Zacharias, Jan; von Zahn, Moritz; Chen, Johannes; Hinz, Oliver

Designing a feature selection method based on explainable artificial intelligence

Electronic Markets

Provided in Cooperation with:
Springer Nature

Suggested Citation: Zacharias, Jan; von Zahn, Moritz; Chen, Johannes; Hinz, Oliver (2022) : Designing a feature selection method based on explainable artificial intelligence, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 32, Iss. 4, pp. 2159-2184, https://doi.org/10.1007/s12525-022-00608-1

Electronic Markets (2022) 32:2159–2184
https://doi.org/10.1007/s12525-022-00608-1

RESEARCH PAPER

Designing a feature selection method based on explainable artificial intelligence

Jan Zacharias[1] · Moritz von Zahn[1] · Johannes Chen[1] · Oliver Hinz[1]

Received: 27 May 2022 / Accepted: 20 October 2022 / Published online: 12 December 2022

Abstract

Nowadays, artificial intelligence (AI) systems make predictions in numerous high stakes domains, including credit-risk assessment and medical diagnostics. Consequently, AI systems increasingly affect humans, yet many state-of-the-art systems lack transparency and thus, deny the individual's "right to explanation". As a remedy, researchers and practitioners have developed explainable AI, which provides reasoning on how AI systems infer individual predictions. However, with recent legal initiatives demanding comprehensive explainability throughout the (development of an) AI system, we argue that the pre-processing stage has been unjustifiably neglected and should receive greater attention in current efforts to establish explainability. In this paper, we focus on introducing explainability to an integral part of the pre-processing stage: feature selection. Specifically, we build upon design science research to develop a design framework for explainable feature selection. We instantiate the design framework in a running software artifact and evaluate it in two focus group sessions. Our artifact helps organizations to persuasively justify feature selection to stakeholders and, thus, comply with upcoming AI legislation. We further provide researchers and practitioners with a design framework consisting of meta-requirements and design principles for explainable feature selection.

Keywords Explainable artificial intelligence · Machine learning · Feature selection · Design science research · SHAP values · Preprocessing

JEL classification C8 · L1

Introduction

Many businesses and organizations create value through artificial intelligence (AI) systems (Müller et al., 2018; Abdel-Karim et al., 2021). Indeed, AI systems have shown to outperform humans in various fields, including breast cancer detection (McKinney et al., 2020), predictive maintenance (Paolanti et al., 2018), and credit-risk assessment (Khandani et al., 2010). Most AI systems apply methods from the field of machine learning (ML) (Kühl et al., 2020). Designing ML-based systems is an extensive process consisting of several stages. After the acquisition of suitable datasets, the data need to be thoroughly pre-processed before being used for model training (Abdel-Karim et al., 2021). One crucial step within data pre-processing is feature selection. Its basic concept is the separation of less relevant features within the dataset from relevant ones in order to save computational capacity, enhance the prediction performance, and make AI systems more understandable (Guyon and Elisseeff, 2003; Reunanen, 2003).

Recent legal initiatives demand transparency throughout the ML pipeline (e.g., the General Data Protection Regulation (GDPR) in the EU and the Algorithmic Accountability Act in the US). As a result, researchers have developed methods to render ML predictions transparent (Bauer et al., 2021a), but prior stages in the ML pipeline remain underexplored. This poses a risk to practitioners, as all stages are

Responsible Editor: Babak Abedin

✉ Jan Zacharias
zacharias@wiwi.uni-frankfurt.de

Moritz von Zahn
vzahn@wiwi.uni-frankfurt.de

Johannes Chen
jchen@wiwi.uni-frankfurt.de

Oliver Hinz
hinz@wiwi.uni-frankfurt.de

[1] Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 1, 60323 Frankfurt Am Main, Germany

presumably subject to new regulations. For example, the current proposal of the Algorithmic Accountability Act in the US views AI systems as multi-stage processes and aims at "impact assessments [for] particular stages" (Algorithmic Accountability Act of 2022). Another example is the GDPR in the EU, which demands practitioners to ensure "transparency with regard to processing of personal data" (General Data Protection, 2018) and thus refers to data pre-processing. Following the GDPR, practitioners should further strictly limit any data (pre-)processing to data that adds value ("principle of data minimization", General Data Protection, 2018). In an ML context, this ultimately implies a careful and well-justified feature selection. In addition, the GDPR calls for an individual's right to explanation, which implies "a more general form of oversight with broad implications for the design, prototyping, field testing, and deployment of data processing systems" (Casey et al., 2019, p. 180). This emphasizes the necessity for transparency not only for final ML outputs but also for prior stages. Overall, with numerous non-governmental guidelines going in the same direction (see, e.g. Jia and Zhang, 2022 for an overview) and further legal frameworks currently in discussion (e.g., the AI Act in the EU), there is an urgent need to extend the current efforts to make AI systems transparent to feature selection.

Besides legal guidelines, novel ethical frameworks also require a transparent and well-justified feature selection. A common ethical framework in the context of AI systems is called "transparency by design" (Felzmann et al., 2020; Plale, 2019; Koulu, 2021). Transparency by design requires AI practitioners to enable transparency throughout the development process proactively. Notably, merely introducing transparency to predictions of already implemented AI systems is regarded as insufficient (Felzmann et al., 2020). According to transparency by design, data processing and analysis (which includes feature selection) is one of three segments where developers should promote transparency. Other ethical frameworks imply similar requirements concerning feature selection (e.g., "trustworthy transparency by design", Zieglmeier and Pretschner, 2021). Hence, to adhere to ethical frameworks and remedy harmful consequences (e.g. disparate impact) of AI systems, a transparent and well-justified feature selection is essential (Kim and Routledge, 2022).

Legal initiatives and ethical considerations require organizations to explain their decisions during the feature selection process to regulators and other interest groups, i.e. developers must be able to justify why they deem certain features as relevant for model training while discarding others (Marcílio and Eler, 2020). However, extant techniques can hardly meet those explainability demands for several reasons: First, most extant techniques reduce a feature's impact to one single global importance value. In reality, a feature's impact can differ severely on the local level. For example, a feature may be irrelevant for the majority of observations, but highly relevant for certain subgroups (Lundberg et al., 2020). Second, extant methods do not reveal the direction of feature impacts (Lundberg et al., 2020). For example, a method may reveal that the age of a borrower is pivotal for predicting creditworthiness, but the nature of this relationship remains unclear: either younger or older age could favour predicted creditworthiness, and it might even be a complex, non-linear relationship. Finally, holistic importance scores also do not show how features interact with each other (Lundberg et al., 2020). In some cases, a feature might be irrelevant for a prediction task on its own while being highly relevant in combination with another predictor (Guyon and Elisseeff, 2003). We argue that in order to fulfill legal and ethical requirements, organizations should use a feature selection method that is explainable in the sense that it provides ML developers such in-depth information about feature impacts. Following this argument, we define explainable feature selection as a process that enables the developer or organization to provide stakeholders persuasive explanations of feature impacts—including a feature's global and local importance, impact direction, and interaction effects—to justify feature selection-related decisions.

Current research on explainable artificial intelligence (XAI) focuses on the interpretability of opaque ML models (Arrieta et al., 2020). The primary goal is to explain the logic behind a model's predictions, which are otherwise incomprehensible for human users (see, e.g., Chakrobartty and El-Gayar, 2021; Fernandez et al., 2022; Cirqueira et al., 2021; Zhang et al., 2020). These XAI techniques mostly address predictions of already trained models; literature that addresses the explainability and justifiability of ML pre-processing, let alone feature selection, is scarce (Marcilio and Eler, 2020). However, neglecting the explainability of ML pre-processing inevitably leads to insufficient explainability and justifiability of the ML-based system as a whole. Consider a ML model predicting the creditworthiness of loan applicants: applying XAI techniques on the final ML model may explain the relationship between the model's input features and the outcomes; however, the exact reasons for incorporating these features in the first step remain unclear to the applicant and other stakeholders. Given the aforementioned legal and ethical requirements for transparency throughout the ML pipeline, we regard this as a major research gap. In this work, we build upon design science research (DSR) to develop a design framework for feature selection based on XAI, which both ensures compliance with ethical and legal frameworks and meets the needs of ML developers. Furthermore, we instantiate this design framework in a running software artifact and evaluate it within focus group sessions with ML practitioners and researchers. With this approach, we aim to answer the following research questions:

1. What design principles should a design framework for an explainable feature selection method include?
2. How do practitioners evaluate a software artifact, which applies the concept of explainable feature selection?

By answering these research questions, we aim to contribute to both theory and practice in three ways. First, with our novel design framework, we contribute a "theory for design and action" (Gregor, 2006) on building systems for explainable feature selection. Second, our work contributes to practice by proposing a running software artifact, which organizations may adopt to implement the concept of explainable feature selection. Third, we contribute to XAI literature by extending the current scope of applications of XAI methods. We demonstrate that developers may not only use XAI methods to justify given ML-based predictions (Adadi and Berrada, 2018; Meske et al., 2022) but may even use XAI to justify the preceding feature selection.

The remainder of this work is structured as follows: In the next section, we provide background information on XAI and feature selection. In the third section, we present related work on XAI in DSR as well as existing tools for feature selection. In section four, we outline our design science methodology. Following that, in the fifth section, we describe the design process of our artifact in detail as well as our evaluation strategy. In the sixth section, we present the empirical evaluation results. In section seven, we discuss the empirical results and their implications. In the final section, we reflect on our work and summarize our main contributions as well as limitations.

Background

Explainable artificial intelligence

Current literature describes various approaches to explain the decisions of opaque ML models. These approaches can be generally categorized as intrinsic interpretability and post-hoc explanations. Intrinsic interpretability aims to develop ML models that are inherently transparent and do not require explicit explanations (Du et al., 2019). Conversely, post-hoc explanations aim to increase the interpretability of black-box models (Arrieta et al., 2020), such as ensemble methods or deep neural networks. In contrast to intrinsic interpretability approaches, post-hoc explanations do not disclose the inner workings of black-box models; instead, they utilize other explanation techniques (Phillips et al., 2020; Lipton, 2018). These techniques include feature attribution methods, textual explanation or visualizations (Lipton, 2018).

Depending on their generalizability, post-hoc explanations can be further divided into model-agnostic and model-specific explanations. Model-agnostic explainability refers to techniques that are applicable to any kind of model (Ribeiro et al., 2016). One example is Local Interpretable Model-Agnostic Explanations (LIME, Ribeiro et al., 2016), which locally approximates a black-box model with an intrinsically interpretable one. In contrast, model-specific explanations are only applicable to specific model types. A well-known example is the embedded feature importance function of tree-based models (Du et al., 2019).

XAI methods can provide global or local explanations. Global methods holistically explain the model at hand (Teso and Kersting, 2019). For instance, global feature attributions aim to reflect how the model weights its input variables. However, focusing solely on global explanations bears the risk of missing useful information. Specifically, for a single prediction of interest, features that are meaningful on the global level are not necessarily important, while features, which seem globally meaningless may be crucial (Ribeiro et al., 2016). By contrast, local methods zoom into model predictions and reveal how features have driven an individual prediction (Du et al., 2019, p. 3).

One of the most popular XAI methods is the local feature attribution method Shapley Additive Explanations (SHAP) (Schlegel et al., 2019; Dunn et al., 2021). SHAP is based on Shapley Values, a game-theoretic approach to find a fair distribution of payouts to each player within a cooperative game according to each player's contribution to the overall payout.[1] Transferred to feature selection, the model output represents the cooperative game and the features represent the players who work together to create the model output. In that sense, SHAP provides local explanations by computing the contribution of each feature to any given prediction, the so-called SHAP value. This is best explained considering the following example: we assume an AI system predicting the risk of credit default (between 0 and 1) using several different features. For each of the features of a given borrower, we can compute a SHAP value indicating how that particular feature has driven the prediction. We assume that the feature "monthly salary" equals 8,000 USD and the corresponding SHAP value is –0.2. This means that the fact that the borrower earns 8,000 USD per month decreases the predicted risk of credit default by 0.2.[2] One can aggregate SHAP values to provide global explanations: To this end, the average absolute contribution of a feature across all predictions is computed. The result thus represents the average importance of a feature for model predictions. SHAP also reveals interaction effects between features, the Shapley interaction index (Lundberg et al., 2018, p. 29). The idea is

[1] for the original paper from Lloyd S. Shapley, see Shapley (1953).

[2] One can compute the SHAP values either in log-odds units or in probability units; the above explained interpretation only applies in case of probability units.

as follows: Marginal contributions cannot only be computed for single features but also for feature pairs. To separate the interaction effect of a feature pair from the feature's main effects, one must subtract the marginal contributions of the single features from the marginal contribution of that pair. The resulting difference represents the interaction effect of the feature pair.

Feature selection

Nowadays, practitioners and researchers deal with vast datasets containing hundreds to several thousands of features (Guyon and Elisseeff, 2003). However, a high-dimensional feature space may also pose problems: The number of training instances required to reach a desired model performance, also called "sample complexity" (Blum and Langley, 1997), grows exponentially with increasing features (Verleysen and François, 2005; Bessa et al., 2017). Yet, in practice, organizations often face the problem of data scarcity: there is either a general lack of training data or available training instances are incomplete or unlabelled (Li et al., 2021). This leads to the so-called "curse-of-dimensionality" (Bach, 2017; Seo and Shneiderman, 2005).

Variables usually vary in their importance for an ML task. Oftentimes, most of the variation in the target variable can be explained using a fraction of the initial input variables (Dunn et al., 2021). In fact, there is even a risk that irrelevant or highly correlated features lead to a substantial loss of performance (Kohavi and John, 1997). To overcome these problems, developers apply feature selection methods within the data pre-processing phase. Feature selection is a dimensionality reduction technique that aims to separate a subset of relevant features from less relevant ones whilst maintaining high prediction performance (Zhang et al., 2018, Chandrashekar and Sahin, 2014). There are various reasons for this task: The training and predictions of the model become computationally less expensive, the prediction power can be enhanced, and the interpretability increases (Guyon and Elisseeff, 2003; Fryer et al., 2021; Bhandari et al., 2020). However, distinguishing relevant features from irrelevant ones is non-trivial which is why numerous methods for feature selection exist (see, e.g., Chandrashekar and Sahin 2014). The choice of the respective method heavily influences the final model's behavior (Abdel-Karim et al., 2021), so it is crucial for ML developers to know and understand the existing methods in order to explain their decisions to other stakeholders including regulators.

Feature selection for supervised learning tasks can be categorized as (i) filter- (ii) wrapper- or (iii) embedded methods (Chandrashekar and Sahin, 2014). Filter methods assess the relevance of features based on inherent characteristics of the data (Kohavi and John, 1997), producing a ranked list of all features (Mlambo et al., 2016). Wrappers, on the other hand, evaluate the relevance of feature subsets using the ML model (Kohavi and John, 1997). The idea is simple: the subset that produces the best prediction performance on the respective model is considered to be the most relevant. Finally, embedded feature selection methods are built-in components of specific ML algorithms and rank features as part of the training process. Common examples are the built-in feature importance values of tree-based algorithms such as decision trees, random forests or gradient boosting (Guyon and Elisseeff, 2003).

At this point, it should be noted that feature selection is not necessarily conducted in a purely data-driven manner; these methods can (and should) be complemented by "domain theory and overarching ethical principles intended to embed fairness in the resulting models" (Maass et al., 2018). For instance, domain experts could expose automatically extracted patterns as spurious (Bentley et al., 2014).

Related work

Explainable artificial intelligence in design science research

Recent literature within the field of DSR shows growing interest for the design of XAI-based artifacts. Meske and Bunde (2022) tackle the problem of hate speech in modern social media. To help social media moderators detect problematic content, the authors propose design principles for an AI-based decision support system and instantiate them in the form of an user interface. In addition, the artifact leverages local XAI techniques to further aid users in their decision making process. The authors evaluate and refine their artifact in three consecutive design cycles. Cirqueira et al. (2021) propose a design framework for an XAI-based decision support system for fraud detection within the financial services industry. They argue that existing XAI-based fraud detection studies neglect a user-centric perspective and, therefore, integrate the concept of user-centricity in their design framework. They use an instantiation of the design framework in the form of a mockup user interface for the artifact evaluation. Zhang et al. (2020) address the problem of fake news on social media in their DSR project. The authors design a ML framework for the prediction of fake financial news and instantiate their framework by developing the system architecture of a fake news detection system. In addition, Zhang et al. (2020) describe the feature selection process during their ML pre-processing: They conduct a wrapper approach by evaluating different feature sets according to the performance of the resulting ML model. After the ML model is completed, the authors leverage a local XAI technique to demonstrate the importance of each feature for specific predictions. Schemmer et al. (2022) identify the need for

explainability techniques for AI-based decision support in high-stakes decision making, such as real estate purchase. To this end, the authors propose design principles for an XAI-based real estate valuation artifact and instantiate it by developing a prototype of an user interface. The explainability component of the artifact includes both global feature importances and example-based explanations of feature impacts.

Despite the growing number of DSR projects that design XAI-based artifacts, we identify a major research gap in the current literature: All related works that we reviewed aim to enhance the explainability of already completed ML models. To the best of our knowledge, there is no work addressing the explainability and justifiability of the pre-processing of ML models. Consequently, we regard the DSR knowledge base for explainable ML pre-processing artifacts as a research gap that needs to be filled.

Tools for feature selection

Several research papers propose (interactive) feature selection tools. One example is the visual analytics tool called INFUSE by Krause et al. (2014): The authors consider the problem of working with high-dimensional datasets where deciding which feature selection method and, subsequently, which classification algorithm to use becomes a non-trivial task. INFUSE offers several feature visualizations and enables the user to interactively add or remove features to compare their impact on the final model quickly. However, although INFUSE enables the user to efficiently compare several feature selection algorithms, none of these algorithms convey information such as a feature's local importances, influence direction or interaction effects. Zhao et al. (2019) aim to enhance users' trust in black-box models and simultaneously improve the model's prediction performance. To this end, they designed the interactive feature selection system FeatureExplorer. Their tool performs feature selection and subsequently model evaluation in an iterative manner with the user. FeatureExplorer conveys information mainly via correlation analyses as well as feature rankings produced by two feature selection algorithms. However, with the main focus of FeatureExplorer being prediction performance improvement, it mostly neglects the explainability of feature selection. The interactive visual analytics tool Prospector by Krause et al. (2016) focuses on explaining ML models by investigating the contributions of single features to the model's output. The tool involves both a global and a local view such that users can inspect the importance of single variables for specific data points. Thus, Prospector aims to explain the association between features and the target variable. However, building on the concept of partial dependence plots, Prospector does not depict local importances or interaction effects concisely.

Overall, the current literature offers several tools that facilitate the feature selection process. However, extant tools mainly focus on the improvement of the prediction performance of ML models. The tools mostly ignore legal and ethical requirements for explainable and justifiable feature selection. To the best of our knowledge, no feature selection tool exists that systematically uses XAI techniques to produce detailed explanations of feature impacts to meet these legal and ethical requirements.

Research methodology

Based on design science principles, we aim to fill the aforementioned research gap by proposing and empirically evaluating a novel method for explainable feature selection. According to Hevner et al. (2004), design science "creates and evaluates [information systems (IS)] artifacts intended to solve identified organizational problems". Accordingly, we create a design framework for general XAI tools intended to resolve the problem of opaque feature selection procedures and the consequential inability to justify feature selection-related decisions. We further instantiate this framework into a running software artifact and empirically evaluate this instantiation. As a methodological foundation, we follow the design research cycle introduced by Kuechler and Vaishnavi (2008).

The design science approach of Kuechler and Vaishnavi (2008) proposes an iterative procedure to ensure continuous evaluation and subsequent adaptation of the artifact. Each iteration of the design cycle consists of five phases: problem awareness, suggestion, development, evaluation, and conclusion. In the problem awareness phase, the researchers may review literature, interview practitioners or investigate related existing artifacts to identify and precisely define the problem to be solved. In the suggestion phase, the researchers derive meta-requirements based on the problem definition and formulate design principles for the artifact that address the meta-requirements (Meth et al., 2015). These design principles should be grounded in scientific theories and the expertise of the design scientists (Hevner et al., 2004). Next, in the development phase, the researchers construct the artifact. While commonly understood as actual software instantiations, artifacts can also take the form of methods, models or constructs (March and Smith, 1995). After its development, one must evaluate the artifact according to objective criteria. This can take the form of expert workshops (Meth et al., 2015), laboratory experiments (Gnewuch et al., 2017), online experiments (Kellner et al., 2021) or quantitative evaluations (Toreini et al., 2022). Finally, the researchers conclude the project and – if necessary – transfer the evaluation results into the next iteration.

Designing explainable feature selection

In our DSR project, we performed one design cycle to develop a prototype for explainable feature selection (see Fig. 1). To be precise, we reviewed both scientific literature and legal documents to derive meta-requirements and design principles. Afterwards, we instantiated our design principles as an actual running software artifact that we evaluated with experts in two focus group sessions. Finally, we concluded the design cycle by analysing the results from the focus groups. Moreover, we completed the first steps of a second design cycle by refining the addressed problem as well as formulating new design principles based on the results of the first cycle. In the following, we explain each phase in more detail.

Awareness of problem

As mentioned before, the need for transparent—and thus explainable—feature selection is driven mainly by legal requirements and ethical considerations. However, organizations should not regard explainability as a mere regulatory hurdle; instead, they should regard it as an opportunity to create more meaningful models. From a developer's point of view, a better understanding of an algorithm's inner workings and structures facilitates the optimization thereof (Meske et al., 2022; Murdoch et al., 2019). This may lead to more accurate ML models and, consequently, to better business outcomes. Explainability might also increase user's trust in AI systems and thereby raise technology adoption (Abedin, 2021).

We identify three main problem areas related to the design of explainable feature selection: First, most extant feature selection methods reduce feature impacts to one holistic value. Second, eliminating features from the dataset may impair prediction performance. Finally, XAI methods often put too little focus on the end-user. Starting with the first problem: Most feature selection techniques—such as the Pearson correlation coefficient (Blum and Langley, 1997) or embedded functions of tree-based algorithms (Strobl et al., 2007)—compute a holistic importance value for each feature. These scores claim global validity, meaning that the ML model would weigh the input variables equivalently across all predictions. However, a feature's contribution usually differs across individual- or subgroups of predictions in reality (Lundberg and Lee, 2017). Additional insights—such as the direction of impact for each feature, and interactions—cannot be represented by global importance values as well (Lundberg and Lee, 2017). This information shortage is problematic when regulators demand

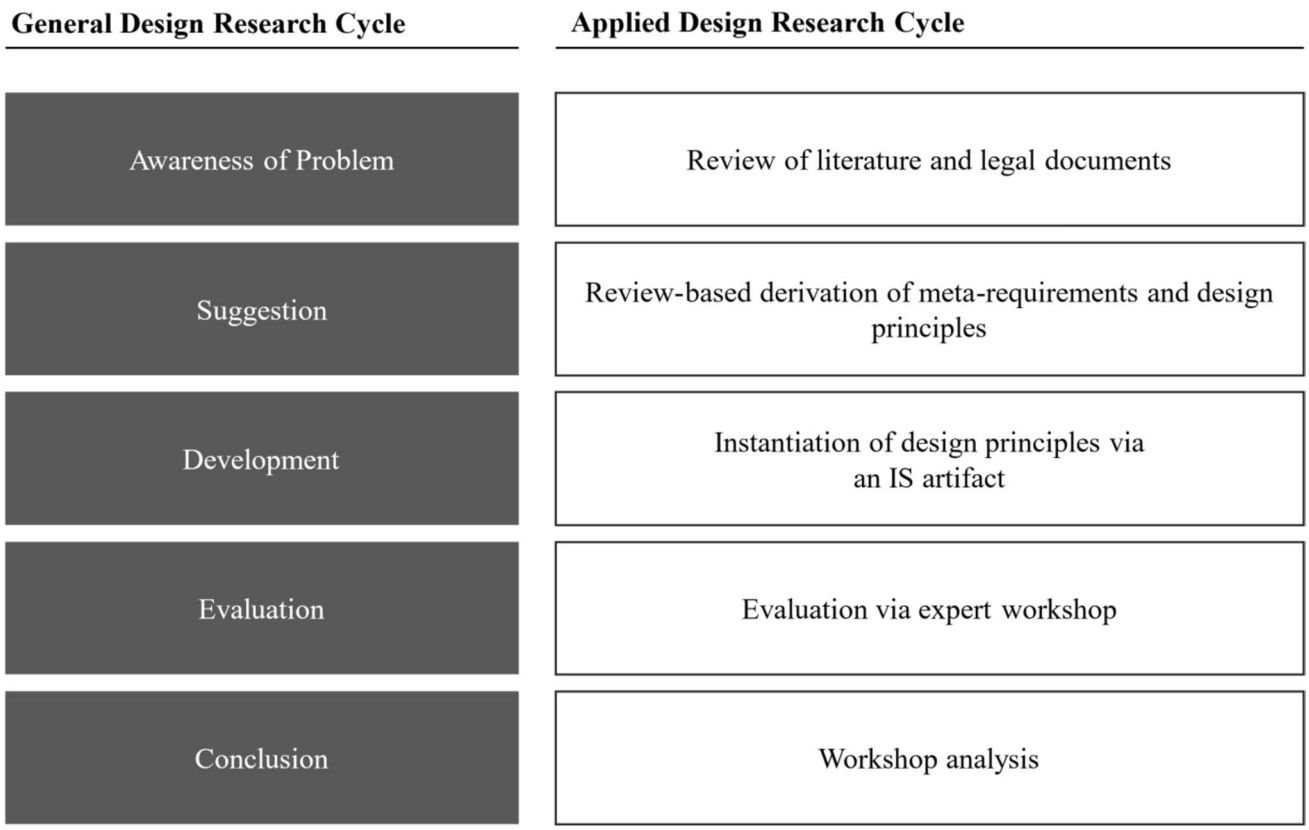

Fig. 1 Our research methodology based on Kuechler and Vaishnavi (2008)

in-depth explanations on how certain features contribute to automated decisions.In regard to the second problem: Removing subsets of a dataset inevitably leads to a loss of information (Abdel-Karim et al., 2021) which, in turn, may impair the predictive performance of the final ML model. If a feature selection method is incapable of efficiently differentiating between relevant and irrelevant features, the resulting performance impairment outweighs the benefits of high explainability. Our third identified problem refers to the user's role in explainability methods: In general, explainability methods aim to inform a human user how a prediction model generates its outputs. Still, researchers often criticize XAI systems for failing to incorporate user-centricity into the system's design (Förster et al., 2020). This may have detrimental consequences, such as reduced trust and lower technology adoption (Förster et al., 2020).

Suggestion

Based on our awareness of the problem, we identified several requirements and design principles that address these issues. In addition, we derived further design principles based on the practitioners' inputs during the focus group analyses in the conclusion stage. Figure 2 presents an overview of the meta-requirements and design principles of both the suggestion stage and the conclusion stage. For the formulation of our design principles, we followed the conceptual schema and terminology for design principles of Gregor et al. (2020). We define the actors of our design principles, to whom we refer in our formulations, as follows:

1. *Implementer*: Software developers and designers who instantiate our design framework within their organiza-

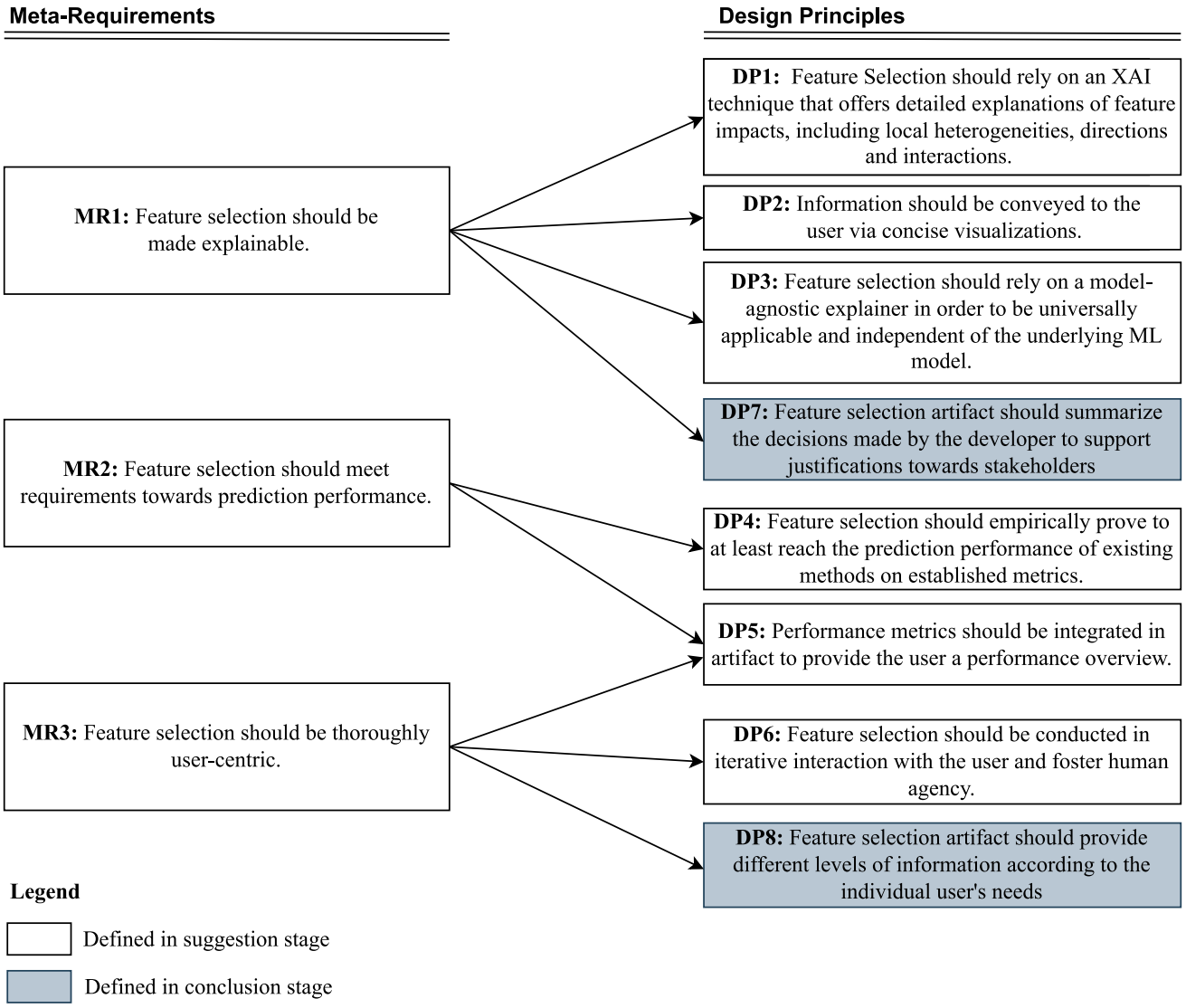

Fig. 2 Design framework consisting of meta-requirements and design principles

tion in order to provide a tool for explainable feature selection.
2. *User*: ML developers who apply the instantiation of our design framework to conduct explainable feature selection.

Subsequently, we provide an overview of the design principles generated during the suggestion stage; the design principles derived in the conclusion stage are presented in the results section.

The first meta-requirement of our proposed feature selection method refers to its explainability. As discussed in the previous sections, a human supervisor should gain detailed insights regarding a feature's impact on the prediction task. This leads to our first meta-requirement (MR):

MR1: Feature selection should be made explainable.

One popular means to produce explanations for opaque models are feature attribution methods (Senoner et al., 2021). Feature attributions represent the contribution of a feature to the model output, and thus aim to reflect the feature's importance for the prediction task. As explained in Sect. 2.3, global feature attributions show the importance of features for the model as a whole whereas their local counterpart depict feature importances on the level of individual predictions (Lundberg and Lee, 2017). Consequently, local feature attributions convey more granular and detailed information than the global methods (Murdoch et al., 2019). A more detailed understanding of a feature's impact may also be achieved when informing the developer about interaction effects and directions of feature effects (Lundberg et al., 2020). Therefore, implementers should integrate in-depth information in an explainable feature selection artifact such that users may incorporate this information into their argumentation when justifying feature selection-related decisions. Taking these aspects together, we define our first design principle (DP) as: Feature selection should rely on an XAI technique that offers detailed explanations of feature impacts, including local heterogeneities, directions and interactions **(DP 1)**. In addition, besides the level of detail, the way information is presented is crucial (Murdoch et al., 2019). Whereas textual or numerical explanation formats are suitable for explaining simple concepts, comprehensible explanations of more complex concepts should rely on visual formats (Wang et al., 2019). The choice of visualization techniques should depend on the concept to be explained. For instance, line charts are suitable for representing raw time series data, bar charts are suitable for representing feature attributions, and partial dependence plots are suitable to demonstrate how feature attributions depend on a feature's value (Wang et al., 2019). Implementers should ensure that users may explain feature selection-related decisions in a way that is comprehensible for the recipients of the explanations. One means to achieve this is the integration of concise visualizations into the artifact. In summary, our second design principle is: Information should be conveyed to the user via concise visualizations **(DP 2)**. To ensure that our artifact is universally applicable, we argue that the implementer should use a model-agnostic XAI technique. This entails several benefits: Nowadays, numerous different prediction models exist and are applicable in practice; using a model-agnostic XAI technique, the user can apply the artifact to any given prediction model (Ribeiro et al., 2016). Moreover, the recipient of the explanation does not require deep technical expertise in specific ML models as the explanations are independent from the inner workings of the respective ML model. Therefore, our third design principle is: Feature selection should rely on a model-agnostic explainer in order to be universally applicable and independent of the underlying ML model **(DP 3)**.

The second meta-requirement addresses the prediction performance issue. Users (i.e., ML developers) can minimize the prediction performance impairment of the ML model when the applied feature selection method successfully separates relevant features from less relevant ones. If a dataset contains features that are entirely irrelevant for the specific prediction task, the user might even improve the model's accuracy (Reunanen, 2003). Thus, we propose our second meta-requirement:

MR2: Feature selection should meet prediction performance requirements.

ML developers typically assess ML models according to established performance metrics. Such metrics steer the development process and facilitate the objective comparison of competing classifiers (Asatiani et al., 2021). We argue that the implementer should use a feature attribution algorithm that has empirically proven to minimize the prediction performance impairment of the final ML model. Otherwise, potential users may not apply the artifact in their organizations. Therefore, our fourth design principle is: Feature selection should empirically prove to at least reach the prediction performance of existing algorithms on established metrics in order to meet performance requirements of the developer **(DP 4)**. In some cases, users might not entirely avoid the negative impact of feature selection on predictive performance. Still, the benefits of reduced dimensionality, i.e. higher explainability, saving of computational resources, etc., may justify a certain prediction performance reduction. The user, however, should be fully aware of potential performance impairments. Therefore, we regard the integration and display of established performance metrics in order to provide the user a performance overview **(DP 5)** as a further design principle.

Finally, we address the user-centricity issue. There are several drivers for user-centricity in XAI systems design:

First, it enhances human agency; if users are not able to effectively develop and interact with XAI systems, they might distrust the system or fall into the automation bias trap (Förster et al., 2020; Herse et al., 2018). Furthermore, the provision of model explanations might convey valuable information to the user; however, in order to leverage those explanations to adapt and improve the system, the user must be in the centre of the development process (Pfeuffer, 2021). This leads to our third meta-requirement:

MR3: Feature selection should be thoroughly user-centric.

One core principle for user involvement is the interaction between knowledgeable humans and the computer during model design (Förster et al., 2020; Kulesza et al., 2015; Pfeuffer, 2021)**.** This human computer interaction can be further enhanced by an iterative mode (Baum et al., 2020). For instance, an XAI method could present explanations for decisions and the developer subsequently controls and corrects these decisions. The XAI system, in turn, considers these corrections in the next explanations, which the developer again controls and so forth. We argue that the implementer should design the artifact in a way that the user can actively steer the feature selection process and have the last say at each decision. This motivates our last design principle: The feature selection process should be conducted in iterative interaction with the user and foster human agency **(DP 6)**.

Development

In this section, we describe the development of our artifact that takes the form of an instantiation of the aforementioned design framework. To this end, we must first determine which feature selection algorithm is most suitable for our task and then embed it into a software artifact that meets our design requirements. As explained in the design principles, we need a local feature attribution method that conveys in-depth information, including holistic feature importances, local feature importances, and interaction effects. In addition, the feature attribution method should be model-agnostic and thus universally applicable. We identified SHAP as the most suitable method since its analyses meet all these requirements.

To assess empirically how SHAP values-based feature selection affects the performance of an ML model, we conducted several simulations on six publicly available datasets (Appendix 2 Table 2 presents an overview of the datasets). To this end, we performed a backward elimination approach based on SHAP feature rankings (for more details, we refer to Appendix 4). We evaluated this approach on two dimensions: First, stability—referring to the sensitivity of feature rankings to small perturbations of the training data—and second, the predictive performance of the final classifier. As comparison benchmarks, we used analogous backward elimination processes based on (i) the Pearson correlation coefficient and (ii) the embedded feature importance function of the XGBoost algorithm. Our results indicate that SHAP-based feature rankings are considerably more stable than those of the XGBoost built-in function, but less stable than the Pearson correlation coefficients. Regarding the predictive performance, SHAP yields at least similar results as the benchmark methods on established performance metrics (see Appendix). These results coincide with related comparative studies that evaluate the predictive performance of SHAP-based feature selection (e.g. Xiaomao et al., 2019; Bhandari et al., 2020; Effrosynidis and Arampatzis 2021).

Having identified SHAP as suitable to perform feature selection, we developed a software artifact that effectively leverages this technique in interaction with a human developer. For this task, we used the Tk GUI toolkit via the programming language python. Although this software artifact represents a prototype of an explainable feature selection tool, it is already operational and usable in practice. Our artifact is split into two parts: In the first part, the tool initially ranks all features according to the global SHAP values and eliminates all features except the k best ones, with k being determined by the user. This represents a radical initial dimensionality reduction of the dataset. After the dataset is reduced to a human-manageable size, the second part—an iterative backward elimination process—follows.

Figure 3 illustrates the artifact's procedure: First, the user needs to specify the test set size for the XGBoost model training, the number of remaining features k, and whether she conducts test set discrimination.[3] Following that, the tool trains an XGBoost model, computes SHAP values and sorts the features according to their global importance scores. Note that our artifact may implement any other ML model besides XGBoost since SHAP is a model-agnostic explainer. Based on the global importance scores, the tool eliminates all features from the dataset except the k highest ranked ones and trains a new XGBoost model based on that reduced dataset. The user can inspect the different SHAP plots of the new XGBoost model as well as performance metrics of the current and the previous models. Figure 4 shows the output of that process. After the developer analyzes all provided pieces of information, she can then decide whether she continues the feature selection and, in case of continuing, which feature she eliminates next. After the elimination, the tool trains a new XGBoost model and presents the same

[3] Test set discrimination refers to the aggregation of local SHAP values to the global importance scores. Instead of averaging the whole test set, one may also select the SHAP values of specific observations as an aggregation basis. For a more detailed explanation, see Appendix 1.

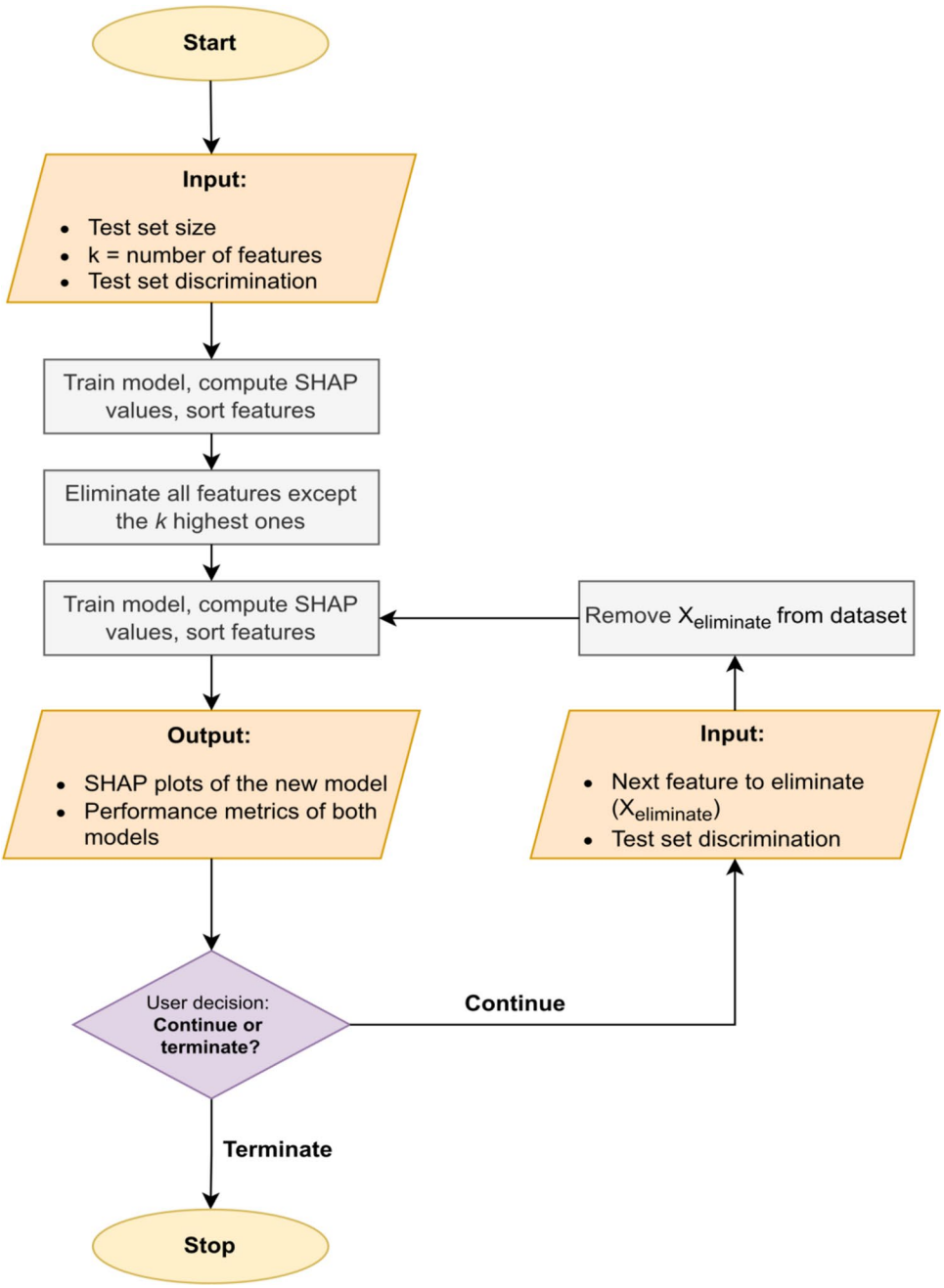

Fig. 3 Flowchart of the artifact procedure

analyses to the developer. At each iteration, the developer can reassess how each feature elimination affects the model performance and feature impacts. As soon as an individually defined stop criterion is triggered (e.g. the intended number of features is reached), the developer can terminate the procedure; the tool returns the final model as well as a list of the feature sets in each iteration for further usage.

Evaluation

We evaluate both our design framework and its instantiation following the Framework for Evaluation in Design Science Research (FEDS, Venable et al., 2016). FEDS describes different evaluation strategies based on two dimensions: The functional purpose as well as the paradigm of the evaluation study. The first dimension, the functional purpose, refers to the reason for evaluation. Does the researcher conduct the evaluation to derive potential improvements of the artifact (*formative evaluations*), or to assess the extent to which the artifact meets performance expectations in practice (*summative evaluations*)? The second dimension, the paradigm of the evaluation, differentiates between artificial evaluation settings on the one hand, such as laboratory experiments, simulations and alike, and naturalistic settings on the other

Fig. 4 Screenshot of the running software artifact

hand, which typically refers to testing in the real environment, such as the organization of the artifact's target group.[4] One goal of the evaluation of our DSR project is the identification of improvement potentials and the derivation of additional design principles for upcoming design cycles. Thus, we decided for a *formative* evaluation strategy. Furthermore, to assess our software artifact without the interference of confounding variables and due to resource constraints, we conducted the evaluation in an *artificial* environment. With our approach, we do not only evaluate the instantiation of our design framework, i.e. the software artifact, but also implicitly evaluate the design framework as such. Thereby, we follow Hevner et al. (2004, p. 84) who argue that the "[a]rtifact instantiation demonstrates feasibility both of the design process and of the designed product". In the following, we describe our evaluation approach in more detail.

To qualitatively evaluate our artifact, we conducted two focus group sessions with subject matter experts from two leading technology consultancy firms as well as researchers (n = 12). We decided to conduct focus group research since

[4] For a detailed explanation of the FEDS, we refer to the original paper of Venable et al. (2016).

Table 1 Descriptive overview of the focus group participants

ID	Job title	Gender	ML experience (in years)	Self-assessed ML competence	Age
1	Data Scientist/ Consultant	male	3	Rather high	25–30
2	Data Scientist/ Consultant	female	3	Rather high	25–30
3	Data Scientist/ Consultant	male	2	Rather high	25–30
4	Data Scientist/ Consultant	male	3	Mediocre	30–35
5	Data Scientist/ Consultant	male	3	Rather high	25–30
6	Data Scientist/ Consultant	male	>3	Rather high	>40
7	Data Scientist/ Consultant	male	<1	Rather low	25–30
8	ML researcher	male	2	Mediocre	18–25
9	Data Scientist/ Consultant	male	>3	Rather high	25–30
10	ML researcher	male	>3	High	25–30
11	Data Scientist/ Consultant	male	3	Mediocre	30–35
12	Data Scientist/ Consultant	male	>3	High	25–30

the IS community regards this method as well-suited for exploring new IS concepts (Belanger, 2012). The purpose of these focus group sessions was to both gather qualitative data on the extent to which our artifact achieves its desired goal of making feature selection explainable and justifiably towards stakeholders, and to discover further design principles for upcoming design cycles. All participants are experienced professionals and work with ML on a daily basis. Table 1 provides an overview of the participants.

The focus group sessions lasted around 1.5—2 h and consisted of four stages. First, we held an introductory presentation explaining the conceptual foundations of SHAP as well as feature selection and presented our artifact. Second, each participant carried out hands-on feature selection tasks using our artifact on their own computer. For this task, we used the German Credit dataset from the UCI Machine Learning Repository.[5] We instructed the participants to conduct at least two feature selection processes. They had approximately fifteen minutes for this procedure. Third, the participants evaluated the artifact on multiple constructs via a Likert scale-based questionnaire. The primary purpose of the questionnaire is creating a basis for the upcoming feedback discussion. In this discussion, we could get more in-depth comments and feedback regarding the constructs and the tool. In addition, we wanted to get a more holistic overview of user perception and derive themes that could inform the design requirements and principles of the second design cycle.

There is no clear consensus among IS researchers on which constructs are most suitable for evaluating design artifacts (Prat et al., 2014). Prat et al. (2014) conducted a comprehensive meta-study covering the most frequently used evaluation constructs within IS research. Our primary evaluation goal is to assess whether our artifact properly fulfills its original purpose, i.e. making feature selection explainable and justifiable. Therefore, we regard the effectiveness (defined as "[…] the degree to which the artifact meets its higher level purpose or goal and achieves its desired benefit in practice" (Venable et al., 2012, p. 426)) as central evaluation criterion. Building upon prior research (e.g., Meth et al., 2015, Venable et al., 2012), we address the effectiveness criterion with five questions that elicit the practitioners' view on how well they understand each feature impact on the ML model and the justifiability of the feature selection process towards stakeholders. To gain a more detailed view on the practitioners' assessment of our artifact, we selected additional evaluation criteria from papers summarized by Prat et al. (2014) and extended them with constructs from two other papers (Chen and Koufaris, 2015; Komiak and Benbasat, 2006). These constructs include, amongst others, perceived usefulness (defined as the "[…] degree to which a person believes that using a particular system would enhance his or her job performance" (Davis, 1989)) and ease of use (defined as "[…] the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989)). To formulate our concrete questionnaire questions, we adapted the constructs of the respective papers according to our use case. All questions are answered on a 5-points Likert scale (1 = fully disagree; 5 = fully agree). An overview of our constructs, our motivation for choosing them, and the papers with the original questions is provided in the Appendix 5.

For the qualitative analysis of the focus group discussion, we first created the transcript of the discussion sessions. Then, we investigated codes and general themes of the discussion via conventional content analysis of the transcript. Conventional content analysis is particularly suited to study the meaning of text data when existing theory and research literature is limited (Hsieh and Shannon, 2005). We based

[5] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

our qualitative analysis on Gioia et al. (2013). Briefly summarized, this approach works as follows: In the first order analysis, we identified all relevant terms and codes; at this point, we strictly adopted the wording of the participants without interpretation. Next, we constructed second order themes by clustering and interpreting these codes. Two researchers conducted the coding and interpretation independently and afterwards discussed their results, guided by an unbiased third researcher, to reach a consensus. We conducted this approach iteratively until we agreed on the first order codes and second order themes. This allowed us to derive new insights from the discussion systematically and formulate new design principles based on those insights.

Results

The qualitative results of our focus group sessions were our focus of the evaluation phase. Figure 5 presents the data structure following Gioia et al. (2013), which depicts the derived codes and themes. In our sample of twelve participants, nine experts participated in the discussion. As discussion guidance, we asked semi-structured questions related to the constructs that we measured in the Likert scale-based questionnaire. The participants, however, were free to talk about any aspect of the tool and were encouraged to make suggestions related to desired features and improvements.

A recurring theme that emerged during the two focus group sessions was the accessibility of the presented tool. Most participants agreed that it was easy to learn for all levels of expertise. In addition, one expert said that because data scientists do not need to implement the tool themselves, the tool would be useful in practice.

> *"I immediately thought this would be helpful, especially for non-professional data scientists, since everything is going towards democratization and 'build your own model'."* (Expert 3)

Expert 3 mentioned that he could imagine the tool to be particularly useful for non-professional data scientists, e.g. people who do not constantly work on data science problems. He stated that currently, many helpful tools exist that support data scientists of all experience levels; our tool, he continues, could be particularly useful for feature selection moving towards democratization of data science. By contrast, another participant found the GUI input fields too restrictive. He suggested adding a coding interface that allows users to input their own code instead of using the input fields.

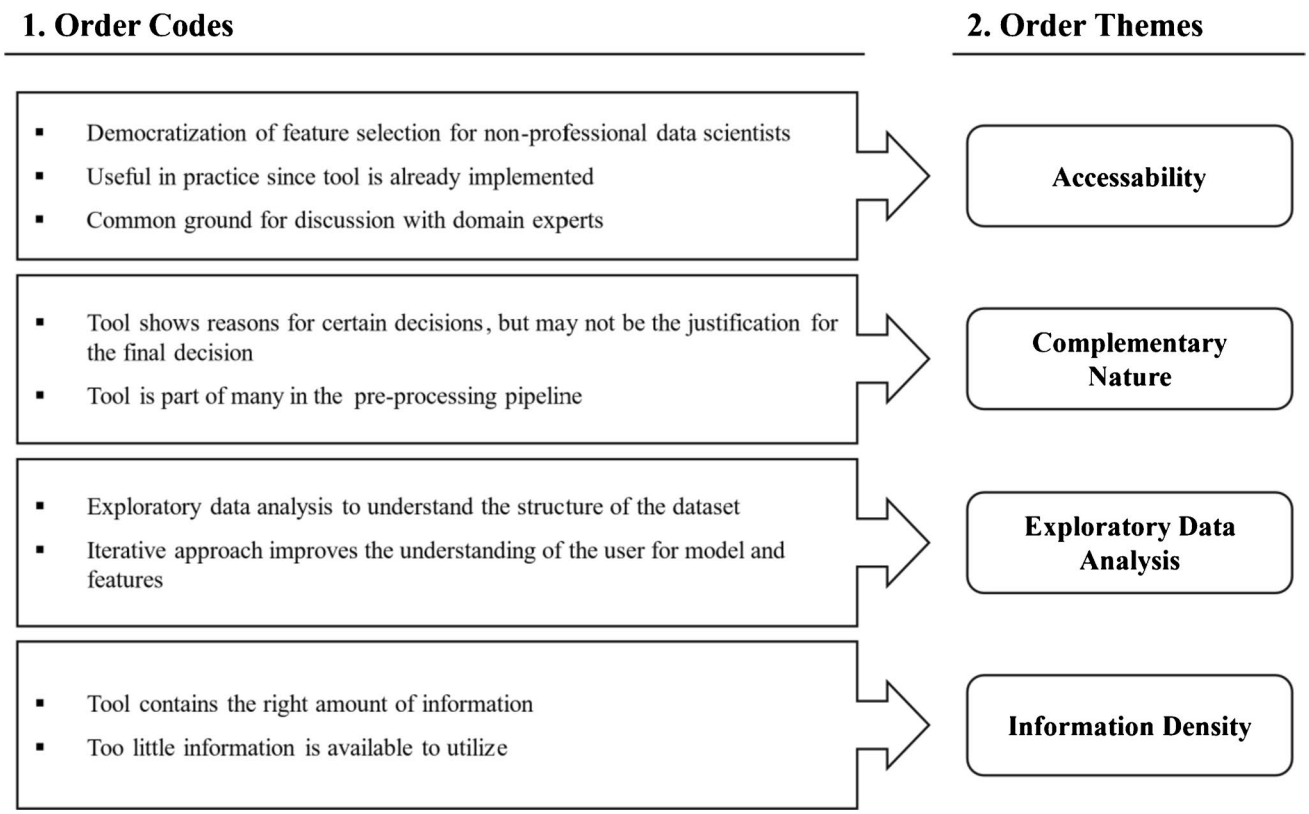

Fig. 5 Data structure of the codes and themes of the focus group discussion following Gioia et al. (2013)

A different facet of accessibility is the ability to lead discussions with domain experts and other stakeholders that do not have a technical background.

> *"Another practical use case is to guide domain experts through [the tool], if you have features where you need external input [...] There are simply problems or tasks that you cannot solve on your own as a data scientist."* (Expert 10)

Since there are many situations in practice where a data scientist is unable to complete the feature selection process just based on technical expertise alone, communication with a domain expert becomes necessary. In these cases, a common ground for discussion has to be created. Expert 10 stated that he could imagine using the tool to talk with domain experts and showing the different local explanations and interactions between the features. Thus, the tool would enable non-technical staff to understand the decision process of feature selection visually. In return, the expert could compare these findings and utilize their domain knowledge to interact with the data scientist.

Many participants further argue that insights generated by this tool facilitate feature selection but are not totally sufficient on their own. In other words, it is well suited as a complementary tool, but feature selection should not rely exclusively on it.

> *"[...] It can be one tool of many. If I have three or four screens, then I have it open on one screen. However, on the remaining screens I have something else open."* (Expert 10)

One participant told us that in order to feel fully confident in the tool, he would need a deeper look into the distributions of the data. Another expert mentioned that the tool successfully shows the impact of features in a visual manner; however, the tool alone would not suffice for the justification of the final decision.

This supportive function is also extended by its explorative nature. Many participants noted that it could be used to explore the data structure and the relationships between features.

> *"I think it is very good for exploratory data analysis. It is also important for the data scientist to understand the structure of your data."* (Expert 6)

However, one of the participants argued that while you are able to explore the features and their interactions, at the end of all iterations, you are unable to see which features were omitted and which were kept. In this case, a report, which summarizes the different iterations of the feature selection process, would be useful. Thus, we propose an additional DP: The artifact should summarize the decisions to support feature selection justification towards stakeholders by the developer **(DP 7)**.

The last theme derived from the discussion refers to the information density of the tool. When designing XAI systems, it is crucial to find a balance between providing the user with extensive information and, on the other hand, avoiding information overload. We interestingly received divergent opinions when asked how the users perceived the information provided by the tool. One of the users perceived the information density to be just right, i.e. the tool showed the right amount of information to him. On the other hand, another user claimed that not enough information was available to him and that more information is always useful.

"I think that was exactly the essential [information]" (Expert 1).

> *"I think the opposite is the case, that there is too little information available to utilize. [...] Every piece of information that you have helps you to understand the underlying model better. That is why I believe that more information is always useful."* (Expert 8)

Since those two participants have different professional backgrounds—expert 1 being a technology consultant and expert 8 being an ML researcher—, we assume that their individual backgrounds influence their need for information. In other words, the professional background seems to moderate the participants' assessment of the information density. Thus, we propose the following DP: The artifact should provide different levels of information according to the individual user's needs **(DP 8)**.

To support our qualitative results further, we quantitatively analyzed the data collected by the Likert scale-based questionnaire. In summary, the quantitative results indicate that our artifact achieves its goal of making feature selection more explainable and justifiable to stakeholders. We present the quantitative results in more detail in Appendix 6.

Discussion

Key findings

The qualitative results of our focus group sessions reveal a variety of benefits of our artifact. Apart from the increased explainability, our tool may make the feature selection task more accessible for intermediate data scientists and thereby contribute to data democratization. According to information systems literature, data democratization might substantially benefit organizations by putting more employees in the position of active contributors to data-driven solutions (Awasthi and George, 2020). Furthermore, our artifact could facilitate the exchange of

information between ML developers and domain experts to enhance the effectiveness of feature selection. Using the SHAP visualizations of our artifact, domain experts can map automatically generated explanations against their domain knowledge and control for potential biases and flaws. Indeed, existing literature suggests that data-driven analyses immensely benefit from the inclusion of domain expertise (Maass et al., 2018; Teso and Kersting, 2019). By interpreting data-driven analyses, domain experts might even extend their domain theory with novel insights (Murdoch et al., 2019; Maass et al., 2018) with the potential to foster human learning (see e.g., Bauer et al., 2021b). Moreover, when discussing our tool, the experts seemed to value the in-depth information on feature impacts and pointed out its suitability for exploratory data analysis. Therefore, developers may use our tool not only to select features, but also to gain a general understanding of their datasets. Overall, the aforementioned benefits and use cases diverge significantly from the purpose that we originally defined for our artifact, namely explainable and justifiable feature selection. This underscores our argument that explainable feature selection may benefit organizations in various ways, beyond meeting legal and ethical requirements.

The quantitative results indicate that our artifact enables developers to better explain and justify feature selection to stakeholders. The participants also regard our artifact as both useful and usable in practice. These results imply that our artifact indeed achieves its general purpose, i.e. making feature selection explainable and justifiable. However, there is still potential for improvement, especially in terms of users' confidence, emotional trust, and satisfaction with the artifact.

Contributions to practice

Our work entails two main contributions to practice. First, by instantiating and empirically evaluating this design framework, we demonstrate the effectiveness and applicability of our artifact. Practitioners that aim to improve the interpretability of their ML-based systems may adopt our running software artifact to make their feature selection processes more explainable and justifiable towards their stakeholders. Thereby, our work helps to promote law-abiding and ethical ML-based systems in practice. Second, our qualitative results show that practitioners may leverage explainable feature selection to foster data democratization and human learning. In organizations, for example, domain experts with non-technical backgrounds could use our artifact to both participate directly in creating ML-based systems and enhance their understanding of related data.

Contributions to theory

Our contribution to literature is twofold. First, we provide prescriptive design knowledge by proposing a design framework consisting of meta-requirements and design principles for the development of explainable feature selection (Fig. 2). This design knowledge complements related DSR works which mainly focus on the explainability of already completed AI systems by addressing the explainability and justifiability of ML pre-processing. This type of contribution is in line with Baskerville et al. (2018), who argue that novel design knowledge constitutes the primary theoretical contribution of DSR. Design knowledge may take the form of design principles, which Baskerville et al. (2018) denote as "nascent design theory". Second, our work contributes to the broader literature in the XAI field. The current literature indicates several application domains for XAI methods with justification towards regulators and other stakeholders as one major application (Adadi and Berrada, 2018). However, existing research limits the scope of justification to the final predictions of a ML model (Chakrobartty and El-Gayar, 2021; Fernandez et al., 2022; Cirqueira et al., 2021; Zhang et al., 2020). This narrow focus leaves all prior stages within the ML pipeline—such as data collection, feature selection and model training—opaque for regulators and other stakeholders. Our work addresses this gap by showing how XAI can justify feature selection, a crucial stage within ML pre-processing.

Limitations and future research directions

Our work does not come without limitations. In our technical analysis, we compared the SHAP-based feature selection approach with two benchmark methods. However, the list of existing feature selection methods is considerably longer. Moreover, we can draw conclusions only regarding the six test datasets that we used. Future work could extend this by comparing SHAP to additional feature selection methods and datasets. Regarding the evaluation of our artifact, we conducted our focus group sessions with twelve participants. However, in order to derive statistically valid quantitative results, a considerably larger number of participants is required. Furthermore, besides the evaluation of our artifact instantiation, we did not conduct a comprehensive evaluation of our design principles as such. Future research could extend and refine our work by performing a structured evaluation of design principles based on the framework of Iivari et al. (2021). This framework proposes a strategy for evaluating the reusability of design principles in practice. Accordingly, this presents an excellent opportunity for fellow researchers to make a distinctive contribution. Furthermore, by initiating the second design science cycle and deriving DPs from the focus groups, we offer the opportunity for future researchers

to build upon our prototype. Future projects aiming to design explainable feature selection could seamlessly connect to our work and realize the improvement opportunities detailed above. Finally, we invite fellow researchers to extend the idea of explainable feature selection to other stages within ML preprocessing, such as data collection and ML model training. Novel methods addressing the explainability of these stages as well could contribute to holistically justifiable ML systems that are prepared to meet current and upcoming explainability requirements.

Conclusion

In this work, we addressed the issue that legal and ethical requirements demand organizations to ensure explainability throughout the development (including pre-processing) and usage of AI systems. However, extant methods for feature selection, a major part of the pre-processing pipeline, lack explainability and make persuasive justification of feature selection difficult. To solve this problem, we developed a design framework of a XAI-based feature selection method and instantiated it as a software artifact. For the evaluation of our artifact, we conducted focus group research and gathered both quantitative and qualitative feedback on established evaluation criteria.

Finally, we reflect on our identified research gap. In the related works section, we identified a lack of design knowledge on explainable ML pre-processing; with our design framework, we take an important step in exploring this promising research field and invite the information systems community to build upon our work. We further argued that existing tools for feature selection solely focus on optimizing prediction performance and mostly neglect the explainability of feature selection. Our work demonstrates that feature selection based on the XAI technique SHAP may achieve both explainability and high prediction performance. Accordingly, organizations could build upon our artifact to both preserve the benefits of ML and meet legal and ethical requirements.

Appendix

Appendix 1: Test set discrimination

Generally, global SHAP importance values are generated by averaging the local SHAP values across the whole test set. As an alternative to averaging across the whole test set, we argue that the aggregation can also be conducted on a selected subset of test observations; we call this concept test set discrimination. This approach is motivated by the work of Blum and Langley (1997). The authors explain that—similar to features—observations often differ in terms of their importance for the training of ML models. Analogous to feature selection, they introduce the concept of example selection as a technique to distinguish between informative and less informative observations.

Test set discrimination works as follows: After obtaining predictions for the test set, a confusion matrix of the predictions is created. Instead of aggregating the local SHAP values across all test observations, one can select specific

Table 2 Datasets for the stability and prediction performance analyses

Dataset	# of observations	# of attributes	# of attributes after cleansing	Target balancing
Diabetes[1]	101766	50	45	1: 0.54 | 0: 0.46
Spam base[2]	4601	58	58	1: 0.39 | 0: 0.61
German credit[3]	1000	21	21	1: 0.30 | 0: 0.70
Credit card fraud detection[4]	307511	122	77	1: 0.08 | 0: 0.92
Breast cancer Wisconsin (diagnostic)[5]	569	32	32	1: 0.63 | 0: 0.37
Taiwanese bankruptcy prediction[6]	6819	96	96	1: 0.03 | 0: 0.97

[1]Url: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008 [Accessed: 19-November-2021]

[2]Url: https://archive.ics.uci.edu/ml/datasets/spambase [Accessed: 19-November-2021]

[3]Url: https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 [Accessed: 19-November-2021]

[4]Url: https://www.kaggle.com/mishra5001/credit-card?select=applicationdata.csv [Accessed: 19-November-2021]

[5]Url: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 [Accessed: 24-November-2021]

[6]Url: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction [Accessed: 24-November-2021]

observations as aggregation basis – e.g. the correctly classified or the true positive instances.

Appendix 2: Datasets

Table 2.

Appendix 3: Stability analysis of feature selection

Kalousis et al. (2005) present a framework for assessing the stability of feature selection algorithms. We base our stability analysis on their work. The stability score measures the stability of feature rankings to changes in the training instances from the same dataset. First, the dataset is split into k equally sized folds. Second, for each fold, the feature ranking is generated based on the importance scores of the respective feature selection technique (e.g., the average absolute SHAP value); this leads to k feature rankings $r=[r_1, \ldots, r_m]$, where *m* denotes the number of available features. Third, the similarity of these rankings is determined. For that purpose, the Spearman's rank correlation coefficient is applied:

$$\rho\left(r, r^{'}\right) = 1 - \frac{6\sum\left(r_j - r_j^{'}\right)^2}{m\left(m^2 - 1\right)}$$

with r_j and r'_j being the rank of feature X_j in two different iterations. Using the Spearman's rank correlation, the similarity of each possible ranking pair is determined; the average similarity across all ranking pairs in turn leads to the final stability score with a value range from 0 to 1. We calculated the stability score on each dataset for our SHAP-based feature selection technique as well as for the two benchmark methods. In order to investigate the effects of the test set discrimination, we further differentiated between the SHAP-based approach based on (i) the whole test set, (ii) the correctly classified observations (hereafter referred to as SHAP-true based feature selection) and (iii) the correctly classified minority-class class observations (hereafter referred to as SHAP-true positive based FS). However, for each dataset we conducted only one version of test set discrimination; which version is chosen depends on the dataset: for highly unbalanced datasets, we chose the SHAP-true positives-based feature selection. For the remaining datasets, we chose the SHAP-true based feature selection. The rationale for this procedure is that for highly unbalanced datasets, the correct classification of minority class-observations is considerably more important than the correct classification of majority class-observations. Table 3 presents the results of the stability analysis.

The feature rankings of the correlation based FS approach are by definition the most stable ones with perfect stability across all datasets. Nonetheless, the SHAP-based approach—with and without test set discrimination—yields highly stable rankings as well. XGBoost Built-In performs the worst on the stability dimension: its feature rankings are relatively unstable on the small datasets but comparable to the SHAP-based approach on the larger datasets.

Appendix 4: Prediction performance analysis of feature selection

In order to assess the impact on prediction performance of each feature selection method, we performed an iterative backward-elimination process and evaluated an ML model after each feature removal until all features are eliminated. Figure 6 describes our performance comparison algorithm for the SHAP-based approach.

We start with a train-validation-test split of the dataset; the training and validation sets are used for ML model training whereas the test set is held out for the performance evaluation. Following that, Synthetic Minority Over-Sampling Technique (SMOTE) is conducted in order to achieve an equal distribution of the target variable. Based on the oversampled training set, a XGBoost model with a learning rate of 0.3 and tree depth of 6 is trained. After the validation set

Table 3 Stability analysis results

Stability score					
Dataset	SHAP	SHAP with test set discrimination		XGB-built-in	Correlation
		SHAP-True	SHAP-True Positive		
Diabetes	0.98	0.97	/	0.96	1
Spam base	0.87	0.85	/	0.76	1
German Credit	0.83	0.85	/	0.46	1
Credit Card Fraud Detection	0.97	/	0.95	0.94	1
Breast Cancer Wisconsin (Diagnostic)	0.83	0.83	/	0.51	1
Taiwanese Bankruptcy Prediction	0.85	/	0.97	0.57	1

Fig. 6 Process of the prediction performance analysis

Algorithm 1: Iterative SHAP based Feature Selection

Input: Dataset, test set size, backward elimination stepsize
Output: Mean and StDev of performance scores per iteration

```
Split dataset into internal_set and test_set
m = number of features of internal_set
Split internal_set into training_set and validation_set
Perform SMOTE oversampling
Train XGBoost classifier on training_set
SHAP_values_initial = Ordered global SHAP values on validation_set
for i ← 1 to m do
    Initiate performance score lists
    Seeds = {1 to 30}
    for i ← 1 to Seeds do
        Split internal_set into training_set and validation_set
        Perform SMOTE oversampling
        Train XGBoost classifier on training_set
        Calculate performance metrics on holdout_set and assign to performance
         score lists
    end
    Calculate mean and stdev of performance score lists
    if m > 25 then
        remove_features = SHAP_values_initial[-stepsize]
    end
    else
        SHAP_values_iterative = Ordered Global SHAP values on
         validation_set
        remove_features = SHAP_values_iterative[-stepsize]
    end
    internal_set = internal_set[-remove_features]
    holdout_set = holdout_set[-remove_features]
end
return(Mean and StDev of performance scores per iteration)
```

prediction using the XGBoost model, the global SHAP value of each feature is computed—either with or without test set discrimination. The test set discrimination is conducted in the same way as in the stability analysis: For highly unbalanced datasets, we selected only the correctly classified minority class-observations (SHAP-true positives based feature selection). For the remaining datasets, we selected the correctly classified observations from both classes (SHAP-true based feature selection). The sorted global SHAP values act as feature ranking $r = (r_1, \ldots, r_m)$. According to this ranking, we performed stepwise backward elimination; at each step, the lowest ranked feature was removed and thirty XGBoost models were trained on the remaining dataset. With each of the thirty XGBoost training procedures, we varied the train-validation split and thus added variability into the models. Next, we estimated the area under the receiver operating curve (ROC AUC) and area under the precision-recall curve (AUPRC) of all thirty XGBoost models on the test set; the average and standard deviation of these metrics are chosen as main prediction performance scores. The evaluation of the benchmark feature selection methods was conducted analogously. The only adaptation pertains to the feature rankings r since each method produces the rankings differently.

The graphs presented in Fig. 7 show the average ROC AUC and AUPRC per backward-elimination step for each feature selection technique. On the x-axis, the sizes of the feature sets which are used for model training are listed in ascending order. On the y-axis, the average ROC AUC and AUPRC for the respective feature sets are shown.

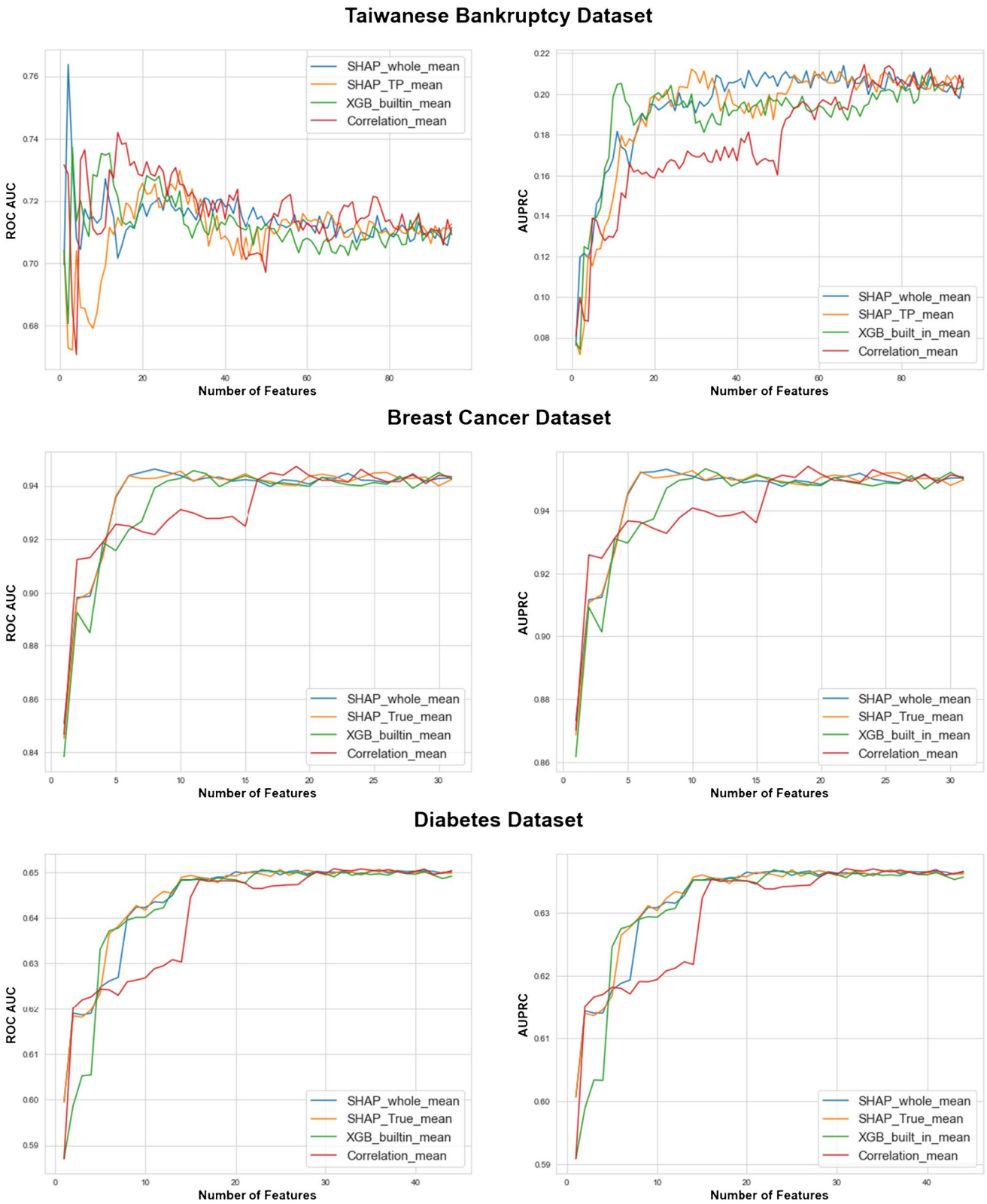

Fig. 7 Prediction performance analysis plots—ROC AUC and AUPRC

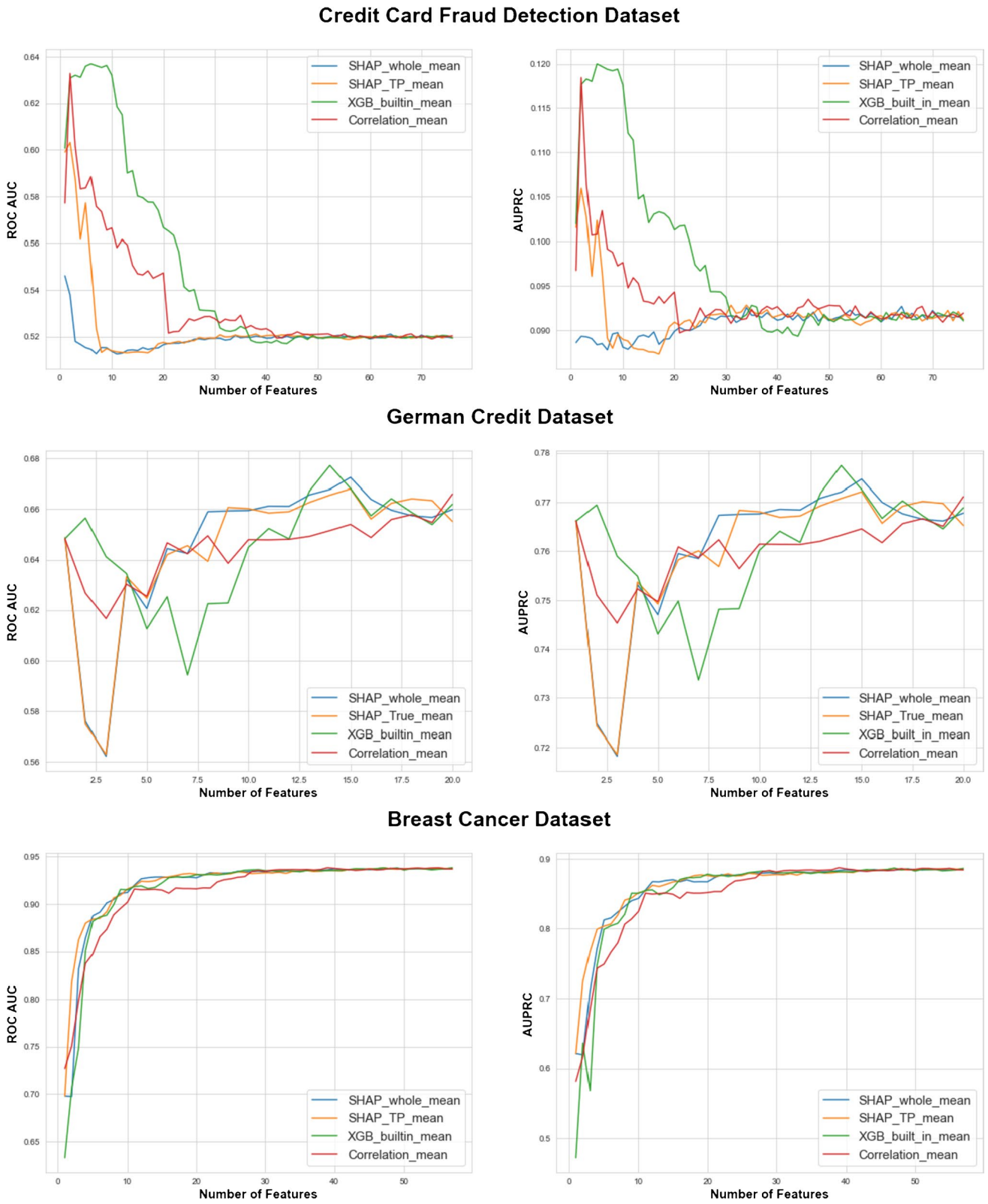

Fig. 7 (continued)

Further, we performed a statistical analysis of the ROC AUC and AUPRC differences between the feature selection methods using the Wilcoxon signed-rank test. The Wilcoxon signed-rank test is a widely used non-parametric procedure for the comparison of two paired samples. It tests the hypothesis that the median of the differences between two paired samples equals zero (Benavoli et al., 2014).

Table 4 shows the results of the Wilcoxon signed-rank tests on the different datasets.

In summary, the prediction performance results show considerable differences across the applied datasets. However, we do neither find unambiguous evidence for prediction performance superiority nor inferiority of SHAP-based FS. In other words, our results imply that SHAP is at least not worse than established methods for the feature selection task.

Appendix 5

Table 5 presents the constructs of our questionnaire, their measures, i.e. the concrete questions listed in the questionnaire, and the source of the respective constructs. Our primary evaluation construct is the effectivess of our artifact, which is defined as "[…] the degree to which the artifact meets its higher level purpose or goal and achieves its desired benefit in practice" (Venable et al., 2012, p. 426). Thereby, we follow the approach of Meth et al. (2015), who describe their artifact's effectiveness and efficiency as central criteria. We further extended our primary construct with six supplementary constructs which we derived from popular DSR papers. The first supplementary construct is the perceived usefulness (Davis, 1989), which is regarded as a crucial criterion when evaluating the contribution to design knowledge bases (Meske and Bunde, 2022). Constructs which are closely linked to user

Table 4 Performance metrics including p-values of Wilcoxon rank-sum tests

H0: Median of differences = 0		AUC		AUPRC	
H1: Median of differences ! = 0		Median difference		Median difference	
Diabetes	SHAP—SHAP True	-1.1E + 11		-9.3E-07	
	SHAP—XGB Built-In	**4.62E-04**	**	**3.41E-04**	**
	SHAP True—XGB Built-In	**6.30E-04**	**	**4.34E-04**	**
	SHAP—Correlation	**6.96E-04**	***	**4.44E-04**	**
	SHAP True—Correlation	**5.67E-04**	***	**3.22E-04**	**
Spam Base	SHAP—SHAP True	-7.83E-05		-1.74E-04	
	SHAP—XGB Built-In	2.05E-04		2.77E-04	
	SHAP True—XGB Built-In	4.58E-04		8.59E-04	
	SHAP—Correlation	**5.57E-04**	**	**1.04E-03**	**
	SHAP True—Correlation	**8.62E-04**	***	**1.87E-03**	***
German Credit	SHAP—SHAP True	5.87E-04		2.39E-04	
	SHAP—XGB Built-In	3.62E-03		1.87E-03	
	SHAP True—XGB Built-In	2.89E-03		1.48E-03	
	SHAP—Correlation	2.84E-03		1.54E-03	
	SHAP True—Correlation	2.84E-03		3.51E-03	
Fraud Detection	SHAP—SHAP True Positives	**-2.80E-04**	*	**-1.93E-04**	***
	SHAP—XGB Built-In	**-7.40E-04**	***	**-3.57E-04**	***
	SHAP True Positives—XGB Built-In	**-1.05E-03**	***	**-4.54E-04**	***
	SHAP—Correlation	**-3.64E-03**	***	**-7.74E-04**	***
	SHAP True Positives—Correlation	**-2.71E-03**	***	**-6.52E-04**	***
Breast Cancer	SHAP—SHAP True	-1.28E-04		-1.76E-04	
	SHAP—XGB Built-In	**1.05E-03**	*	**5.33E-04**	*
	SHAP True—XGB Built-In	**2.01E-03**	**	**1.68E-03**	**
	SHAP—Correlation	8.78E-04		7.25E-04	
	SHAP True—Correlation	1.39E-03		9.01E-04	
Bankruptcy	SHAP—SHAP True Positives	-4.75E-05		5.91E-04	
	SHAP—XGB Built-In	**2.30E-03**	*	**6.00E-03**	***
	SHAP True Positives—XGB Built-In	1.25E-03		**4.01E-03**	**
	SHAP—Correlation	**-3.35E-03**	**	**1.97E-02**	***
	SHAP True Positives—Correlation	**-4.53E-03**	***	**1.37E-02**	***

Table 5 Constructs and measures

Construct	Measure	Source
Effectiveness	E1: Through the program I better understand which features are relevant for the ML model	Venable et al., 2012; Meth et al., 2015
	E2: The program lets me better understand dependencies between features	
	E3: Through the program I gain a better general understanding of model preprocessing	
	E4: Through the program, I am able to justify feature selection to customers, regulators, or other stakeholders	
	E5: I consider the presented feature selection method applicable in practice	
Perceived Usefulness	PU1: The program gives me better control over the feature selection process	Davis, 1989; Meske and Bunde, 2022
	PU2: Overall, I find the program useful in practice	
Understandability	U1: The analyses are presented in a clear and understandable manner	Adipat et al., 2011
Ease of Use	EoU1: I consider the operation of the program easy to learn	Adipat et al., 2011
Confidence	C1: I am confident that I have made a good selection of features	Chen and Koufaris, 2015
	C2: I feel safe relying on this program for my feature selection	
Emotional Trust	ET1: I feel comfortable relying on this program for my feature selection	Komiak and Benbasat 2006
Satisfaction	S1: I am satisfied with the decision making process during this Feature Selection	Chen and Koufaris, 2015
	S2: I am satisfied with the choices I made during this Feature Selection	

acceptance of information systems are understandability, ease of use, and trustworthiness (Meske and Bunde, 2022); to be more precise, we adapted trustworthiness to emotional trust (Komiak and Benbasat, 2006). Finally, we added the constructs confidence and satisfaction from Chen and Koufaris (2015) to our questionnaire to get a more holistic idea of the user perception towards our artifact.

Appendix 6

Figure 8 concisely summarizes our quantitative results on the construct level and Table 6 presents these results granularly on the measured level. One crucial construct that we used is effectiveness. Effectiveness is defined as "the degree to which the artifact meets its higher level purpose or goal and achieves its desired benefit in practice" (Venable et al., 2012, p. 426). In our case, the terms'higher level purpose or goal' refer to high explainability of the feature selection process and, consequently, the facilitation of justifying feature selection related decisions to stakeholders. On average, effectiveness has a rating of 3.93 on a scale from 1 to 5. More precisely, measure E1 has the highest ranking with a mean of 4.42. This shows that the participants unanimously gained a better understanding of feature relevances. The participants also agreed that the tool helps justify feature selection to stakeholders (E4 = 4.01). Further, the participants gave particularly high ratings for the constructs 'perceived usefulness' (4.21) and 'ease of use' (4.33).

The understandability was rated slightly worse with an average of 3.75. Notably, not all participants perceived the

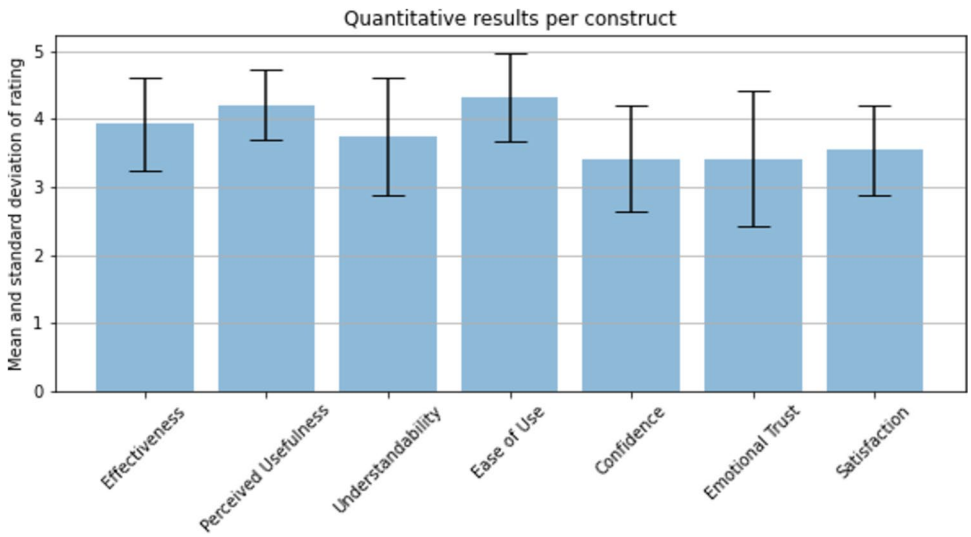

Fig. 8 Quantitative results of the focus group questionnaire on construct level

Table 6 Quantitative results of the focus group questionnaire on the measured level

Construct	Overall mean	Measure	Mean	Standard deviation
Effectiveness	3.93	E1	4.42	0.67
		E2	3.83	0.58
		E3	3.25	1.14
		E4	4.01	0.51
		E5	4.01	0.51
Perceived Usefulness	4.21	PU1	4.25	0.45
		PU2	4.17	0.58
Understandability	3.75	U1	3.75	0.87
Ease of Use	4.33	EoU1	4.33	0.65
Confidence	3.42	C1	3.50	0.67
		C2	3.33	0.89
Emotional Trust	3.42	ET1	3.42	0.10
Satisfaction	3.54	S1	3.50	0.67
		S2	3.58	0.67

information presented in the tool as perfectly understandable. XAI systems are prone to overload users with details, which should be avoided (Kulesza et al., 2015). In contrast, the ease of use has a high ranking with 4.33. The lowest ratings are given to confidence, emotional trust and satisfaction. This shows that users neither feel fully comfortable with relying solely on that artifact when performing feature selection nor are they fully satisfied with their achieved results.

Overall, the results indicate that the tool achieves its desired goal of making features selection explainable. However, there is still room for improvement.

Funding Open Access funding enabled and organized by Projekt DEAL. This research has received funding from the Volkswagen Foundation as part of the initiative Artificial Intelligence and its Impact on Tomorrow's World.

References

Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems-a bibliographic review and open research issues. *Electronic Markets, 31*(3), 643–670. https://doi.org/10.1007/s12525-021-00459-2

Abedin, B. (2021). Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective. *Internet Research*. https://doi.org/10.1108/INTR-05-2020-0300

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adipat, B., Zhang, D., & Zhou, L. (2011). The effects of tree-view based presentation adaptation on mobile web browsing. *MIS Quarterly,* 99–121. https://doi.org/10.2307/23043491

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Lopez, S.-G., Molina, D., Benjaminsh, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *Journal of the Association for Information Systems, 22*(2), 8. https://doi.org/10.17705/1jais.00664

Awasthi, P., & George, J. (2020). A case for data democratization. *Proceedings of theAmericas Conference on Information Systems (AMCIS), 23.*

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research, 18*(1), 629–681.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems, 19*(5), 3. https://doi.org/10.17705/1jais.00495

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021a). Expl(AI)n it to me–explainable AI and information systems research. *Business & Information Systems Engineering, 63*(2), 79–82. https://doi.org/10.1007/s12599-021-00683-2

Bauer, K., von Zahn, M., & Hinz, O. (2021b). Expl(Ai)Ned: The impact of explainable artificial intelligence on cognitive processes. *SAFE Working Paper No. 315*. https://ssrn.com/abstract=3872711

Baum, T., Herbold, S., & Schneider, K. (2020). GIMO: A multi-objective anytime rule mining system to ease iterative feedback from domain experts. *Expert Systems with Applications: X, 8*, 100040. https://doi.org/10.1016/j.eswax.2020.100040

Belanger, F. (2012). Theorizing in information systems research using focus groups. *Australasian Journal of Information Systems*, 17(2). https://doi.org/10.3127/ajis.v17i2.695

Benavoli, A., Corani, G., Mangili, F., Zaffalon, M., & Ruggeri, F. (2014). A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. *International conference on machine learning* (pp. 1026–1034). PMLR.

Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences, 37*(1), 63. https://doi.org/10.1017/S0140525X13000289

Bessa, M. A., Bostanabad, R., Liu, Z., Hu, A., Apley, D. W., Brinson, C., Chen, W., & Liu, W. K. (2017). A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality. *Computer Methods in Applied Mechanics and Engineering*, *320*, 633–667. https://doi.org/10.1016/j.cma.2017.03.037

Bhandari, S., Kukreja, A. K., Lazar, A., Sim, A., & Wu, K. (2020). Feature selection improves tree-based classification for wireless intrusion detection. *Proceedings of the 3rd International Workshop on Systems and Network Telemetry and Analytics* (pp. 19–26). https://doi.org/10.1145/3391812.3396274

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1–2), 245–271. https://doi.org/10.1016/S0004-3702(97)00063-5

Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking explainable machines: The GDPR's' right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Tech. LJ, 34*, 143. https://doi.org/10.15779/Z38M32N986

Chakrobartty, S., & El-Gayar, O. (2021). Explainable artificial intelligence in the medical domain: A systematic review. *Proceedings of theAmericas Conference on Information Systems (AMCIS)*.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, C. W., & Koufaris, M. (2015). The impact of decision support system features on user overconfidence and risky behavior. *European Journal of Information Systems, 24*(6), 607–623. https://doi.org/10.1057/ejis.2014.30

Cirqueira, D., Helfert, M., & Bezbradica, M. (2021). Towards design principles for user-centric explainable AI in fraud detection. *International Conference on Human-Computer Interaction* (pp. 21–40). Springer, Cham. https://doi.org/10.1007/978-3-030-77772-2_2

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340. https://doi.org/10.2307/249008

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM, 63*(1), 68–77. https://doi.org/10.1145/3359786

Dunn, J., Mingardi, L., & Zhuo, Y. D. (2021). Comparing interpretability and explainability for feature selection. *arXiv preprint* arXiv:2105.05328. *https://doi.org/10.48550/arXiv.2105.05328*

Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics, 61*, 101224. https://doi.org/10.1016/j.ecoinf.2021.101224

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics, 26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Fernandez, C., Provost, F., & Han, X. (2022). Explaining data-driven decisions made by AI systems: The counterfactual approach. *MIS Quarterly, 46*(3), 1635–1660. https://doi.org/10.25300/MISQ/2022/16749

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Fostering human agency: A process for the design of user-centric XAI systems. *ICIS 2020 Proceedings*, 12.

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access, 9*, 144352–144360. https://doi.org/10.1109/ACCESS.2021.3119110

General Data Protection Regulation (GDPR). (2018). *General data protection regulation (GDPR) – final text neatly arranged*. [online]. Available at: https://gdpr-info.eu. Accessed Feb 2022.

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. *Proceedings of the InternationalConference on Information Systems (ICIS)*.

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, 611–642. https://doi.org/10.2307/25148742

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems*, *21*(6), 2. https://doi.org/10.17705/1jais.00649

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182. https://doi.org/10.1162/153244303322753616

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge W., & Williams, M. A. (2018). Do you trust me, blindly? Factors influencing trust towards a robot recommender system. *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 7–14). https://doi.org/10.1109/ROMAN.2018.8525581

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75–105. https://doi.org/10.2307/25148625

H.R.6580 – *Algorithmic Accountability Act of 2022*. https://doi.org/10.2139/ssrn.4135237

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Iivari, J., Rotvit Perlt Hansen, M., & Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems*, *30*(3), 286–303. https://doi.org/10.1080/0960085X.2020.1793697

Jia, K., & Zhang, N. (2022). Categorization and eccentricity of AI risks: A comparative study of the global AI guidelines. *Electronic Markets*, *32*(1), 1–13. https://doi.org/10.1007/s12525-021-00480-5

Kalousis, A., Prados, J., & Hilario, M. (2005). Stability of feature selection algorithms. *Fifth IEEE International Conference on Data Mining (ICDM'05)* (pp. 8). IEEE. https://doi.org/10.1109/ICDM.2005.135

Kellner, D., Lowin, M., von Zahn, M., & Chen, J. (2021). Towards designing a user-centric decision support system for predictive maintenance in SMEs. *INFORMATIK 2021. Gesellschaft für Informatik,* 1255–1260. https://doi.org/10.18420/informatik2021-104

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kim, T. W., & Routledge, B. R. (2022). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly, 32*(1), 75–102. https://doi.org/10.1017/beq.2021.3

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 941–960. https://doi.org/10.2307/25148760

Koulu, R. (2021). Crafting digital transparency: Implementing legal values into algorithmic design. *Critical Analysis of Law, 8*(1), 81–100.

Krause, J., Perer, A., & Bertini, E. (2014). INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1614–1623. https://doi.org/10.1109/TVCG.2014.2346482

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5686–5697). https://doi.org/10.1145/2858036.2858529

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems, 17*(5), 489–504. https://doi.org/10.1057/ejis.2008.40

Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2020). Machine learning in artificial intelligence: Towards a common understanding. *arXiv preprint* arXiv:2004.04686. *https://doi.org/10.48550/arXiv.2004.04686*

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126–137). https://doi.org/10.1145/2678025.2701399

Li, J., Yan, X. S., Chaudhary, D., Avula, V., Mudiganti, S., Husby, H., Shahjouei, S., Afshar, A., Stewart, W. F., Yeasin, M., Zand, R., & Abedi, V. (2021). Imputation of missing values for electronic health record laboratory data. *NPJ Digital Medicine*, *4*(1), 1–14. https://doi.org/10.1038/s41746-021-00518-0

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint* arXiv:1802.03888. *https://doi.org/10.48550/arXiv.1802.03888*

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1. https://doi.org/10.17705/1jais.00526

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Marcílio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing shap values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)* (pp. 340–347). IEEE. https://doi.org/10.1109/SIBGRAPI51738.2020.00053

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D. Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Meske, C., & Bunde, E. (2022). Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers*, 1-31. https://doi.org/10.1007/s10796-021-10234-5

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association for Information Systems*, *16*(9), 2. https://doi.org/10.17705/1jais.00408

Mlambo, N., Cheruiyot, W. K., & Kimwele, M. W. (2016). A survey and comparative study of filter and wrapper feature selection techniques. *International Journal of Engineering and Science (IJES), 5*(8), 57–67.

Müller, O., Fay, M., & Vom Brocke, J. (2018). The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *Journal of Management Information Systems, 35*(2), 488–509. https://doi.org/10.1080/07421222.2018.1451955

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E., & Loncarski, J. (2018). Machine learning approach for predictive maintenance in industry 4.0. *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)* (pp. 1–6). IEEE. https://doi.org/10.1109/MESA.2018.8449150

Pfeuffer, N. (2021). Explainability in interactive machine learning: Novel avenues for information systems research. *Proceedings of the Pacific Asia Conference on Information Systems (PACIS),* 231. https://aisel.aisnet.org/pacis2021/231

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four principles of explainable artificial intelligence. *Gaithersburg, Maryland*.

Plale, B. (2019). Transparency by design in eScience research. *2019 15th International Conference on eScience (eScience)* (pp. 428–431). IEEE. https://doi.org/10.1109/eScience.2019.00055

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design-science research- A holistic view. *Pacific Asia Conference on Information Systems (PACIS), 23*, 1–16.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research,* 3, 1371–1382. https://dl.acm.org/doi/https://doi.org/10.5555/944919.944978

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Schemmer, M., Hemmer, P., Kühl, N., & Schäfer, S. (2022). Designing resilient AI-based robo-advisors: A prototype for real estate appraisal. *17th International Conference on Design Science Research in Information Systems and Technology*, 1st–3rd June 2022, St. Petersburg, FL.

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. A. (2019). Towards a rigorous evaluation of xai methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 4197–4201). IEEE. https://doi.org/10.1109/ICCVW.2019.00516

Senoner, J., Netland, T., & Feuerriegel, S. (2021). Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science*. https://doi.org/10.1287/mnsc.2021.4190

Seo, J., & Shneiderman, B. (2005). A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization, 4*(2), 96–113. https://doi.org/10.1057/palgrave.ivs.9500091

Shapley, S. (1953). A value for n-person games. Contributions to the Theory of Games II. *Annals of Mathematical Studies, 28*. Princeton University Press.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 1–21. https://doi.org/10.1186/1471-2105-8-25

Teso, S., & Kersting, K. (2019). Explanatory interactive machine learning. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 239–245). https://doi.org/10.1145/3306618.3314293

Toreini, P., Langner, M., Maedche, A., Morana, S., & Vogel, T. (2022). Designing attentive information dashboards. *Journal of the Association for Information Systems*, *2021*. https://doi.org/10.17705/1jais.00732

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research.

International conference on design science research in information systems (pp. 423–438). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_31

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. *International work-conference on artificial neural networks* (pp. 758–770). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11494669_93

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3290605.3300831

Xiaomao, X., Xudong, Z., & Yuanfang, W. (2019). A comparison of feature selection methodology for solving classification problems in finance. *Journal of Physics: Conference Series* (vol. 1284, No. 1, p. 012026). IOP Publishing. https://doi.org/10.1088/1742-6596/1284/1/012026

Zhang, L., Mistry, K., Lim, C. P., & Neoh, S. C. (2018). Feature selection using firefly optimization for classification and regression models. *Decision Support Systems, 106*, 64–85. https://doi.org/10.1016/j.dss.2017.12.001

Zhang, X., Du, Q., & Zhang, Z. (2020). An explainable machine learning framework for fake financial news detection. *International Conference on Information Systems* (*ICIS*).

Zhao, J., Karimzadeh, M., Masjedi, A., Wang, T., Zhang, X., Crawford, M. M., & Ebert, D. S. (2019). Featureexplorer: Interactive feature selection and exploration of regression models for hyperspectral images. *2019 IEEE Visualization Conference (VIS)* (pp. 161–165). IEEE. https://doi.org/10.1109/VISUAL.2019.8933619

Zieglmeier, V., & Pretschner, A. (2021). Trustworthy transparency by design. *arXiv preprint.* https://doi.org/10.48550/arXiv.2103.10769