Review

A Literature Review of Explainable Tabular Data

Helen O'Brien Quinn *, Mohamed Sedky , Janet Francis , Michael Streeton

Posted Date: 8 August 2024

doi: 10.20944/preprints202408.0556.v1

Keywords: Explainable Artificial Intelligence (XAI); tabular data, interpretable; machine learning (ML)

Review

Literature Review of Explainable Tabular Data

Helen O'Brien Quinn [1,*], Mohamed Sedky [1,2], Janet Francis [1,2] and Michael Streeton [2]

[1] Staffordshire University, College Road, Stoke on Trent, ST4 2DE, United Kingdom; mhs2@staffs.ac.uk (M.S.); j.francis@staffs.ac.uk (J.F.)
[2] Connexica Ltd, Unit D, Dyson Court, Dyson Way, Stafford Technology Park, Staffordshire, ST18 0LQ, United Kingdom; mike.streeton@connexica.com
* Correspondence: helen.o'brienquinn@research.staffs.ac.uk

Abstract: Explainable Artificial Intelligence (XAI) plays a vital role in increasing transparency and trust in machine learning models, particularly when applied to tabular data which is used in domains such as finance, healthcare, and marketing. This paper presents an extensive survey of XAI techniques used with tabular data and aims to analyze recent research developments since 2021. The survey classes and describes several XAI techniques pertinent to tabular data, it identifies challenges specific to this domain, and explores potential applications and emerging trends. Future research directions are outlined, concentrating on the need for clear definitions of terminology used, data security, user-centric explanations, enhanced interaction, robust evaluation metrics, and advancements in adversarial example-based analysis. The aim is to contribute to the evolving field of XAI, and provide insights for effective, trustworthy, and transparent decision-making using tabular data.

Keywords: Explainable Artificial Intelligence (XAI); tabular data; interpretable; machine learning (ML)

1. Introduction

Artificial intelligence (AI) and machine learning (ML) algorithms are used to build models capable of achieving impressive performance with regards to predictions or classification accuracy in a wide range of domains. However, they typically have a complex, black-box structure that prevents users from gaining a better understanding of the data or the task. As a result, the field of Explainable Artificial Intelligence (XAI) has developed and grown, generating a lot of research in recent years. It is a fast-moving research field, and one in which XAI tries to show the workings of black box algorithms in a more transparent and more easily understood way for users who have a variety of diverse levels of knowledge. This is especially important when the data and models are used to make safety critical decisions, such as in medical diagnostics or in financial risk assessments, where incorrect decisions could be made due to bias and false correlations in the data and the model [1]. Another driver for the development of XAI is the right to an explanation, for users, which is enshrined in law in the European Union General Data Protection Regulations (GDPR).

Understanding how and why ML models make the decisions they do, allows for a deeper trust to be fostered in AI, and better ethical AI solutions because not only is there trust but also accountability [2]. Making explanations easier to understand can function as a driver for the adoption of further AI [3].

Tabular data is the main type of data that companies and businesses deal with, it can be structured or semi-structured, much of it is in tables, spreadsheets, databases, and data warehouses. In the case of tabular data, XAI becomes particularly important because many industries and applications rely on this type of data for decision-making, such as in finance, healthcare, marketing, and many other areas [4,5]. XAI is the interface between ML models and human decision makers, it allows for ethical, effective, and trustworthy decision making in diverse domains [3,6]. XAI increases trust, but it is not enough nor needed for trust. A complete understanding of a model does not stop

 doi:10.20944/preprints202408.0556.v1

it from being untrustworthy if it is faulty. Yet models that are not understood can and are trusted by people, for example a smart TV.

Explainable AI is an important field of research in which several surveys have been carried out, with some examining its specific role with regards to tabular data [5], a survey of explainable Supervised Machine Learning [2], a survey on XAI and natural language explanations [7], revealing black box logic [8], a methods-centric overview with concrete examples [4], a systematic survey of surveys on methods and concepts [9], from approaches, limitations and applications aspects [10]. Sahakyan et al. compiled a survey of XAI used with tabular data, claiming that to their knowledge it was the first survey of this type [5]. Tabular data is used in a myriad of different disciplines, which makes it surprising that there are not more specific XAI methods for it.

This research article seeks to build on the work done by Sahakyan et al., by providing an up-to-date survey of XAI techniques relevant to tabular data by thoroughly analyzing previous studies. Initially, several digital library databases were searched using search strings based on 'Explainable artificial intelligence,' 'XAI,' 'survey,' and 'tabular data,' and limited to the years 2021 to 2023, these being the years following Sahakyan et al.'s survey on XAI for tabular data [5]. The list of references of the selected studies was examined to include other papers that might not have been retrieved from the selected electronic databases. The one hundred and twenty-eight articles found were then whittled down using the search string 'tabular,' giving a total of twelve articles. A further search of the twelve articles using the search string 'Sahakyan' left two articles. When the field was further narrowed to articles published in 2023, the total number of articles published was fifty seven, five of those mention tabular data, and two of those articles mention Sahakyan et al. Table 1 shows the search results by year and from this the interest in explainable tabular data is slowly growing. Tabular data is found in a range of different disciplines and XAI research is taking place, in finance, health, management, statistics, environment, energy, law, engineering, and computing.

Table 1. shows the breakdown of the papers by the search terms used.

Search terms	Number of papers
XAI AND explainable artificial intelligence	128
XAI AND explainable artificial intelligence AND 2021	28
XAI AND explainable artificial intelligence AND 2022	43
XAI AND explainable artificial intelligence AND 2023	57
2021 AND tabular	2
2022 AND tabular	5
2023 AND tabular	5
2021 AND survey (in title)	5
2022 AND survey (in title)	1
2023 AND survey (in title)	8
2021 AND survey AND tabular	1
2022 AND survey AND tabular	6
2023 AND survey AND tabular	11
2021 AND survey AND tabular AND Sahakyan (Sahakyan's article)	1
2022 AND survey AND tabular AND Sahakyan	0
2023 AND survey AND tabular AND Sahakyan	2

The objectives of this survey are:

To analyze the various techniques, inputs, and methods used to build XAI models, since 2021, to determine if any newly proposed model is better for tabular data than the already described models. This research article builds on the work done by Sahakyan et al. by providing an up-to-date survey of XAI techniques pertinent to tabular data, meticulously analyzing recent studies. An extensive literature review was conducted, using digital library databases, focusing on the phrases 'Explainable artificial intelligence,' 'XAI,' 'survey,' and 'tabular data' from 2021 to 2024. The objectives of this survey are threefold:

 doi:10.20944/preprints202408.0556.v1

1. To analyze the various techniques, inputs, and methods used to build XAI models since 2021, aiming to identify superior models for tabular data.
2. To identify the need, challenges, and opportunities in XAI for tabular data.
3. To explore evaluation methods and metrics used to assess the effectiveness of XAI models specifically concerning tabular data.

The document is structured as follows: Section 2 introduces the theoretical foundations of explainability, providing an overview of fundamental concepts underpinning XAI. Section 3 delves into existing techniques for Explainable Tabular Data Analysis, categorizing and describing different XAI techniques suitable for tabular data analysis. Section 4 highlights challenges and gaps in Explainable Tabular Data Analysis. Section 5 explores applications of Explainable Tabular Data Analysis and its impact on decision-making and user trust in various domains. Section 6 looks ahead to future directions and emerging trends, outlining potential research avenues in explainable tabular data analysis. Finally, Section 7 concludes by summarizing the key findings from the literature review.

2. Background

XAI is a way of making it easier for users to understand how models obtain their results. The main aim is to make the workings of a black box model transparent [11]. This has meant that XAI research has burgeoned in recent years, and with it many ways to explain AI systems and their predictions. These methods, when integrated with an AI system may, but not always, meet the legal demands of legislation and guidelines such as the GDPR's right-to-explanation [12] and DARPA's policies [13].

Conducting XAI research surveys is not a simple task because, not only are there different definitions for the same word, but the taxonomies used are also different, meaning that it is difficult to compare like with like [14–16]. Barbiero et al have tried to remedy this by devising a theoretical basis for XAI taxonomies [17], and Ali et al have proposed their own taxonomy, as have Graziani et al., [1,15]. The problem of multiple different taxonomies continues.

The meanings of the terminology around XAI vary, with many terms having no one agreed meaning [5,16]. Unfortunately, the various methods created have been used in isolation with the result that all the methods have not been brought together to form one universally agreed way to explain all AI systems [17].

The X in XAI is referred to as 'explainable,' however the ML community often appears to use the words explainability and interpretability interchangeably to explain transparency and trust. These words, whilst not having agreed formal definitions, are slightly different. Vilone and Longo define interpretability as being "The capacity to provide or bring out the meaning of an abstract concept" [18]. Rudin described interpretability as being a domain specific idea and so maintained that it could not have an all-purpose definition [19]. Barredo Arrieta et al., define interpretability as being the ability to explain or to provide the meaning in a human understandable way [20]. Sahakyan et al viewed explainability as a way of showing the features that have contributed to the result of a specific instance [5]. Barredo Arrieta et al also agreed with this definition, describing it as any procedure or action conducted by the model to make clear its internal workings, meaning it provided a rationale for its decision [20]. They also viewed explainability as being an active process, whilst interpretability has been considered as being passive. Barbiero et al. argued that without agreed explainable AI definitions based in mathematics, there are likely to be a plethora of taxonomies containing unnecessary information, the wrong types of questions and a narrowing of the field for future research directions [17].

Transparency is another term that has no one agreed definition. Often transparency is seen as an umbrella term frequently used with XAI and can be taken to mean the ability to understand an algorithm and its decision making, through such things as simulatability, decomposability, and algorithmic transparency [5,16,20,21]. Haresamudram et al. argue that this is all algorithm based and should have a much broader definition, which includes the whole life cycle. The definition would consist of algorithmic transparency, interaction transparency, and social transparency [21].

 doi:10.20944/preprints202408.0556.v1

How deep learning systems make their decisions is often known as the 'black-box problem' because although it is not clear how they make decisions, they give more accurate results. Wadden explains that the 'black-box problem' is not unique to AI, it occurs in other fields. Agency is a black-box in the field of youth sociology, in race and social issues, fertility and childbirth are black-boxes [22]. The use of this term in other fields can lead to misunderstandings, particularly when looking to use XAI in, for example, a medical setting. Burrell split the opacity problem into three parts, the first they referred to as self-protection and concealment, where algorithms are deliberately opaque to keep proprietary information from others [23]. The second opacity was due to the code used being indecipherable to most members of the public, and the third opacity was that which comes from increasingly complex systems, where learning alters the original algorithm.

Sahakyan et al. also break this problem into three distinct parts, model explanation problems, model inspection problems, and outcome explanation problems [5]. They posit that a model explanation can be obtained by using a more transparent model to simulate the black-box's behavior. A model inspection explanation can be generated in visual or textual format of some properties of the model or its output. And an outcome explanation can be the way an explanation of a particular instance has been generated, or how it can be altered.

Markus et al. proposed, like Sahakyan et al., three separate sorts of explanations. The first, model-based explanations, where a simpler model is used to explain the more complex model. The second, attribution-based explanations, where the model is explained in terms of its input features. The third, example-based explanations, where individual instances are examined and used to explain how the model works [24].

Brożek et al. approach this problem from a legal and psychological perspective, breaking it down into four distinct parts, the opacity problem, the strangeness problem, the unpredictability problem, and the justification problem [25]. They argue that 'the human mind is much more of a black box than the most sophisticated machine learning algorithm' [25]. In the case of the algorithms, how they work is known, even if how they arrived at their decision is not. How the human mind works is still largely unknown. They argue that the opacity problem is not a significant problem, as the workings of algorithms are much better known than the workings of the human mind. They contend that algorithms have been designed to uncover patterns beyond the grasp of the human mind. From their perspective, while explainability is valuable, it should not hold a higher priority than human decision-making. They stated that it is the unpredictability of algorithmic decisions that leads to human skepticism. They refer to this phenomenon as 'strangeness' and suggest that explanations should be approached from a psychological perspective rather than a purely technical one.

Comprehensibility is another term often used when explaining AI. This is often understood as the ability of a learning algorithm to show its learned knowledge in a way which is understandable to humans. Others refer to the quality of the language used by an explainability method [18]. Allgaier et al. opted for a more technical approach and proposed measures to facilitate the examination of code, like providing percentage breakdowns for training, validation, and testing datasets. They argued that the comprehensibility of the XAI system becomes increasingly reliant on domain knowledge as the AI application becomes more specialized. Additionally, they recommended explaining variable names rather than relying on abbreviations, particularly in the context of tabular data, and cautioned against excessive length in explanations, such as overly complex decision trees [14]. Ali et al. found that for comprehensibility, explanations need to be short and readable [1].

Just as there are many ways of generating XAI explanations, there are also many ways of measuring the quality of the explanations. Usually, different XAI methods yield differing explanations on the same dataset and model, the 'best' metric is then selected from all the metrics. If this is unachievable a new XAI method might be written [26]. What do you measure to ensure the quality of explanations? Various aspects of explanations are checked, Li et al., devised a method of assessing consistency and run time efficiency of explanations [26]. Nauta et al., viewed explanation assessment as a twelve faceted problem and developed a twelve-property quality evaluation [27]. Their twelve properties were correctness, consistency, completeness, continuity, contrastively, covariate completeness, compactness, composition, confidence, context, coherence, and

 doi:10.20944/preprints202408.0556.v1

controllability. Lopes et al., looked at evaluation methods as being either human centered or computer centered and devised their own taxonomy for evaluations [28]. They looked at trust, explanation usefulness and satisfaction, understandability, performance, interpretability, and fidelity. Baptista et al., evaluated SHapley Additive exPlanations (SHAP) explanations via the established metrics of monotonicity, 'trendability,' and 'prognosability,' finding that SHAP did trend the metrics, but that model complexity could be a problem [29]. Fouladgar et al., examined the sensitivities of time series XAI models, they found that the sensitivities varied according to how the hyperparameters were set [30]. Oblizanov et al., used synthetic data with LIME and SHAP, and devised a Metrics Calculation Tool. The tool used a modified faithfulness metric, a monotonicity metric, and an incompleteness metric. SHAP and LIME were comparable in terms of accuracy of explanation, however the explanations were not as accurate when used with a decision tree, compared to when they were used with linear regression [31].

3. Existing Techniques for Explainable Tabular Data Analysis

Sahakyan et al explained that explanations could be divided into two groups, those that are model specific and those that are model agnostic. The model specific explanations are obtained by exploiting the model's features and parameters, they also use the model's structure and its weights. Their main disadvantage is that they cannot usually be generalized and used on other types of models [5]. The model agnostic explanations are generated after a model has been trained. They can be used to explain a variety of diverse types of models, but their main disadvantage is that the explanations may not be faithful to the underlying AI as they are produced by a different model.

Models are often categorized into two main types of explanations: local and global. Local explanations pertain to elucidating a single prediction, whereas global explanations aim to expound upon the entire model's behavior [32]. Taxonomies are established based on the distinct approaches adopted for providing explanations, which encompass functioning-based, result-based, conceptual, and mixed methods [32].

The functioning-based approach employs perturbation techniques, involving the manipulation of a model's inputs to identify the most influential features, primarily focusing on local explanations. Additionally, it can leverage the model's structural components, such as gradients, to determine feature importance. This approach also encompasses meta-explanations, architectural modifications, and illustrative examples. The result-based approach also relies on feature importance, surrogate models, and examples. The surrogate model is a simplified rendition of the original model. The conceptual approach classifies explanations according to overarching concepts, such as applicability, scope, problem type, and output type. The mixed approach represents a hybrid, incorporating elements from various other explanation approaches [32].

The explanations can be further divided into five techniques or methods [33]. The first is explanation by simplification, where the explanation is generated through distillation and rule extraction. The second is explanation by feature relevance, where the explanation is generated by measuring the amount of influence a feature has on a prediction. The third is visual explanation, where explanations are in the form of pictures or plots providing information about the model's predictions. The fourth is explanation by concept, where a human-relatable concept is used to generate explanations [34]. The fifth is explanation by example, where an explanation is generated by a factual example that justifies the outcome predicted [35].

There are many ways of generating explanations and many different taxonomies [32]. Figure 1 shows a taxonomy based on the variety of types of explanation and is the most like Sahakyan et al. 's approach [33].

Several surveys on XAI and several aspects of XAI have been conducted recently [1,4,7,8,17,36–47], to name a few. Surveys on evaluating XAI have been conducted by [27] who looked at quantitative XAI evaluation methods, and [26] who defined two metrics, consistency, and efficiency to evaluate XAI methods.

 doi:10.20944/preprints202408.0556.v1

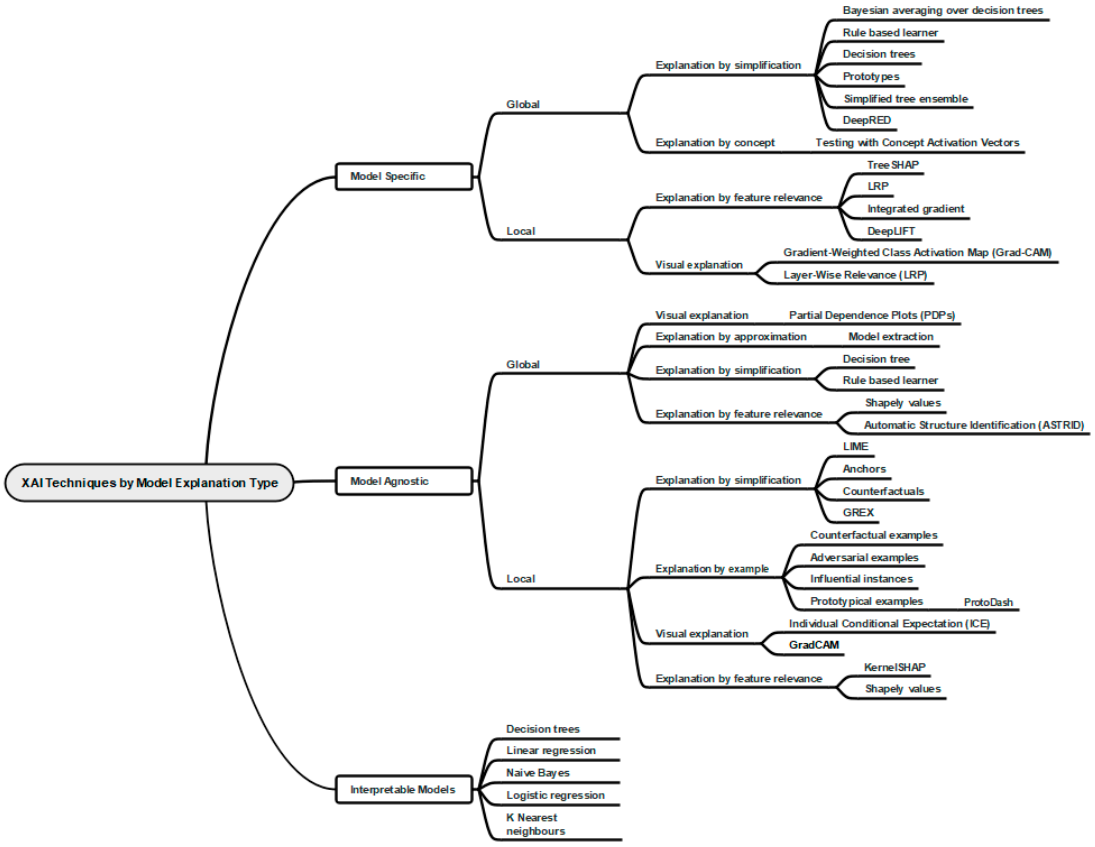

Figure 1. A taxonomy of XAI techniques by explanation type.

4. Challenges and Gaps in Explainable Tabular Data Analysis

Explainable Tabular Data Analysis plays a critical role in enhancing the transparency and interpretability of machine learning models applied to structured data. Despite its importance, several challenges and gaps persist in this domain, hindering its widespread adoption and effectiveness.

In their survey, Sahakyan et al delved into the varied types of explanation and their adaptability to different data types. Their findings revealed that many explanation techniques were primarily developed for image and audio data or tailored to specific AI models, rendering them inadequately suited for tabular data types [5]. XAI techniques are tailored for model types or data forms, and many of these techniques can be complex and challenging to use. Moreover, these techniques face issues related to scalability and coverage. This disparity stems from the inherent attributes of tabular data, where the structure is defined by individual cells serving as fundamental units within the table. Each row in a tabular dataset encapsulates unique attributes, offering a row-centric perspective, while the columns consist of consistent data types, be it numeric or categorical values. Unlike image data, which exhibits spatial and geometric structures, the arrangement of columns in tabular data does not influence the integrity of the underlying information, leading to distinct challenges in constructing adaptable and effective explanation methods for this specific data format.

The challenges inherent in tabular data are multifaceted, arising from issues such as data quality, outliers, missing values, imbalanced classes, and high dimensionality. Despite the abundance of data in tabular datasets, a sizable portion may be of inferior quality, introducing challenges for accurate analysis and modelling. Outliers, representing extreme values within the data, can skew results and impact the performance of machine learning models. Additionally, the presence of missing values further complicates the analysis process, requiring robust imputation techniques to manage the gaps in the data effectively. Imbalanced class distributions pose challenges in classification tasks, where certain classes are underrepresented, potentially leading to biased model outcomes. Moreover, the

 doi:10.20944/preprints202408.0556.v1

high dimensionality of tabular data, characterized by many features, can overwhelm traditional machine learning algorithms and hinder model performance.

Tabular data, by its very nature, reflects the real-world phenomena it aims to capture. If the data is incomplete, biased, or contains inaccuracies, then the insights and explanations derived from that data are inherently limited. For example, if a dataset is missing key variables that are important predictors of the target outcome, the XAI techniques may end up highlighting less relevant features as being important, leading to incomplete or potentially misleading explanations.

Similarly, if the data contains systematic biases, such as under representation of certain demographic groups, the explanations generated may overlook or fail to account for these biases, failing to provide a comprehensive and fair understanding of the model's decision-making process. This can occur deliberately and is known as Fairness hacking. It is the unethical practice of adding or removing sensitive attributes from the testing to lead outsiders to believe that the results are fair [48]. This practice also facilitates the illegitimate exploitation of the many definitions of fairness. This happens when tests are performed using many different fairness metrics, but only the metrics that produce positive results are reported. During the process of fairness hacking, the fairness shortcomings in ML algorithms are hidden by reporting that these algorithms are fair by certain metrics, while the metrics that have produced negative results are suppressed. This practice undermines transparency and accountability of explanations, as it allows organizations to misrepresent the true fairness of their AI systems. It is important fairness testing is conducted comprehensively and rigorously, with all results being reported.

Noisy, or erroneous data can introduce confounding factors, making it challenging for XAI methods to disentangle the true underlying relationships and provide accurate explanations.

Consequently, the quality of explanations can be heavily dependent on the quality and completeness of the data used to train the machine learning models and generate the explanations. Careful data curation, validation, and feature engineering become essential prerequisites for deploying effective and trustworthy XAI systems, particularly in high-stakes applications where the explanations play a crucial role in decision-making. Addressing data quality concerns should be a key consideration in the development and deployment of XAI techniques for tabular data analysis.

The scalability of explainable techniques poses a significant challenge in the context of tabular data, especially as modern datasets continue to increase in both size and complexity. For example, in healthcare, tabular patient records encompass a wide range of medical information, including diagnoses, treatment history, laboratory results, and imaging data. As these datasets expand, existing explainable methods may struggle to efficiently process and interpret the burgeoning volume of tabular healthcare data [49]. Moreover, large-scale tabular datasets, containing thousands or even millions of records, are increasingly prevalent in various domains such as financial transactions, e-commerce sales, and scientific research. This necessitates the development of scalable explainable techniques capable of delivering timely and resource-efficient explanations without compromising accuracy, particularly when analyzing vast tabular datasets to detect anomalies in financial transactions or uncover potential disease markers in extensive genomic data. An example of existing XAI used with tabular data is local Interpretable Model-agnostic Explanations (LIME). It generates explanations for specific instances, not for the whole dataset and it can be resource-intensive with complex models or vast datasets. Therefore, addressing the scalability gap is crucial to ensure that explainable AI techniques remain practical and effective in navigating and extracting meaningful insights from complex tabular datasets across diverse domains, especially as the volume and complexity of tabular data continue to grow at an unprecedented rate.

Neural networks, known for their ability to manage complex patterns in data, face limitations when dealing with such data quality issues as those found in tabular datasets. Consequently, extensive pre-processing steps are essential to address outliers, missing values, and other quality-related issues before leveraging neural networks for analysis. Another critical challenge lies in the inherent complexity of learning the intricate relationships between features in each dataset iteration, necessitating the relearning of structures and patterns in each training cycle. This constant relearning process renders traditional inductive biases used in deep learning less effective when applied to tabular data.

 doi:10.20944/preprints202408.0556.v1

Furthermore, the presence of categorical features in tabular data adds another layer of complexity, as deep learning algorithms traditionally struggle with processing this type of data efficiently. Ensuring accurate feature importance and stability in predictions becomes imperative, as small fluctuations in feature values can significantly impact the overall model predictions and result in unreliable outcomes [50]. The interplay of these challenges underscores the need for advanced pre-processing techniques, specialized model architectures, and robust interpretability methods to effectively navigate and extract meaningful insights from complex tabular datasets [51]. Determining how each feature contributes to the model predictions in the context of tabular data can be intricate, especially when using complex machine learning algorithms. Explaining the significance of each feature accurately for decision-making is not straightforward.

It is common for machine learning models that are used in tabular data analysis to include random forests and boosting techniques, as documented by Tjoa & Guan [52]. However, Borisov et al.'s observations indicated that the most prevalent types of eXplainable Artificial Intelligence (XAI) explanations for tabular data encompass feature highlighting explanations and counterfactual explanations [50]. Despite their effectiveness, it is worth noting that these models may not offer the same level of interpretability as simpler models such as single decision trees or logistic regression. Moreover, as the capabilities of deep learning algorithms continue to advance, working with tabular data remains a persistent challenge.

Table 2 and Figure 2 show some of the XAI methods used with tabular data, along with their pros and cons. The types of XAI have been classed into five groups, counterfactual explanations, feature importance, feature interactions, decision rules, and simplified models. Counterfactual explanations are generated by very slightly altering the input data into a model and seeing what effect it has on the output. 'If this input is altered, will it affect the output'? These explanations can inform what needs to be changed to achieve a desired result [53]. Feature importance involves evaluating the amount each feature contributes to the model's predictions [54]. The features are then shown on a graph in descending order. Feature interactions occur when the effect of one feature on the prediction, made by a model, depends on the value of another feature. This means that the combined effect of two or more features on the target variable is not merely the sum of their individual effects [54]. Decision rules are logical statements that typically take the form 'If [Condition], then [Outcome].' It specifies a conditional relationship where the outcome or decision is determined if the conditions specified in the rule are satisfied [55]. The final group are simplified models. Hassija et al. refer to transferring 'dark knowledge,' complex and hidden insights from sophisticated black-box models to simpler ones like decision trees. This transfer allows these streamlined models to match the predictive capabilities of more complex systems while enhancing interpretability [54].

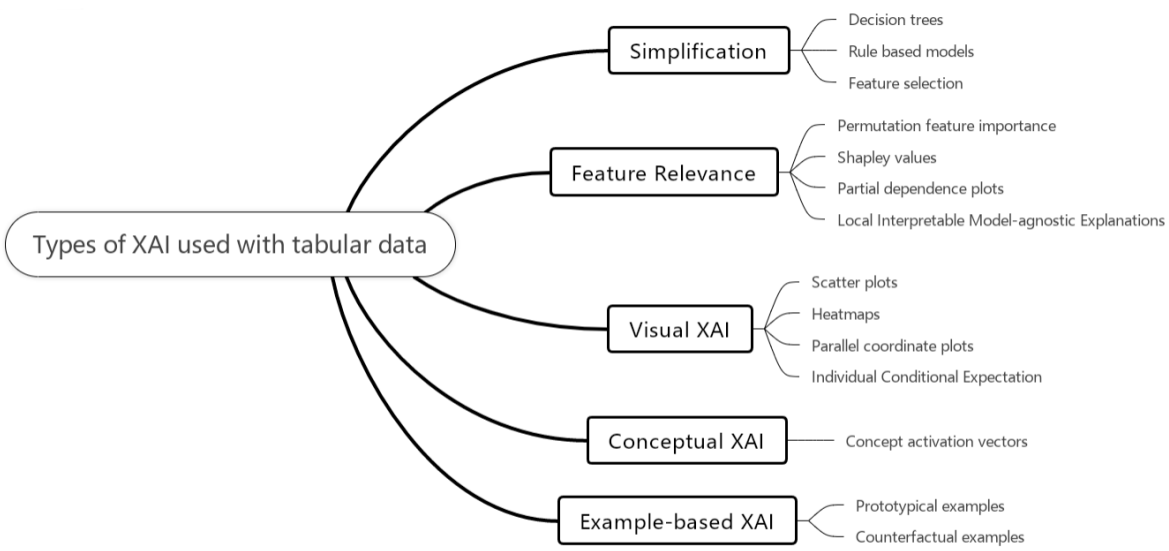

Figure 2. The main types of XAI used with tabular data.

 doi:10.20944/preprints202408.0556.v1

Borisov et al. further pinpointed three fundamental challenges pertaining to tabular data and deep learning, namely inference, data generation, and interpretability [50]. Traditional machine learning models such as decision trees or logistic regression are inherently more interpretable compared to complex models like neural networks commonly used with tabular data. Therefore, explaining the predictions of these more advanced models becomes a challenge. Their studies revealed that decision tree ensembles outperformed deep learning models in the context of tabular data analysis. Additionally, their research included assessing the interpretability of these models, leading to the conclusion that improved benchmarks for the claimed interpretability characteristics and their pragmatic applicability are needful requirements. This research underscores the imperative for comprehensive assessments and benchmarks to validate the interpretability of models and their suitability for real-world applications. These findings encourage further research and development to address the challenges associated with deep learning algorithms and their application to tabular data [50].

In the realm of tabular data, there is a significant lack of universally accepted benchmark datasets and evaluation frameworks. Unlike the fields of computer vision and natural language processing, which have well-established standard benchmark datasets and evaluation protocols, the landscape of tabular data remains unstandardized. Two benchmarks used on XAI for tabular data are synthetic tabular datasets and intrusion detection datasets. Synthetic datasets are datasets that mimic the structure and complexity of real-world tabular data, providing standardized testing environments for evaluating XAI techniques [5]. Whereas the intrusion detection datasets are real-world datasets from domains like network security, such as flow-based network traffic data, used to assess the performance of XAI approaches in practical applications [5].

Without clear benchmarks, comparing the performance and effectiveness of various XAI techniques for tabular data becomes challenging. Researchers and practitioners typically use their own custom datasets or evaluation metrics, complicating meaningful comparisons and assessments of the generalizability of proposed methods. This lack of standardization also impedes the development of best practices, making it difficult to identify the most promising XAI approaches for specific tabular data problems.

To address the challenge of evaluating the practical usefulness of interpretability methods, there is a growing emphasis on creating improved benchmarks for assessing the interpretability characteristics of models used with tabular data [56]. These benchmarks help validate the effectiveness of interpretability methods in real-world applications, ensuring their reliability and utility.

Additionally, when dealing with tabular data, the black-box nature of certain machine learning models, such as deep neural networks, presents a fundamental gap in transparency [54]. In such cases, post-hoc explainable techniques like SHAP and LIME may offer insights into model predictions but may not fully capture the underlying mechanisms of these complex models, particularly when applied to extensive tabular datasets [57].

Moreover, the lack of consensus and standardization in assessing these explanations adds another layer of complexity. The absence of consistent evaluation methods and terminology in explainable AI (XAI) further complicates the comprehensive assessment of explanations, especially in the context of diverse disciplines dealing with extensive tabular datasets. Therefore, addressing scalability and standardization gaps becomes crucial to ensure the practicality and effectiveness of explainable AI techniques when analyzing large and complex tabular datasets, such as those found in healthcare, finance, and scientific research [58].

Tabular data often consist of multiple features or attributes with varying types and scales, requiring XAI techniques to address the complexity and dimensionality of the data in a distinct manner [59]. Non-tabular data, such as images or text, have different properties and may need specialized approaches for effective interpretation. The scalability of XAI techniques varies notably between tabular and non-tabular data. Tabular data presents a particularly pronounced scalability challenge due to the growing volume and intricacy of datasets across diverse domains.

The XAI used for tabular data has similarities but also differences when compared to the XAI used with other forms of data. The key differences discussed are interpretability techniques, where feature importance and local data instances are used more commonly with tabular data due to the

 doi:10.20944/preprints202408.0556.v1

different nature of the data [59]. The inherent structure of tabular data, with rows and columns, introduces unique challenges and opportunities for XAI techniques compared to the non-tabular formats of other data types. For example, the interaction between features in tabular data may require specific methods to interpret and explain the model's decision process [42].

For tabular data, XAI techniques often focus on interpreting the importance and interactions of the various features or variables within the dataset. Methods such as Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and feature importance analysis are commonly employed to identify the key drivers behind the model's outputs. These techniques leverage the structured and quantitative nature of tabular data to provide insights into the model's decision-making process.

In contrast, XAI for image or text data may involve different approaches, such as visualizing the regions of an image that are most salient to the model's prediction, or highlighting the specific words or phrases that contribute the most to a text classification outcome. These modality-specific techniques take advantage of the unique properties of the data, such as spatial relationships in images or semantic dependencies in text.

Despite these differences, there are some common themes and approaches that span multiple data types. For example, the concept of counterfactual explanations, which explore "what-if" scenarios to understand how changes in the input would affect the model's output, can be applied to both tabular and unstructured data.

While XAI methods for tabular data share some similarities with those for other data formats, the unique challenges and characteristics of tabular data necessitate the development of specialized techniques to provide comprehensive and actionable explanations. Understanding these nuances is crucial for selecting and applying the most appropriate XAI approaches for a given problem and data context.

In summary, while explainable techniques hold promise for enhancing the transparency and trustworthiness of machine learning models applied to tabular data, several challenges and gaps must be addressed to realize their full potential. Overcoming these obstacles will require interdisciplinary collaboration among researchers, practitioners, and policymakers to develop innovative solutions and establish best practices.

Table 2. Types of XAI for tabular data and their pros and cons.

Summary of XAI types					
Type of XAI	Description	Examples	Pros	Cons	Evaluation
Counterfactual explanations	Counterfactual explanations generate similar input instances that lead to different predicted results, providing insight into the model's initial prediction rationale.	DiCE WatcherCF GrowingSpheresCF	Causal insight – understand the causal relationship between input features and predictions. Personalized explanations – tailors individualized insights for better insights. Decision support – aids decision making with actionable	Complexity – generating counterfactuals is computationally intensive, particularly for complex models and high-dimensional data. Model specificity – effectiveness is influenced by the underlying model's characteristics. Interpretation – conveying implications can	Alignment with predicted outcome – ensuring the generated counterfactual instances closely reflect the intended predicted outcome. Proximity to original instance – maintaining similarity to the original instance whilst altering the fewest features possible. Diverse outputs – capable of

 doi:10.20944/preprints202408.0556.v1

Summary of XAI types					
Type of XAI	Description	Examples	Pros	Cons	Evaluation
			outcome-focused changes	necessitate domain expertise.	producing multiple diverse counterfactual explanations. Feasible feature values – the counterfactual features should be practical and adhere to the data distribution.
Feature importance	Feature importance techniques quantify the relative contribution of each input feature to the model's predictions	Permutation Importance Gain Importance. SHAP Feature Importance LIME	Helps in feature selection and model interpretability. Provides insight into the most influential features driving the model's decisions.	May not capture complex feature interactions. Can be sensitive to data noise and model assumptions.	Relative importance – rank features based on their contribution to the model's prediction. Stability – ensure consistency of feature importance over different subsets of the data or re-trainings of the model. Model impact - Assessing the influence of individual features on the model's predictive performance
Feature interactions	Feature interaction analysis looks at how the combined effect of multiple input features influences the model's predictions.	Partial Dependence plots Accumulated Local Effects plots. Interaction Values Individual Conditional Expectation Plots	Reveals intricate and synergistic connections among features. Enhances insight into the model's decision-making mechanism.	Visualizing and interpreting features can be difficult, especially when dealing with high-dimensional data. The computational complexity grows as the number of interacting features increases.	Non-linear relationships – uncovers and visualizes complex, non-linear interactions among the features. Holistic insight – provides a comprehensive understanding of how features collectively impact the model's predictions. Predictive power – evaluates the combined effects of interacting features

 doi:10.20944/preprints202408.0556.v1

Summary of XAI types					
Type of XAI	Description	Examples	Pros	Cons	Evaluation
					on the model 's performance.
Decision rules	Decision rules are if-then-else statements that describe the model's decision logic in a human-interpretable format	Decision Trees Rule-Based Models Anchors	Provides clear and intuitive insights into the model's predictions. Easily understood by non-technical stakeholders	Might struggle to capture complex relationships in the data, leading to oversimplification. Can be prone to overfitting, reducing generalization performance	Transparency – offers clear and interpretable explanations of the conditions and criteria used for decision making. Understandability – ensures ease of understanding by non-technical stakeholders and experts alike. Model adherence – check that decision rules capture accurately the model's decision logic without oversimplification.
Simplified models	Simplified models are interpretable machine learning models that approximate the behavior of a more complex black-box model	Generalized Additive Models. Interpretable Tree Ensembles.	Gives a balance between model interpretability and model complexity. Offers global insights into the model's decision-making process	Might not capture the total complexity of the underlying data generating process. Needs careful model choice and tunning to maintain a good trade-off between interpretability and accuracy.	Balance of complexity – achieves an optimal compromise between model simplicity and predictive performance. Interpretable representation – ensures that the offers transparent and intuitive insights into the original complex model's behavior. Fidelity to original model - Assesses the extent to which the simplified model captures the key characteristics and patterns of the original complex model.

5. Applications of Explainable Tabular Data Analysis

Explainable Tabular Data Analysis has various uses across different domains where understanding and explaining the results of machine learning models used on tabular data are

 doi:10.20944/preprints202408.0556.v1

crucial. XAI is an interdisciplinary research field, and its applications can be found in finance, healthcare, autonomous vehicles and many more areas [11].

Examples of explainable tabular data analysis use cases in the financial sector are identity verification in client onboarding, transaction data analysis, fraud detection in claims management, anti-money laundering monitoring, price differentiation in car insurance, automated analysis of legal documents, and the processing of loan applications, to name a few [60].

In the finance industry, banks and lending institutions use explainable tabular data analysis to explain credit scoring models. Customers can receive explanations for not only why their credit applications were accepted or rejected, but also to help them understand what factors influenced the decision. The customers can then be given an idea of what they need to do to improve their credit scores.

In healthcare, explainable tabular data analysis can be applied to explain the model predictions used for patient risk assessment, diagnosis, or treatment recommendations. Doctors and patients can benefit from the explanations in making personalized healthcare decisions.

Explainable tabular data analysis can be used in the field of fraud detection, to provide explanations for why a transaction or an anomaly was flagged as potentially fraudulent. This helps fraud analysts understand the model's decision and take appropriate actions in a timely fashion.

It is used in business to explain why a customer is predicted to churn. This can include highlighting the key features contributing to the churn risk, allowing companies to take targeted actions to retain their customers. In Human Resources (HR) departments it can be employed to understand why employees are likely to leave a company. This can help with retention strategies and the improvement of workplace conditions. Insurance companies use this type of analysis to explain how premiums are calculated and why certain people receive higher or lower rates. This transparency can build and increase trust with policyholders.

Manufacturing companies can use it to explain why a product failed a quality control check. This can lead to work, process improvements, and reduce the number of defects. In logistics and supply chain management, this sort of analysis can explain decisions related to inventory management, shipping routes, and order fulfilment, assisting with better decision-making. E-commerce platforms use this type of analysis to explain product recommendations to customers. It can highlight the features or historical behavior that influenced the recommendations.

It can be applied by utility companies to explain how they forecast energy consumption for customers. This transparency can help consumers make informed decisions about their energy usage. Explainable tabular data analysis can also be used to explain why certain legal decisions or compliance assessments have been made. This can be critical where transparency and fairness are essential. In the education sector, it can provide show why students receive grades or recommendations for courses. This can help teachers tailor their teaching methods. It can be applied to explain predictions related to environmental factors, such as air quality, weather forecasting, and pollutant levels.

In these applications, explainable tabular data analysis not only improves model transparency but also enhances decision-making, fosters trust, and assists in compliance with regulations, making it a valuable tool to use across a wide range of industries and use cases.

6. Future Directions and Emerging Trends

A significant amount of earlier research has concentrated on improving the quality of explanations and offering suggestions for future research. Nevertheless, there is still uncertainty surrounding the terminology employed in the field of XAI, highlighting the need for future research to establish standardized definitions to encourage a broader acceptance of XAI. In the area of tabular data, challenges arise due to the frequent occurrence of multicollinearity or substantial inter-correlations among features.

The work done by Vellido and Martín-Guerrero as far back as 2012 highlighted an important challenge in the field of Explainable Artificial Intelligence (XAI) - the lack of publicly available tabular datasets that possess both annotated labels and concept-related attributes [61]. Alkhatib et al. noted that this gap in accessible data has hindered progress in specific research efforts focused on XAI

 doi:10.20944/preprints202408.0556.v1

techniques for tabular data. In 2023, they encountered obstacles in locating publicly accessible tabular datasets that possessed both annotated labels and concept-related attributes [62]. The creation of such datasets would encourage specific research into XAI and tabular data.

Currently an area of research focus is on methods of adversarial example-based analysis, this is being done on natural language processing tasks and on images. Examination of these methods on tabular data could be a useful area of research [42]. Explanations can always be improved and created to be more user focused, this would involve the development and use of intelligent interfaces, capable of interacting with the user and generating relevant explanations [63]. In critical decision areas such as medicine for example, the expert user would need to be able to interact with the system and would need to know that any mapping of 'causability' and explainability are extremely reliable [42]. This would give two areas of further research, relevant explanations and user interaction with the model or system. Table 3 and Figure 3 give further suggestions for research ideas.

Security has a significant role to play in promoting trust in the use of XAI. People want to know that their data is safely stored, and that results from models using their data are only available to a restricted set of people. This is also an obligation under GDPR, where personal data should not be stored longer than necessary for the purposes for which it was collected, this is the principle of storage limitation. The research and development of XAI is particularly important in the context of the GDPR, which requires the implementation of appropriate technical and organizational safeguards to ensure the security of personal data against unauthorized or unlawful processing, accidental loss, destruction, or damage. Under GDPR fairness hacking would come under unlawful processing.

Table 3. Suggested areas of further research.

Possible research areas	Suggestions
Hybrid Explanations [64]	Combining multiple XAI techniques to provide more comprehensive and robust explanations for tabular data models. [64] Integrating global and local interpretability methods to offer both high-level and instance-specific insights.
Counterfactual Explanations	Generating counterfactual examples that show how the model's predictions would change if certain feature values were altered. [65] Helping users understand the sensitivity of the model to different feature inputs and how to achieve desired outcomes.
Causal Inference [66]	Incorporating causal reasoning into XAI methods to better understand the underlying relationships and dependencies in tabular data. [66] Identifying causal features that drive the model's predictions, beyond just correlational relationships.
Interactive Visualizations	Developing interactive visualization tools that allow users to explore and interpret the model's behavior on tabular data [63] Enabling users to interactively adjust feature values and observe the corresponding changes in model outputs [63]
Scalable XAI Techniques [4]	Designing XAI methods that can handle the growing volume and complexity of tabular datasets across various domains [67]. Improving the computational efficiency and scalability of XAI techniques to support real-world applications.
Domain-specific XAI	Tailoring XAI approaches to the specific needs and requirements of different industries and applications that rely on tabular data, such as finance, healthcare, and manufacturing. Incorporating domain knowledge and constraints to enhance the relevance and interpretability of explanations. [68]
Automated Explanation Generation [69]	Developing AI-powered systems that can automatically generate natural language explanations for the model's decisions on tabular data. [70].

 doi:10.20944/preprints202408.0556.v1

Possible research areas	Suggestions
	Bridging the gap between the technical aspects of the model and the end-user's understanding [70].

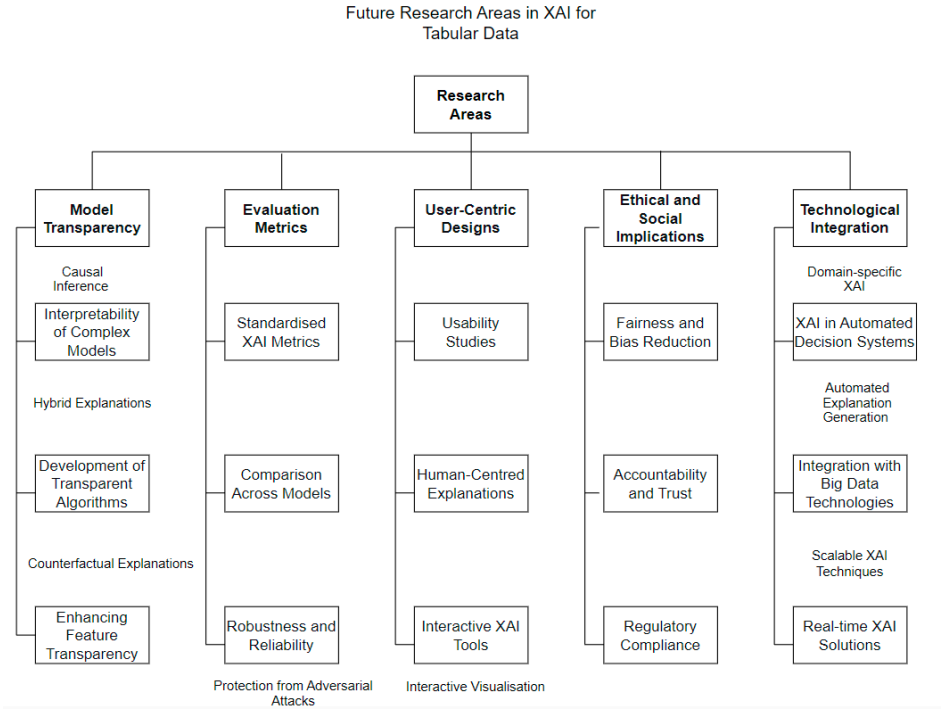

Figure 3. Future research areas in XAI for tabular data.

Another aspect of data security under GDPR is safety from adversarial attacks, as these are constantly changing, it is an area of constant research [8]. LIME and SHAP explanations for tabular data can be manipulated by exploiting their reliance on perturbing input data for estimation. The proposed attack substitutes a biased black-box model with a model surrogate to effectively conceal bias, for instance, from auditors. An out-of-distribution detector is trained to divide the input data so that the black-box model's predictions remain biased within the in-distribution, but its behavior on the perturbed data is controlled. This allows the explanations to seem fair, even though the underlying model may be biased. The aim is to use the fundamental mechanisms of LIME and SHAP, which rely on perturbing the input data to estimate the feature importances and give explanations. By carefully crafting a model surrogate that exhibits the desired behavior on the perturbed data, the real nature of the black-box model's biases can effectively be hidden. This technique presents a real and concerning threat, as it could enable the masking of harmful biases in high-stakes AI systems from regulatory oversight and public scrutiny. It emphasizes the need for robust and comprehensive testing of XAI methods to ensure they are not susceptible to such manipulative attacks [71]. SHAP is the XAI technique most susceptible to adversarial attack [71].

The newer EU AI Act takes a risk-based approach and classifies A.I. into four categories with each being deemed less risky than the previous one. Depending upon the level of risk, different requirements apply. The AI Act places particular focus on high-risk AI systems, which are required to meet stringent requirements, including transparency and explainability. This means that providers of high-risk AI systems must ensure that their models can provide understandable explanations for their decisions and predictions. High-risk AI systems, such as those used in healthcare and credit scoring, are expected to be transparent and have human oversight, however many such systems use black box algorithms and tabular data. This aspect of the requirements, as described in Articles 13 and 14, can be met using XAI. Users of high-risk AI systems have the right to receive clear information about how the AI system works, which necessitates the development and integration of XAI techniques to fulfil these obligations.

 doi:10.20944/preprints202408.0556.v1

For intermediate systems like recommender systems and chatbots, users should be informed of the capabilities and limitations of the system, and in particular users should be told that they are interacting with a bot [72].

To facilitate meaningful human oversight, the AI Act requires that explanations be provided in a manner that humans can comprehend. This enhances the importance of XAI in making AI models interpretable and understandable. There is a requirement that AI systems, particularly high-risk ones, are designed in a way that allows for the traceability of decisions. XAI techniques help by providing detailed insights into how different inputs affect outputs, making it easier to audit and trace decisions.

The AI Act emphasizes the importance of ensuring that AI systems do not perpetuate biases or discrimination. For tabular data, XAI methods can help detect and explain biases in the system, allowing steps to promote fairness to be taken. Further research areas will no doubt become apparent as the Act is used more widely [73].

Existing XAI has few evaluation metrics to demonstrate how effective the explanations are. This could be a vast area of research, particularly when explanations are generated for a wide range of users, all with their own specific needs. There needs to be an accepted way of measuring not only the quality, but also the satisfaction of explanations. This would benefit users and researchers as it would allow for a better comparison between models.

Future research in XAI should focus on establishing clear definitions, enhancing user-focused explanations, incorporating user interaction, ensuring data security, and devising robust evaluation metrics. Adversarial example-based analysis, user-specific explanations, and addressing security concerns are emerging areas of interest in XAI research [42,63]. Another key area is that of intra-metric measurements to avoid the selection of different favorable metrics and omitting the unfavorable metrics.

7. Conclusions

Explainable Artificial Intelligence has emerged as an important field of research in recent years, driven by the need to understand the complex, black-box nature of AI and machine learning models. XAI aims to provide more transparent and easily understandable explanations of how these models make decisions. This is crucial for building trust and enabling ethical decision-making, especially in safety-critical domains like healthcare and finance. Given that tabular data is the predominant data format used in many industries, XAI becomes particularly important for these types of datasets. While several surveys have been conducted on XAI approaches, methods, and applications, there is a surprising lack of XAI techniques specifically tailored for tabular data, despite its widespread use across diverse disciplines. Overall, XAI is seen as a main driver for the broader adoption and trust in AI-powered decision-making systems.

The machine learning, and XAI community often uses the terms "explainability" and "interpretability" interchangeably when referring to the goal of making AI systems more transparent and trustworthy. There is still a lack of agreed-upon formal definitions, which has led to the proliferation of different taxonomies and frameworks in the XAI field. Establishing agreed-upon mathematical definitions for these concepts could help provide a more unified foundation for XAI research and development.

The "black-box problem" is a multifaceted challenge, and researchers have proposed various frameworks for categorizing and addressing the various aspects of model opacity. The "black-box problem" also has legal, psychological, and technical dimensions, and researchers are exploring various approaches to improve the comprehensibility and explainability of AI systems beyond just the technical aspects.

Assessing the quality of XAI explanations is a multi-faceted challenge, with researchers exploring diverse technical and human-centric metrics and frameworks to tackle this problem. There are model-specific explanations that exploit internal model structures, and model-agnostic explanations that are more generally applicable but may be less faithful to the underlying model. Taxonomies of explanation approaches further categorize the different explanations into five key techniques for

 doi:10.20944/preprints202408.0556.v1

generating XAI explanations. These are simplification, feature relevance, visual, conceptual, and example-based, which can be further organized into different taxonomical frameworks.

Despite the critical importance of Explainable Tabular Data Analysis, there are significant challenges and gaps that persist in this domain, stemming from the inherent properties of tabular data and the limitations of current XAI techniques. Addressing these challenges is crucial for enhancing the transparency and interpretability of machine learning models applied to structured data.

There are significant scalability challenges of explainable techniques for tabular data, as well as the limitations of neural networks in effectively managing the unique characteristics and quality issues present in tabular datasets. The complex challenges posed by categorical features, feature importance, and stability in tabular data, require the development of advanced techniques, specialized architectures, and robust interpretability methods to effectively analyze and extract insights from complex tabular datasets.

There is also a critical need for improved benchmarks and assessment frameworks to validate the interpretability and practical usefulness of machine learning models, especially when applied to complex tabular datasets, which have unique challenges and scalability issues associated with applying XAI techniques, and there is a need for greater standardization and interdisciplinary efforts to address these gaps.

There is a critical need for further research into XAI techniques that are robust against adversarial attacks. If malicious actors can exploit the vulnerabilities of prominent explanation methods to obscure harmful biases, it poses a grave threat to the integrity and trustworthiness of AI systems, particularly in high-stakes domains. Developing XAI approaches that are impervious to such manipulative tactics is paramount, as it would ensure that the explanations provided are a faithful and accurate reflection of the model's true decision-making process. This, in turn, would bolster transparency, accountability and public trust in the deployment of AI, allowing for more rigorous auditing and oversight. The pursuit of XAI techniques resistant to adversarial attacks is a crucial frontier in using the full potential of AI while safeguarding against its misuse.

The key research challenges and future directions include standardizing terminology, addressing data availability, adapting adversarial analysis to tabular data, and developing more user-focused and reliable explanations, particularly for critical applications. The exploration and advancement of XAI must consider the security and regulatory requirements, particularly under the GDPR and the EU AI Act, while also addressing the need for robust evaluation metrics and user-focused explanations.

Explainable AI has emerged as a critical field of research, driven by the need to build trust and enable ethical decision-making in the face of complex, black-box AI models. As tabular data remains the predominant format used across many industries, the development of XAI techniques tailored for structured data is of paramount importance. However, the research community still faces significant challenges in establishing agreed-upon definitions, addressing the unique characteristics of tabular data, and devising robust evaluation frameworks. Addressing these gaps is crucial not only for enhancing the transparency and interpretability of AI systems, but also for ensuring compliance with evolving security and regulatory requirements, such as those outlined in the GDPR and the EU AI Act. By focusing on user-centric explanations, robust evaluation metrics, and the incorporation of security best practices, the XAI research community can pave the way for the broader adoption and trust in AI-powered decision-making systems across a wide range of safety-critical domains.

Author Contributions: Conceptualization, H.O'B.Q; Methodology, H.O'B.Q.; software, H.O'B.Q.; investigation, H.O'B.Q.; writing-original draft preparation, H.O'B.Q.; writing-review and editing, M.S., H.O'B.Q.; Supervision, M.S.; J.F., funding acquisition, M.S; M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by Connexica Ltd and Innovate UK [Grant Reference: KTP10021376].

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

 doi:10.20944/preprints202408.0556.v1

References

1. Ali, S., Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N. & Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*(January 2023), 101805, **1-52**. https://doi.org/10.1016/j.inffus.2023.101805.
2. Burkart, N. & Huber, M. F. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 2021, *70*, 245–317. https://doi.org/10.1613/JAIR.1.12228
3. Weber, L.; Lapuschkin, S.; Binder, A. & Samek, W. Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, *92*(November 2022), 154–176. https://doi.org/10.1016/j.inffus.2022.11.013
4. Marcinkevičs, R. & Vogt, J. E. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *February 2022*, 1–32. https://doi.org/10.1002/widm.1493
5. Sahakyan, M., Aung, Z., & Rahwan, T. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access*, 2021, *9*, 135392-135422. https://doi.org/10.1109/ACCESS.2021.3116481
6. Alicioglu, G. & Sun, B. A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers and Graphics (Pergamon)*, 2022, *102*, 502–520. https://doi.org/10.1016/j.cag.2021.09.002
7. Cambria, E.; Malandri, L.; Mercorio, F; Mezzanzanica, M, & Nobani, N. A survey on XAI and natural language explanations. *Information Processing and Management*, 2023, *60*(1), 103111, **1-16**. https://doi.org/10.1016/j.ipm.2022.103111
8. Chinu & Bansal, U. Explainable AI: To Reveal the Logic of Black-Box Models. In *New Generation Computing* (Issue February), 2023, **53-87**. Springer Link Nature. https://doi.org/10.1007/s00354-022-00201-2
9. Schwalbe, G. & Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, (2023), **1-59**. https://doi.org/10.1007/s10618-022-00867-8
10. Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R. & Yao, N. Survey on Explainable **AI:** From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems*, 2023, ***3(3)***, **161-188**. https://doi.org/10.1007/s44230-023-00038-y
11. Hamm, P.; Klesel, M.; Coberger, P. & Wittmann, H. F. Explanation matters: An experimental study on explainable AI. *Electronic Markets*, 2023, *33*(1), **1-21**. https://doi.org/10.1007/s12525-023-00640-9
12. Lance, E. Ways That the GDPR Encompasses Stipulations for Explainable AI or XAI. ***SSRN, Stanford Center for Legal Informatics*** (April 15, 2022), **1-7**. https://ssrn.com/abstract=4085089
13. Gunning, D.; Vorm, E.; Wang, J. Y. & Turek, M. DARPA 's explainable AI **(XAI)** program: A **retrospective.** *Applied AI Letters*, 2021, *2*(4), 1–11. https://doi.org/10.1002/ail2.61
14. Allgaier, J.; Mulansky, L.; Draelos, R. L. & Pryss, **R.** How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*, 2023, *143*(May), 102616, **1-13**. https://doi.org/10.1016/j.artmed.2023.102616
15. Graziani, M.; Dutkiewicz, L.; Calvaresi, D.; Amorim, J. P.; Yordanova, K.; Vered, M.; Nair, R.; Abreu, P. H.; Blanke, T.; Pulignano, V.; Prior, J. O.; Lauwaert, L.; Reijers, W.; Depeursinge, A.; Andrearczyk, V. & Müller, H. (2023). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. In *Artificial Intelligence Review*. 2023, (Vol. 56, Issue 4), **3473–3504**. Springer Netherlands. https://doi.org/10.1007/s10462-022-10256-8
16. Bellucci, M.; Delestre, N.; Malandain, N. & Zanni-Merk, C. Towards a terminology for a fully contextualized XAI. *Procedia Computer Science*, 2021, *192*, 241–250. https://doi.org/10.1016/j.procs.2021.08.025
17. Barbiero, P.; Fioravanti, S.; Giannini, F.; Tonda, A.; Lio, P. & Di Lavore, E. *Categorical Foundations of Explainable AI: A Unifying Formalism of Structures and Semantics*. **(eds) Explainable Artificial Intelligence. xAI** 2024. **Communications in Computer and Information Science, vol 2155. 185–206**, **Springer, Cham. https://doi.org/10.1007/978-3-031-63800-8_10**
18. Vilone, G., & Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 2021, *76*(May), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009
19. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

20. Barredo Arriet**a,** A.**;** Díaz-Rodríguez, N.**;** Del Ser, J.**;** Bennetot, A.**;** Tabik, S.**;** Barbado, A.**;** Garcia, S.**;** Gil-Lopez, S.**;** Molina, D.**;** Benjamins, R.**;** Chatila, R. & Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020**,** *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
21. Haresamudra**,** K.**;** Larsson, S. & Heintz, F. (2023). Three Levels of AI Transparency. *Computer*, 2023, *56*(2), 93–100. https://doi.org/10.1109/MC.2022.3213181
22. Wadde**n,** J. J. Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics*, 2021**,** *48*(10), 764–768. https://doi.org/10.1136/medethics-2021-107529
23. Burrel**l,** J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, 2016**,** *3*(1), 1–12. https://doi.org/10.1177/2053951715622512
24. Markus**,** A. F.**;** Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*(July 2021), 103655**,** **1-11**. https://doi.org/10.1016/j.jbi.2020.103655
25. Brożek**,** B.**;** Furman, M.**;** Jakubiec, M. & Kucharzyk, B. The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law*, 2023**,** 32:427–440. https://doi.org/10.1007/s10506-023-09356-9
26. Li**,** D.**;** Liu, Y.**;** Huang, J. & Wang, Z. A Trustworthy View on Explainable Artificial Intelligence Method Evaluation. *Computer*, 2023**,** *56*(4), 50–60. https://doi.org/10.1109/MC.2022.3233806
27. Nauta**,** M.**;** Trienes, J.**;** Pathak, S.**;** Nguyen, E.**;** Peters, M.**;** Schmitt, Y.**;** Schlötterer, J.**;** van Keulen, M. & Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 2023**,** *55**(13)***,** 295**,** **1-42**. https://doi.org/10.1145/3583558
28. Lopes**,** P.**;** Silva, E.**;** Braga, C.**;** Oliveira, T. & Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-**Centered** Methods. *Applied Sciences (Switzerland)*, 2022**,** *12*(19)**,** **1-31**. https://doi.org/10.3390/app12199423
29. Baptista, M. L.**;** Goebel, K. & Henrique**s,** E. M. P. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artificial Intelligence*, 2022**,** *306*, 103667**,** **1-22**. https://doi.org/10.1016/j.artint.2022.103667
30. Fouladga**r,** N.**;** Alirezaie, M. & Framling, K. Metrics and Evaluations of Time Series Explanations: An Application in Affect Computing. *IEEE Access*, 2022**,** *10*, 23995**-**24009. https://doi.org/10.1109/ACCESS.2022.3155115
31. Oblizano**v,** A.**;** Shevskaya, N.**;** Kazak, A.**;** Rudenko, M. & Dorofeeva, A. (2023). Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. *Applied System Innovation*, 2023**,** *6*(1), 1–13. https://doi.org/10.3390/asi6010026
32. Speit**h,** T. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *ACM International Conference Proceeding Series*, 2022**,** 2239–2250. https://doi.org/10.1145/3531146.3534639
33. Kurdziole**k** M. Explaining the Unexplainable: Explainable AI (XAI) for UX. *User Experience Magazine*, 2022**,** ***(Issue 22.3)*, Available online:**.https://uxpamagazine.org/explaining-the-unexplainable-explainable-ai-xai-for-ux/ **(Accessed on 20 August 20, 2023)**
34. Kim**,** B.**;** Wattenberg, M.**;** Gilmer, J.**;** Cai, C.**;** Wexler, J.**;** Viegas, F. & Sayres, R. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML* 2018, *6*, 4186–4195. https://proceedings.mlr.press/v80/kim18d/kim18d.pdf
35. Kenny**,** E. M. & Keane, M. T. Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems*, 2021**,** *233*, 107530**,** **1-14**. https://doi.org/10.1016/j.knosys.2021.107530
36. Alfeo, A. L.**;** Zippo, A. G.**;** Catrambone, V.**;** Cimino, M. G. C. A.**;** Toschi, N. & Valenza, G. From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine*, 2023**,** *236*, 107550**,** **1-12**. https://doi.org/10.1016/j.cmpb.2023.107550
37. An, J.**;** Zhan**g,** Y. & Joe, I. Specific-Input LIME Explanations for Tabular Data Based on Deep Learning Models. *Applied Sciences (Switzerland)*, 2023**,** *13*(15)**,** **1-19**. https://doi.org/10.3390/app13158782
38. Bharat**i,** S.**;** Mondal, M. R. H. & Podder, P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When. *IEEE Transactions on Artificial Intelligence*, 2023**,** **(Vol 5, Issue 4), 1429 - 1442**. https://doi.org/10.1109/TAI.2023.3266418

39. Chaddad, A., Peng, J., Xu, J., & Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors*, 2023, *23*(2), 1–19. https://doi.org/10.3390/s23020634
40. Chamola V.; Hassija, V.; Sulthana, A. R.; Ghosh, D.; Dhingra, D. & Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access*, 2023, *11*(July), 78994- 79015. https://doi.org/10.1109/ACCESS.2023.3294569
41. Chen, X. Q.; Ma, C. Q.; Ren, Y. S.; Lei, Y. T.; Huynh, N. Q. A. & Narayan, S. Explainable artificial intelligence in finance: A bibliometric review. *Finance Research Letters*, 2023, *56*(May), 104145, **1-10**. https://doi.org/10.1016/j.frl.2023.104145
42. Di Martino F. & Delmastro, F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data. In *Artificial Intelligence Review* 2023, (Vol. 56, Issue 6). Springer Netherlands, 56:5261–5315 https://doi.org/10.1007/s10462-022-10304-3
43. Kok, I.; Okay, F. Y.; Muyanli, O. & Ozdemir, S. Explainable Artificial Intelligence (XAI) for Internet of Things: A Survey. *IEEE Internet of Things Journal*, 2023, *10*(16), 14764–14779. https://doi.org/10.1109/JIOT.2023.3287678
44. Haque, A. B.; Islam, A. K. M. N. & Mikalef, P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, 2023, *186*(PA), 122120, **1-19**. https://doi.org/10.1016/j.techfore.2022.122120
45. Sahoh, B. & Choksuriwong, A. The role of explainable Artificial Intelligence in high-stakes decision-making systems: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 2023, *14*(6), 7827–7843. https://doi.org/10.1007/s12652-023-04594-w
46. Saranya, A. & Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 2023, *7*(April), 100230, **1-14**. https://doi.org/10.1016/j.dajour.2023.100230
47. Sosa-Espadas, C. E.; Orozco-del-Castillo, M. G.; Cuevas-Cuevas, N. & Recio-Garcia, J. A. IREX: Iterative Refinement and Explanation of classification models for tabular datasets. *SoftwareX*, 2023, *23*, 101420, **1-7**. https://doi.org/10.1016/j.softx.2023.101420
48. Meding, K. & Hagendorff, T. Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms. *Philosophy and Technology*, 2024, *37*(1), 1–22. https://doi.org/10.1007/s13347-023-00679-8
49. Batko, K. & Ślęza, A. The use of Big Data Analytics in healthcare. *Journal of Big Data*, 2022, *9*(1), **1-24**. https://doi.org/10.1186/s40537-021-00553-4
50. Borisov, V.; Leemann, T.; Sessler, K.; Haug, J.; Pawelczyk, M. & Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, ***June*** 1–22. https://doi.org/10.1109/TNNLS.2022.3229161
51. Mbanaso, Uche M.; Abrahams, L. & Okafor, K.C. Data Collection, Presentation and Analysis. In: Research Techniques for Computer Science, Information Systems and Cybersecurity. *Springer, Cham* 2023, **115–138**. https://doi.org/10.1007/978-3-031-30031-8_7
52. Tjoa, E. & Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
53. Gajcin, J. & Dusparic, I. Redefining Counterfactual Explanations for Reinforcement Learning: Overview, Challenges and Opportunities. *ACM Computing Surveys*, 2024, *56*(9), **219:1-219:33**. https://doi.org/10.1145/3648472
54. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M. & Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, **2024**, *16*(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8
55. Lötsch, J.; Kringel, D. & Ultsch, A. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics*, **2022**, *2*(1), 1–17. https://doi.org/10.3390/biomedinformatics2010001
56. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L. & Zhong, **C.** Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, **2022**, *16*(none), 1–85. https://doi.org/10.1214/21-ss133
57. Zhong, X.; Gallagher, B.; Liu, S.; Kailkhura, B.; Hiszpanski, A., & Han, T. Y. J. Explainable machine learning in materials science. *Npj Computational Materials*, **2022**, *8*(1), 1–19. https://doi.org/10.1038/s41524-022-00884-7

58. Ekanayake, I. U.; Meddage, D. P. P.; & Rathnayake, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, 2022, *16*(April), e01059, **1-20**. https://doi.org/10.1016/j.cscm.2022.e01059
59. Hossain, M. I.; Zamzmi, G.; Mouton, P. R.; Salekin, M. S., Sun, Y., & Goldgof, D. Explainable AI for Medical Data: Current Methods, Limitations, and Future Directions. *ACM Computing Surveys*, 2023, *ACM 0360-0300/2023/12-ART* **1-41,** https://doi.org/10.1145/3637487
60. Leijnen, S.; Kuiper, O.; & van der Berg, M. Impact Your Future Xai in the Financial Sector a Conceptual Framework for Explainable Ai (Xai). *Hogeschool Utrecht, Lectoraat Artificial Intelligence, Whitepaper,* 2020, *v1.1*, 1-24 https://www.hu.nl/onderzoek/projecten/uitlegbare-ai-in-de-financiele-sector
61. Vellido, A.; Martín-Guerrero, J. D.; & Lisboa, P. J. G. Making machine learning models interpretable. ESANN 2012 Proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, **April** 2012, 163–172. https://www.esann.org/sites/default/files/proceedings/legacy/es2012-7.pdf
62. Alkhatib, A.; Ennadir, S.; Boström, H.; & Vazirgiannis, M. Interpretable Graph Neural Networks for Tabular Data. *ICLR 2024 Data-centric Machine Learning Research (DMLR) Workshop*, 2023, **1–35. https://openreview.net/pdf/60ce21fd5bcf7b6442b1c9138d40e45251d03791.pdf**
63. Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 2023, *263*, 110273, **1-24**. https://doi.org/10.1016/j.knosys.2023.110273
64. Pawlicki, M.; Pawlicka, A.; Kozik, R.; & Choraś, M. Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity. *Neurocomputing*, 2024, *590*(February), 127759, **1-13**. https://doi.org/10.1016/j.neucom.2024.127759
65. de Oliveira, R. M. B.; & Martens, D. A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences (Switzerland)*, 2021, *11*(16), **7274**, **1-28**. https://doi.org/10.3390/app11167274
66. Bienefeld, N.; Boss, J. M.; Lüthy, R.; Brodbeck, D.; Azzati, J.; Blaser, M.; Willms, J.; & Keller, E. Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals. *Npj Digital Medicine*, 2023, *6*(1), 1–7. https://doi.org/10.1038/s41746-023-00837-4
67. Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In: Koprinska, I., *et al.* ECML PKDD 2020 Workshops. ECML PKDD 2020. Communications in Computer and Information Science, vol 1323. Springer, Cham. **417–431** https://doi.org/10.1007/978-3-030-65965-3_28
68. Hartog, P. B. R.; Krüger, F.; Genheden, S., & Tetko, I. V. Using test-time augmentation to investigate explainable AI: inconsistencies between method, **model,** and human intuition. *Journal of Cheminformatics*, 2024 *16*(1), 1–20 https://doi.org/10.1186/s13321-024-00824-1
69. Srinivasu, P. N., Sandhya, N., Jhaveri, R. H., & Raut, R. (2022). From Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies. *Mobile Information Systems*, 2022, **167821, 1-20** https://doi.org/10.1155/2022/8167821
70. Rong, Y.; Leemann, T.; Nguyen, T. T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; & Kasneci, E. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, *46*(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846
71. Baniecki, H., & Biecek, P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 2024, *107*, *102303* **1-14.** https://doi.org/10.1016/j.inffus.2024.102303
72. Panigutti, C.; Hamon, R.; Hupont, I.; Fernandez Llorca, D.; Fano Yela, D.; Junklewitz, H.; Scalzo, S.; Mazzini, G.; Sanchez, I.; Soler Garrido, J.; & Gomez, **E.** The role of explainable AI in the context of the AI Act. *ACM International Conference Proceeding Series*, 2023 1139–1150. https://doi.org/10.1145/3593013.3594069
73. Madiega, T., & Chahri, S. *EU Legislation in Progress: Artificial intelligence act*. *March*. 2024, **1-12** https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf