, vol. 13, no. 9, p. 5809, May 2023, Available at https://doi.org/10.3390/app13095809

[28] G. Ras, N. Xie, M. van Gerven, Marcel, D. Doran, Explainable Deep Learning: A Field Guide for the Uninitiated. *Journal of Artificial Intelligence Research*, vol. 73, pp.329-397, 2022, Available at https://doi.org/10.1613/jair.1.13200

[29] R. R. Hoffman, S. T. Mueller Shane, G. Klein, J. Litman, Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, vol. 5, 2023, Available at https://doi.org/10.3389/fcomp.2023.1096257

[30] F. Bodria, F. Giannotti, R. Guidotti, et al. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, vol.37, pp. 1719–1778, 2023, Available at https://doi.org/10.1007/s10618-023-00933-9

[31] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri and F. Turini, Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, vol. 34, no. 6, pp. 14-23, 1 Nov.-Dec. 2019, Available at https://doi.org/10.1109/MIS.2019.2957223

[32] D. Alvarez-Melisì, T. S. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Proc. of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, pp. 7786–7795, 2018, Available at https://dl.acm.org/doi/10.5555/3327757.3327875

[33] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, C.-C. Tu, Leveraging Latent Features for Local Explanations. *Proc. of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. Association for Computing Machinery, New York, NY, USA, pp. 1139–1149, 2021, Available at https://doi.org/10.1145/3447548.3467265

[34] R. Brandt, D. Raatjens, G. Gaydadjiev, Precise Benchmarking of Explainable AI Attribution Methods, *arXiv e-prints*, 2023, Available at https://doi.org/10.48550/arXiv.2308.03161

[35] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4). Available at: https://doi.org/10.1002/widm.1312

[36] X.-Q. Chen, C.-Q. Ma, Y.-S. Ren, Y.-T. Lei, N.Q.A. Huynh, and S. Narayan (2023). Explainable artificial intelligence in finance: A bibliometric review. Finance Research Letters, 56, 104145. Available at: https://doi.org/10.1016/j.frl.2023.104145

[37] R. Setchi, M.B. Dehkordi, and J.S. Khan (2020). Explainable Robotics in Human-Robot Interactions. Procedia Computer Science, 176, 3057-3066. Available at: https://doi.org/10.1016/j.procs.2020.09.198

[38] M. Ryo (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. Artificial Intelligence in Agriculture, 6, 257-265. Available at: https://doi.org/10.1016/j.aiia.2022.11.003

[39] D. Gunning, and D. Aha (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44-58. Available at: https://doi.org/10.1609/aimag.v40i2.2850

[40] M. Liao, S.S. Sundar, and J.B. Walther (2022). User Trust in Recommendation Systems: A comparison of Content-Based, Collaborative and Demographic Filtering. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), Article 486, 1–14. Available at: https://doi.org/10.1145/3491102.3501936

[41] M. Ebers (2021). Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s). Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence. Available at: http://dx.doi.org/10.2139/ssrn.3901732

[42] E. Tjoa, and C. Guan (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813. Available at: https://doi.org/10.1109/TNNLS.2020.3027314

[43] A. Kirimtat, O. Krejcar, A. Kertesz, and M.F. Tasgetiren (2020). Future Trends and Current State of Smart City Concepts: A Survey. IEEE Access, 8, 86448-86467. Available at: https://doi.org/10.1109/ACCESS.2020.2992441

[44] S. Thys, W. V. Ranst and T. Goedemé (2019). Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 49-55. Available at: https://doi.org/10.1109/CVPRW.2019.00012

[45] E. Balkir, S. Kiritchenko, I. Nejadgholi, and Kathleen Fraser (2022). Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pages 80–92, Seattle, U.S.A.. Association for Computational Linguistics. Available at: http://dx.doi.org/10.18653/v1/2022.trustnlp-1.8

[46] Mandeep, A. Agarwal, A. Bhatia, A. Malhi, P. Kaler and H. S. Pannu. (2022). Machine Learning Based Explainable Financial Forecasting. 4th International Conference on Computer Communication and the Internet (ICCCI), Chiba, Japan, 2022, pp. 34-38. Available at: https://doi.org/10.1109/ICCCI55554.2022.9850272

[47] S. K. Jagatheesaperumal, Q. -V. Pham, R. Ruby, Z. Yang, C. Xu and Z. Zhang, "Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions," in IEEE Open Journal of the Communications Society, vol. 3, pp. 2106-2136, 2022. Available at: https://doi.org/10.1109/OJCOMS.2022.3215676

[48] K. Simonyan, A. Vedaldi, and A. Zisserman (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings. Available at: https://doi.org/10.48550/arXiv.1312.6034

[49] M.D. Zeiler, R. Fergus (2014). Visualizing and Understanding Convolutional Networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham. Available at: https://doi.org/10.1007/978-3-319-10590-1_53

[50] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi (2018). A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93 (September 2019), 42 pages. Available at: https://doi.org/10.1145/3236009

[51] C. Molnar (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. Available at: https://christophm.github.io/interpretable-ml-book/

[52] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, Volume 99, 2023, 101805, ISSN 1566-2535. Available at: https://doi.org/10.1016/j.inffus.2023.101805

[53] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning", arXiv e-prints, 2017. Available at: https://doi.org/10.48550/arXiv.1702.08608

[54] T. Hastie and R. Tibshirani (1986). Generalized Additive Models. Statistical Science 1, no. 3, 297–310. Available at: http://www.jstor.org/stable/2245459

[55] M.T. Ribeiro, S. Singh, and C. Guestrin (2016). "Why should I trust you?: Explaining the Predictions of Any Classifier." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Kdd San Francisco, CA, 1135–44. New York, NY: Association for Computing Machinery

[56] S. M. Lundberg and S.-I.Lee, "A Unified Approach to Interpreting Model Predictions", Advances in Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774. Available at: https://dl.acm.org/doi/10.5555/3295222.3295230

[57] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations", In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), 2020, pp. 607–617. Available at: https://doi.org/10.1145/3351095.3372850

[58] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019). Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences, 2019. Available at: https://doi.org/10.1073/pnas.1900654116

[59] G. Plumb, D. Molitor, and A. Talwalkar (2018). Model Agnostic Supervised Local Explanations. Neural Information Processing Systems.

[60] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV). Available at: https://doi.org/10.1109/ICCV.2017.74

[61] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences", In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML), 2017, pp. 3145–3153. Available at: https://dl.acm.org/doi/10.5555/3305890.3306006

[62] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise", In Proceedings of the 2017 International Conference on Machine Learning (ICML), Workshop on

Visualization for Deep Learning. Available at: https://arxiv.org/abs/1706.03825

[63] J. Friedman, T. Hastie, and R. Tibshirani (2001). The Elements Of Statistical Learning. Springer, New York. Available at: https://web.stanford.edu/Ÿhastie/ElemStatLearn/printings/ESLIIprint12.pdf

[64] T. Jiang and A.B. Owen, "Quasi-regression for visualization and interpretation of black box functions", 2002, Stanford University. Available at: https://artowen.su.domains/reports/qregvis.pdf

[65] P. Cortez and M.J. Embrechts (2011). Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE. Available at: https://doi.org/10.1109/CIDM.2011.5949423

[66] Z. J. Wang et al., "CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1396-1406, Feb. 2021. Available at: https://doi.org/10.1109/TVCG.2020.3030418

[67] B. K. Iwana, R. Kuroki and S. Uchida, "Explaining Convolutional Neural Networks using Softmax Gradient Layer-wise Relevance Propagation," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 4176-4185. Available at: https://doi.org/10.1109/ICCVW.2019.00513

[68] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network", 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp. 1–6. Available at: https://doi.org/10.1109/ICEngTechnol.2017.8308186

[69] T. Evans, C. O. Retzlaff, C. Geißler, M. Kargl, M. Plass, H. Müller, T.R. Kiehl, N. Zerbe, and A. Holzinger (2022). The explainability paradox: Challenges for xAI in digital pathology. Future Generation Computer Systems, Volume 133. Available at: https://doi.org/10.1016/j.future.202

[70] Hinton, G., Vinyals, O., and Dean, J.. Distilling the Knowledge in a Neural Network. ArXiv, 2015. Available at: 10.48550/arXiv.1503.02531

[71] Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., Baesens, B. (2008). Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. In: Diederich, J. (eds) Rule Extraction from Support Vector Machines. Studies in Computational Intelligence, vol 80. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/978-3-540-75390-2_2

[72] Núñez, H., Angulo, C. & Català, A. Rule-Based Learning Systems for Support Vector Machines. Neural Process Lett 24, 1–18 (2006). Available at: https://doi.org/10.1007/s11063-006-9007-8

[73] P. Cortez and M. J. Embrechts, "Opening black box Data Mining models using Sensitivity Analysis," 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, 2011, pp. 341-348. Available at: https://doi.org/10.1109/CIDM.2011.5949423

[74] Alex Goldstein, Adam Kapelner, Justin Bleich & Emil Pitkin (2015) Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation, Journal of Computational and Graphical Statistics, 24:1, 44-65. Available at: https://doi.org/10.1080/10618600.2014.907095

[75] J. Tan, M. Ung, C. Cheng, and C.S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders", Pacific Symposium on Biocomputing vol. 20, 2015, pp. 132–43. Available at: https://doi.org/10.1142/9789814644730_0014

[76] Goebel, R. et al. (2018). Explainable AI: The New 42? In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science (), vol 11015. Springer, Cham. Available at: https://doi.org/10.1007/978-3-319-99740-7_21

[77] A. Datta, S. Sen and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2016, pp. 598-617, Available at: https://doi.org/10.1109/SP.2016.42

[78] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions", In Proceedings of the 34th International Conference on Machine Learning (ICML), Volume 70, pp. 1885–1894. Available at: https://dl.acm.org/doi/10.5555/3305381.3305576

[79] R. El Shawi, Y. Sherif, M. Al-Mallah and S. Sakr, "Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 2019, pp. 275-280. Available at: https://doi.org/0.1109/CBMS.2019.00065.

[80] A. Yadu, P. K. Suhas and N. Sinha, "Class Specific Interpretability in CNN Using Causal Analysis," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 3702-3706. Available at: https://doi.org/10.1109/ICIP42928.2021.9506118.

[81] B. Malolan, A. Parekh and F. Kazi, “Explainable Deep-Fake Detection Using Visual Interpretability Methods,” 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 2020, pp. 289-293. Available at: https://doi.org/10.1109/ICICT50521.2020.00051.

[82] R. Jiang, Y. Xue and D. Zou, “Interpretability-Aware Industrial Anomaly Detection Using Autoencoders,” in IEEE Access, vol. 11, pp. 60490-60500, 2023. Available at: https://doi.org/10.1109/ACCESS.2023.3286548.

[83] M. P. Neto and F. V. Paulovich, “Explainable Matrix - Visualization for Global and Local Interpretability of Random Forest Classification Ensembles,” in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1427-1437, Feb. 2021. Available at: https://doi.org/10.1109/TVCG.2020.3030354.

[84] R. H. Valencia Tenorio, "Neuroevolved Binary Neural Networks". PhD thesis. The University of Auckland, 2020. Available at: https://researchspace.auckland.ac.nz/bitstream/handle/2292/57055/Valencia%20Tenorio-2020-thesis.pdf?sequence=1