

Explainable Artificial Intelligence Approaches: A Survey

Sheikh Rabiul Islam, University of Hartford, William Eberle, Tennessee Tech University, Sheikh Khaled Ghafoor, Tennessee Tech University, Mohiuddin Ahmed, Edith Cowan University

Abstract—The lack of explainability of a decision from an Artificial Intelligence (AI) based “black box” system/model, despite its superiority in many real-world applications, is a key stumbling block for adopting AI in many high stakes applications of different domain or industry. While many popular Explainable Artificial Intelligence (XAI) methods or approaches are available to facilitate a human-friendly explanation of the decision, each has its own merits and demerits, with a plethora of open challenges. We demonstrate popular XAI methods with a mutual case study/task (i.e., credit default prediction), analyze for competitive advantages from multiple perspectives (e.g., local, global), provide meaningful insight on quantifying explainability, and recommend paths towards responsible or human-centered AI using XAI as a medium. Practitioners can use this work as a catalog to understand, compare, and correlate competitive advantages of popular XAI methods. In addition, this survey elicits future research directions towards responsible or human-centric AI systems, which is crucial to adopt AI in high stakes applications.

Index Terms—Explainable Artificial Intelligence, Explainability Quantification, Human-centered Artificial Intelligence, Interpretability.

1 INTRODUCTION

ARTIFICIAL Intelligence (AI) has become an integral part of many real-world applications. Factors fueling the proliferation of AI-based algorithmic decision making in many disciplines include: (1) the demand for processing a variety of voluminous data, (2) the availability of powerful computing resources (e.g., GPU computing, cloud computing), and (3) powerful and new algorithms. However, most of the successful AI-based models are “black box” in nature, making it a challenge to understand how the model or algorithm works and generates decisions. In addition, the decisions from AI systems affect human interests, rights, and lives; consequently, the decision is crucial for high stakes applications such as credit approval in finance, automated machines in defense, intrusion detection in cybersecurity, etc. Regulators are introducing new laws such as European Union’s General Data Protection Regulation (GDPR) ¹ [1] aka “right to explanation” [2], US government’s “Algorithmic Accountability Act of 2019” ² [3], or U.S. Department of Defense’s Ethical Principles for Artificial Intelligence ³ [4]) to tackle primarily fairness, accountability, and transparency-related risks with automated decision making systems.

XAI is a re-emerging research trend, as the need to advocate these principles/laws, and promote the explainable decision-making system and research, continues to increase. Explanation systems were first introduced in the early ’80s to explain the decisions of expert systems. Later, the focus of the explanation systems shifted towards human-computer systems (e.g., intelligent tutoring systems) to provide better cognitive support to users. The primary reason for the renewed interest in XAI research has stemmed from recent advancements in AI and ML, and their application to a

wide range of areas, as well as prevailing concerns over the unethical use, lack of transparency, and undesired biases in the models. Many real-world applications in the Industrial Control System (ICS) greatly increase the efficiency of industrial production from the automated equipment and production processes [5]. However, in this setting, the use of ‘black box’ is still not in a favorable position due to the lack of explainability and transparency of the model and decisions.

According to [6] and [7], XAI encompasses Machine Learning (ML) or AI systems/tools for demystifying black models internals (e.g., what the models have learned) and/or for explaining individual predictions. In general, explainability of an AI model’s prediction is the extent of transferable qualitative understanding of the relationship between model input and prediction (i.e., selective/suitable causes of the event) in a recipient friendly manner. The term “explainability” and “interpretability” are being used interchangeably throughout the literature. To this end, in the case of an intelligent system (i.e., AI-based system), it is evident that explainability is more than interpretability in terms of importance, completeness, and fidelity of prediction. Based on that, we will use these terms accordingly where appropriate.

Due to the increasing number of XAI approaches, it has become challenging to understand the pros, cons, and competitive advantages, associated with the different domains. In addition, there are lots of variations among different XAI methods, such as whether a method is global (i.e., explains the model’s behavior on the entire data set), local (i.e., explains the prediction or decision of a particular instance), ante-hoc (i.e. involved in the pre training stage), post-hoc (i.e. works on already trained model), or surrogate (i.e. deploys a simple model to emulate the prediction of a “black box” model). However, despite many reviews on XAI methods, there is still a lack of comprehensive analysis of

1. <https://www.eugdpr.org>

2. <https://www.senate.gov>

3. <https://www.defense.gov>

XAI when it comes to these methods and perspectives.

Some of the popular work/tools on XAI are LIME, DeepVis Toolbox, TreeInterpreter, Keras-vis, Microsoft InterpretML, MindsDB, SHAP, Tensorboard WhatIf, Tensorflow’s Lucid, Tensorflow’s Cleverhans, etc. However, a few of these work/tools are model specific. For instance, DeepVis, keras-vis, and Lucid are for a neural network’s explainability, and TreeInterpreter is for a tree-based model’s explainability. At a high level, each of the proposed approaches have similar concepts, such as feature importance, feature interactions, shapely values, partial dependence, surrogate models, counterfactual, adversarial, prototypes and knowledge infusion. However, despite some visible progress in XAI methods, the quantification or evaluation of explainability is under-focused, and in particular, when it comes to human study-based evaluations.

In this paper, we (1) demonstrate popular methods/approaches towards XAI with a mutual task (i.e., credit default prediction) and explain the working mechanism in layman’s terms, (2) compare the pros, cons, and competitive advantages of each approach with their associated challenges, and analyze those from multiple perspectives (e.g., global vs local, post-hoc vs ante-hoc, and inherent vs emulated/approximated explainability), (3) provide meaningful insight on quantifying explainability, and (4) recommend a path towards responsible or human-centered AI using XAI as a medium. Our survey is only one among the recent ones (See Table 1) which includes a mutual test case with useful insights on popular XAI methods (See Table 4).

TABLE 1
Comparison with other Surveys

Survey	Reference	Mutual test case
Adadi et al., 2018	[8]	×
Mueller et al., 2019	[9]	×
Samek et al., 2017	[6]	×
Molnar et al., 2019	[10]	×
Staniak et al., 2018	[11]	×
Gilpin et al., 2018	[12]	×
Collaris et al., 2018	[13]	×
Ras et al., 2018	[1]	×
Dosilovic et al., 2018	[14]	×
Tjoa et al., 2019	[15]	×
Dosi-Valez et al., 2017	[16]	×
Rudin et al., 2019	[17]	×
Arrieta et al., 2020	[18]	×
Miller et al., 2018	[19]	×
Zhang et al., 2018	[20]	×
This Survey		✓

We start with a background of related works (Section 2), followed by a description of the test case in Section 3, and then a review of XAI methods in Section 4. We conclude with an overview of quantifying explainability and a discussion addressing open questions and future research directions towards responsible for human-centered AI in Section 5.

2 BACKGROUND

Research interests in XAI are re-emerging. The earlier works such as [21], [22], and [23] focused primarily on explaining the decision process of knowledge-based systems and expert systems. The primary reason behind the renewed

interest in XAI research has stemmed from the recent advancements in AI, its application to a wide range of areas, the concerns over unethical use, lack of transparency, and undesired biases in the models. In addition, recent laws by different governments are necessitating more research in XAI. According to [6] and [7], XAI encompasses Machine Learning (ML) or AI systems for demystifying black models internals (e.g., what the models have learned) and/or for explaining individual predictions.

In 2019, Mueller et al. presents a comprehensive review of the approaches taken by a number of types of “explanation systems” and characterizes those into three generations: (1) first-generation systems—for instance, expert systems from the early 70’s, (2) second generation systems—for instance, intelligent tutoring systems, and (3) third generation systems—tools and techniques from the recent renaissance starting from 2015 [9]. The first generation systems attempt to clearly express the internal working process of the system by embedding expert knowledge in rules often elicited directly from experts (e.g., via transforming rules into natural language expressions). The second generations systems can be regarded as the human-computer system designed around human knowledge and reasoning capacities to provide cognitive support. For instance, arranging the interface in such a way that complements the knowledge that the user is lacking. Similar to the first generation systems, the third generation systems also attempt to clarify the inner workings of the systems. But this time, these systems are mostly “black box” (e.g., deep nets, ensemble approaches). In addition, nowadays, researchers are using advanced computer technologies in data visualizations, animation, and video, that have a strong potential to drive the XAI research further. Many new ideas have been proposed for generating explainable decisions from the need of primarily accountable, fair, and trust-able systems and decisions.

There has been some previous work [10] that mentions three notions for quantification of explainability. Two out of three notions involve experimental studies with humans (e.g., domain expert or a layperson, that mainly investigate whether a human can predict the outcome of the model) [24], [25], [26], [27], [28]. The third notion (proxy tasks) does not involve a human, and instead uses known truths as a metric (e.g., the less the depth of the decision tree, the more explainable the model).

Some mentionable reviews on XAI are listed in Table 1. However, while these works provide analysis from one or more of the mentioned perspectives, a comprehensive review considering all of the mentioned important perspectives, using a mutual test case, is still missing. Therefore, we attempt to provide an overview using a demonstration of a mutual test case or task, and then analyze the various approaches from multiple perspectives, with some future directions of research towards responsible or human-centered AI.

3 TEST CASE

The mutual test case or task that we use in this paper to demonstrate and evaluate the XAI methods is *credit default prediction*. This mutual test case enables a better understanding of the comparative advantages of different

XAI approaches. We predict whether a customer is going to default on a mortgage payment (i.e., unable to pay monthly payment) in the near future or not, and explain the decision using different XAI methods in a human-friendly way. We use the popular Freddie Mac [29] dataset for the experiments. Table 2 lists some important features and their descriptions. The description of features are taken from the data set’s [29] user guide.

We use well-known programming language R’s package “iml” [30] for producing the results for the XAI methods described in this review.

4 EXPLAINABLE ARTIFICIAL INTELLIGENCE METHODS

This section summarizes different explainability methods with their pros, cons, challenges, and competitive advantages primarily based on two recent comprehensive surveys: [31] and [16]. We then enhance the previous surveys with a multi-perspective analysis, recent research progresses, and future research directions. [16] broadly categorize methods for explanations into three kinds: Intrinsically Interpretable Methods, Model Agnostic Methods, and Example-Based Explanations.

4.1 Intrinsically Interpretable Methods

The convenient way to achieve explainable results is to stick with intrinsically interpretable models such as Linear Regression, Logistic Regression, and Decision Trees by avoiding the use of “black box” models. However, usually, this natural explainability comes with a cost in performance.

In a **Linear Regression**, the predicted target consists of the weighted sum of input features. So the weight or coefficient of the linear equation can be used as a medium of explaining prediction when the number of features is small.

$$y = b_0 + b_1 * x_1 + ... + b_n * x_n + \epsilon \quad (1)$$

In Formula 1, y is the target (e.g., chances of credit default), b_0 is a constant value known as the intercept (e.g., .33), b_i is the learned feature’s weight or coefficient (e.g., .33) for the corresponding feature x_i (e.g., credit score), and ϵ is a constant error term (e.g., .0001). Linear regression comes with an interpretable linear relationship among features. However, in cases where there are multiple correlated features, the distinct feature influence becomes indeterminable as the individual influences in prediction are not additive to the overall prediction anymore.

Logistic Regression is an extension of Linear Regression to the classification problems. It models the probabilities for classification tasks. The interpretation of Logistic Regression is different from Linear Regression as it gives a probability between 0 and 1, where the weight might not exactly represent the linear relationship with the predicted probability. However, the weight provides an indication of the direction of influence (negative or positive) and a factor of influence between classes, although it is not additive to the overall prediction.

Decision Tree-based models split the data multiple times based on a cutoff threshold at each node until it reaches a leaf node. Unlike Logistic and Linear Regression,

it works even when the relationship between input and output is non-linear, and even when the features interact with one another (i.e., a correlation among features). In a Decision Tree, a path from the root node (i.e., starting node) (e.g., credit score in Figure 1) to a leaf node (e.g., default) tells how the decision (the leaf node) took place. Usually, the nodes in the upper-level of the tree have higher importance than lower-level nodes. Also, the less the number of levels (i.e., height) a tree has, the higher the level of explainability the tree possesses. In addition, the cutoff point of a node in the Decision Trees provides counterfactual information—for instance, increasing the value of a feature equal to the cutoff point will reverse the decision/prediction. In Figure 1, if the credit score is greater than the cutoff point 748, then the customer is predicted as non-default. Also, tree-based explanations are contrastive, i.e., a “what if” analysis provides the relevant alternative path to reach a leaf node. According to the tree in Figure 1, there are two separate paths (credit score \rightarrow delinquency \rightarrow non-default; and credit score \rightarrow non-default) that lead to a non-default classification.

However, tree-based explanations cannot express the linear relationship between input features and output. It also lacks smoothness; slight changes in input can have a big impact on the predicted output. Also, there can be multiple different trees for the same problem. Usually, the more the nodes or depth of the tree, the more challenging it is to interpret the tree.

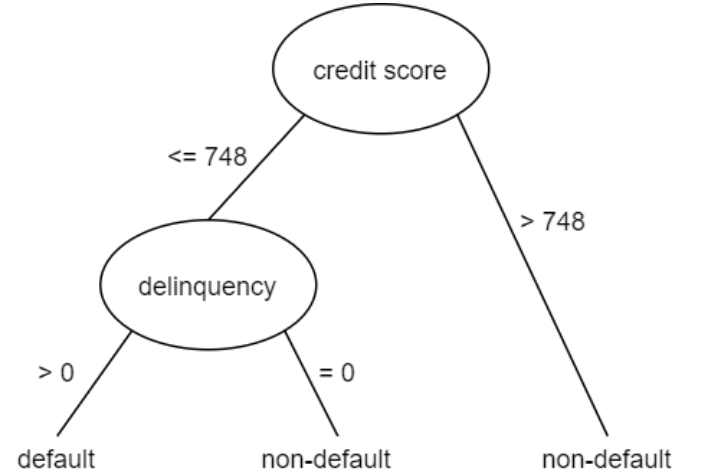


Fig. 1. Decision Trees

Decision Rules (simple IF-THEN-ELSE conditions) are also an inherent explanation model. For instance, “IF credit score is less than or equal to 748 AND if the customer is delinquent on payment for more than zero days (condition), THEN the customer will default on payment (prediction)”. Although IF-THEN rules are straightforward to interpret, it is mostly limited to classification problems (i.e., does not support a regression problem), and inadequate in describing linear relationships. In addition, the **RuleFit** algorithm [32] has an inherent interpretation to some extent as it learns sparse linear models that can detect the interaction effects in the form of decision rules. Decision rules consist of the combination of split decisions from each of the decision paths. However, besides the original features, it also learns some new features to capture the interaction effects of

TABLE 2
Dataset description

Feature	Description
creditScore	A number in between 300 and 850 that indicates the creditworthiness of the borrowers.
originalUPB	Unpaid principle balance on the note date.
originalInterestRate	Original interest rate as indicated by the mortgage note.
currentLoanDelinquencyStatus	Indicates the number of days the borrower is delinquent.
numberOfBorrower	Number of borrower who are obligated to repay the loan.
currentInterestRate	Active interest rate on the note.
originalCombinedLoanToValue	Ratio of all mortgage loans and appraised price of mortgaged property on the note date.
currentActualUPB	Unpaid principle balance as of latest month of payment.
defaulted	Whether the customer was default on payment (1) or not (0.)

original features. Usually, interpretability degrades with an increasing number of features.

Other interpretable models include the extension of linear models such as **Generalized Linear Models (GLMs)** and **Generalized Additive Models (GAMs)**; they help to deal with some of the assumptions of linear models (e.g., the target outcome y and given features follow a Gaussian Distribution; and no interaction among features). However, these extensions make models more complex (i.e., added interactions) as well as less interpretable. In addition, a **Naïve Bayes Classifier** based on Bayes Theorem, where the probability of classes for each of the features is calculated independently (assuming strong feature independence), and **K-Nearest Neighbors**, which uses nearest neighbors of a data point for prediction (regression or classification), also fall under intrinsically interpretable models.

4.2 Model-Agnostic Methods

Model-agnostic methods separate explanation from a machine learning model, allowing the explanation method to be compatible with a variety of models. This separation has some clear advantages such as (1) the interpretation method can work with multiple ML models, (2) provides different forms of explainability (e.g., visualization of feature importance, linear formula) for a particular model, and (3) allows for a flexible representation—a text classifier uses abstract word embedding for classification but uses actual words for explanation. Some of the model-agnostic interpretation methods include Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Accumulation Local Effects (ALE) Plot, Feature Interaction, Feature Importance, Global Surrogate, Local Surrogate (LIME), and Shapley Values (SHAP).

4.2.1 Partial Dependence Plot (PDP)

The partial Dependence Plot (PDP) or PD plot shows the marginal effect of one or two features (at best three features in 3-D) on the predicted outcome of an ML model [33]. It is a global method, as it shows an overall model behavior, and is capable of showing the linear or complex relationships between target and feature(s). It provides a function that depends only on the feature(s) being plotted by marginalizing over other features in such a way that includes the interactions among them. PDP provides a clear and causal interpretation by providing the changes in prediction due to changes in particular features. However, PDP assumes features under the plot are not correlated with the remaining

features. In the real world, this is unusual. Furthermore, there is a practical limit of only two features that PD plot can clearly explain at a time. Also, it is a global method, as it plots the average effect (from all instances) of a feature(s) on the prediction, and not for all features on a specific instance. The PD plot in Figure 2 shows the effect of credit score on prediction. Individual bar lines along the X axis represent the frequency of samples for different ranges of credit scores.

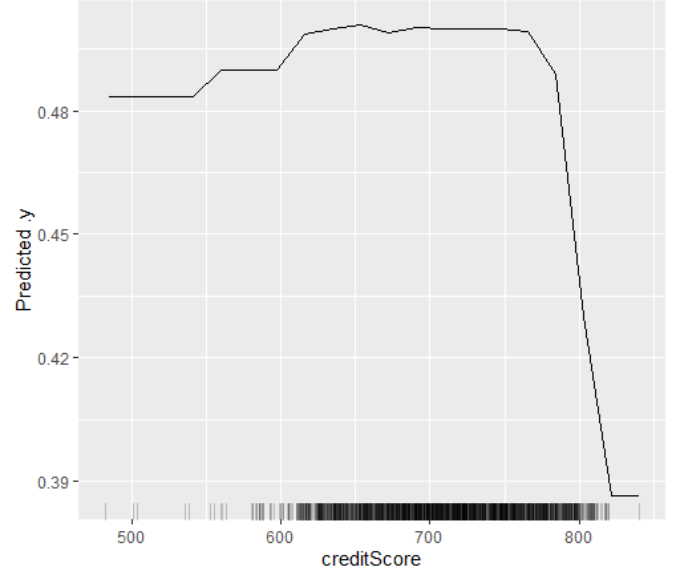


Fig. 2. Partial Dependence Plot (PDP)

4.2.2 Individual Conditional Expectation (ICE)

Unlike PDP, ICE plots one line per instance showing how a feature influences the changes in prediction (See Figure 3. The average on all lines of an ICE plot gives a PD plot [34] (i.e., the single line shown in the PD plot in Figure 2). Figure 4, combines both PDP and ICE together for a better interpretation.

Although ICE curves are more intuitive to understand than a PD plot, it can only display one feature meaningfully at a time. In addition, it also suffers from the problem of correlated features and overcrowded lines when there are many instances.

4.2.3 Accumulated Local Effects (ALE) Plot

Similar to PD plots (Figure 2, ALE plots (Figure 5 describe how features influence the prediction on average. However,

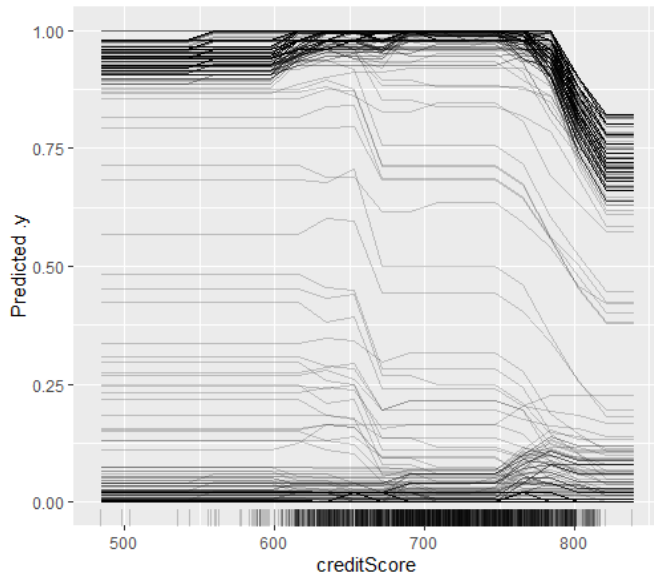


Fig. 3. Individual Conditional Expectation (ICE)

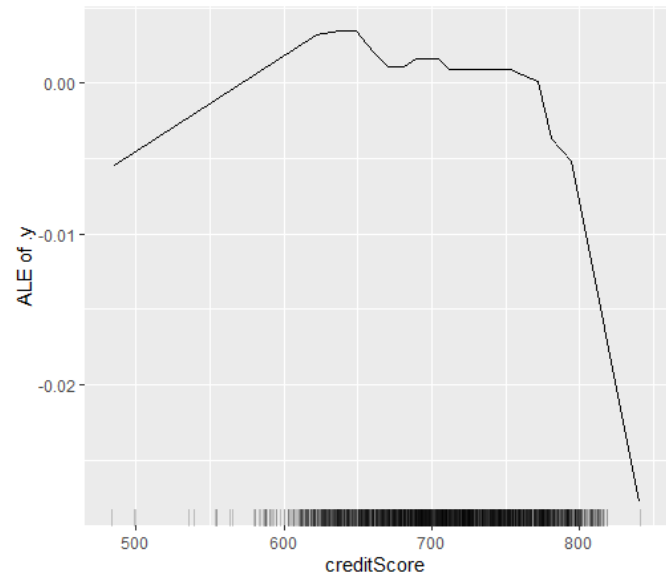


Fig. 5. Accumulated Local Effects (ALE) Plot

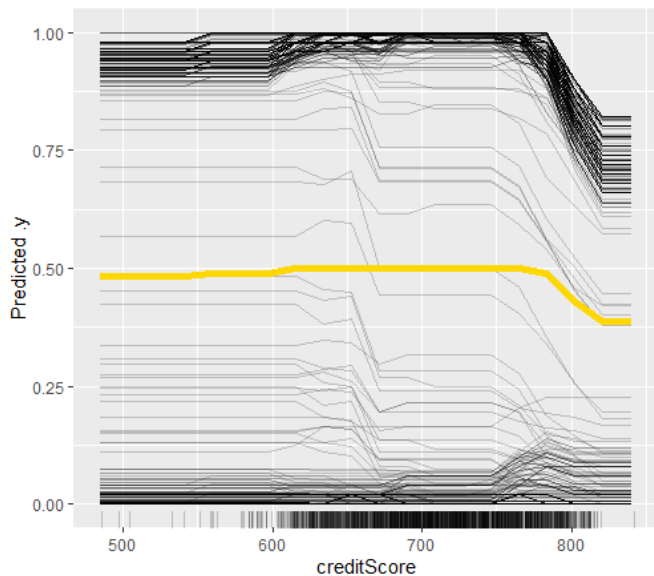


Fig. 4. PDP and ICE combined together in the same plot

unlike PDP, ALE plot reasonably works well with correlated features and is comparatively faster. Although ALE plot is not biased to the correlated features, it is challenging to interpret the changes in prediction when features are strongly correlated and analyzed in isolation. In that case, only plots showing changes in both correlated features together make sense to understand the changes in the prediction.

4.2.4 Feature Interaction

When the features interact with one another, individual feature effects do not sum up to the total feature effects from all features combined. An H-statistic (i.e., Friedman's H-statistic) helps to detect different types of interaction, even with three or more features. The interaction strength between two features is the difference between the *partial dependence function for those two features together* and the sum

of the *partial dependence functions for each feature separately*. Figure 6 shows the interaction strength of each participating feature. For example, *current Actual UPB* has the highest level of interaction with other features, and *credit score* has the least interaction with other features. However, calculating feature interaction is computationally expensive. Furthermore, using sampling instead of the entire dataset usually shows variances from run to run. 6,

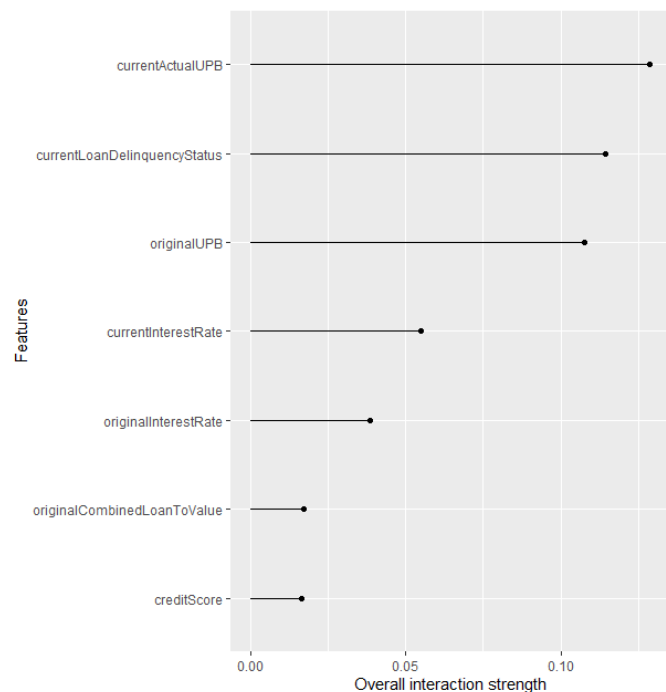


Fig. 6. Feature interaction

4.2.5 Feature Importance

Usually, the feature importance of a feature is the increase in the prediction error of the model when we permute

the values of the feature to break the true relationship between the feature and the true outcome. After shuffling the values of the feature, if errors increase, then the feature is important. [35] introduced the permutation-based feature importance for Random Forests; later [36] extended the work to a model-agnostic version. Feature importance provides a compressed and global insight into the ML model's behavior. For example, Figure 7 shows the importance of each participating feature, *current Actual UPB* possess the highest feature importance, and *credit score* possess the lowest feature importance. Although feature importance takes into account both the main feature effect and interaction, this is a disadvantage as feature interaction is included in the importance of correlated features. We can see that the feature *current Actual UPB* possesses the highest feature importance (Figure 7), at the same time it also possesses the highest interaction strength 6. As a result, in the presence of interaction among features, the feature importance does not add up to total drop-in of performance. Besides, it is unclear whether the test set or training set should be used for feature importance, as it demonstrates variance from run to run in the shuffled dataset. It is necessary to mention that feature importance also falls under the global methods.

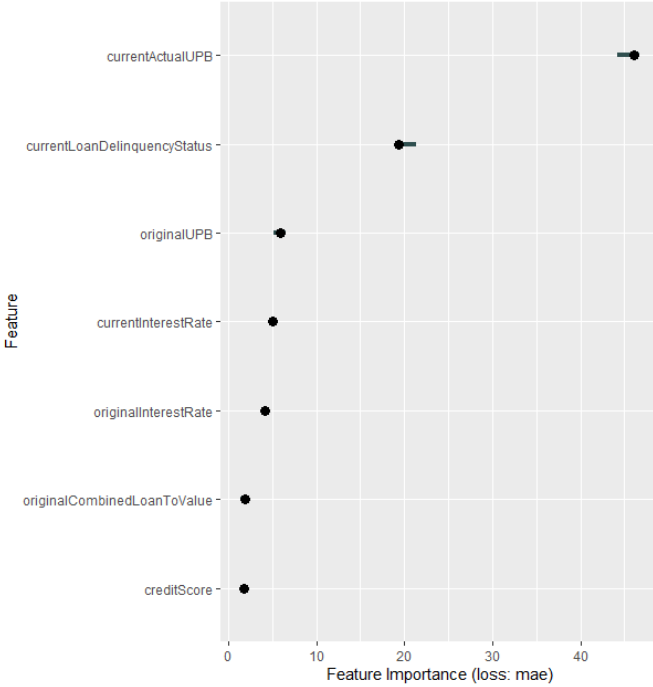


Fig. 7. Feature importance

4.2.6 Global Surrogate

A global surrogate model tries to approximate the overall behavior of a “black box” model using an interpretable ML model. In other words, surrogate models try to approximate the prediction function of a black-box model using an interpretable model as correctly as possible, given the prediction is interpretable. It is also known as a meta-model, approximate model, response surface model, or emulator. We approximate the behavior of a Random Forest using CART decision trees (Figure 8). The original black box

model could be avoided given the surrogate model demonstrates a comparable performance. Although a surrogate model comes with interpretation and flexibility (i.e., such as model agnosticism), diverse explanations for the same “black box” such as multiple possible decision trees with different structures, is a drawback. Besides, some would argue that this is only an illusion of interpretability.

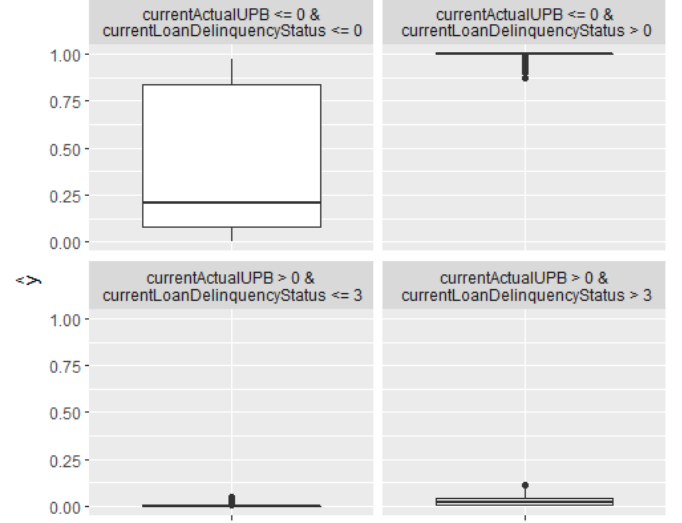


Fig. 8. Global surrogate

4.2.7 Local Surrogate (LIME)

Unlike global surrogate, local surrogate explains individual predictions of black-box models. Local Interpretable Model-Agnostic Explanations (LIME) was proposed by [37]. LIME trains an inherently interpretable model (e.g., Decision Trees) on a new dataset made from the permutation of samples and the corresponding prediction of the black box. Although the learned model can have a good approximation of local behavior, it does not have a good global approximation. This trait is also known as local fidelity. Figure 9 is a visualization of the output from LIME. For a random sample, the black box predicts that a customer will default on payment with a probability of 1; the local surrogate model, LIME also predict that the customer will default on the payment, however, the probability is 0.99, that is little less than the black box models prediction. LIME also shows which feature contributes to the decision making and by how much. Furthermore, LIME allows replacing the underlying “black box” model by keeping the same local interpretable model for the explanation. In addition, LIME works for tabular data, text, and images. As LIME is an approximation model, and the local model might not cover the complete attribution due to the generalization (e.g., using shorter trees, lasso optimization), it might be unfit for cases where we legally need complete explanations of a decision. Furthermore, there is no consensus on the boundary of the neighborhood for the local model; sometimes, it provides very different explanations for two nearby data points.

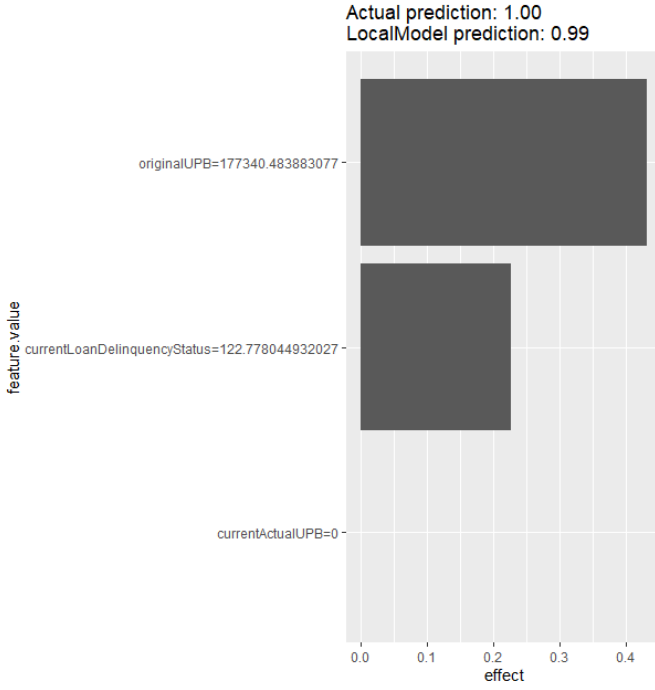


Fig. 9. Local Interpretable Model-Agnostic Explanations (LIME)

4.2.8 Shapley Values

Shapley is another local explanation method. In 1953, Shapley [38] coined the Shapley Value. It is based on coalitional game theory that helps to distribute feature importance among participating features fairly. Here the assumption is that each feature value of the instance is a player in a game, and the prediction is the overall payout that is distributed among players (i.e., features) according to their contribution to the total payout (i.e., prediction). We use Shapely values (See Figure 10) to analyze the prediction of a random forest model for the credit default prediction problem. The actual prediction for a random sample is 1.00, the average prediction from all samples in the data set is 0.53, and their difference .47 ($1.00 - 0.53$) consists of the individual contributions from the features (e.g., *Current Actual UPB* contributes 0.36). The Shapely Value is the average contribution in prediction over all possible coalition of features, which make it computationally expensive when there is a large number of features—for example, for k number of features, there will be 2^k number of coalitions. Unlike LIME, Shapely Value is an explanation method with a solid theory that provides full explanations. However, it also suffers from the problem of correlated features. Furthermore, the Shapely value returns a single value per feature; there is no way to make a statement about the changes in output resulting from the changes in input. One mentionable implementation of the Shapely value is in the work of [39] that they call SHAP.

4.2.9 Break Down

The Break Down package provides the local explanation and is loosely related to the partial dependence algorithm with an added step-wise procedure known as “Break Down” (proposed by [11]). It uses a greedy strategy to identify

and remove features iteratively based on their influence on the overall average predicted response (baseline) [40]. For instance, from the game theory perspective, it starts with an empty team, then adds feature values one by one based on their decreasing contribution. In each iteration, the amount of contribution from each feature depends on the features values of those are already in the team, which is considered as a drawback of this approach. However, it is faster than the Shapley value method due to the greedy approach, and for models without interactions, the results are the same [31]. Figure 11 is a visualization of *break down* for a random sample, showing contribution (positive or negative) from each of the participating features towards the final prediction.

4.3 Example-Based Explanations

Example-Based Explanation methods use particular instances from the dataset to explain the behavior of the model and the distribution of the data in a model agnostic way. It can be expressed as “X is similar to Y and Y caused Z, so the prediction says X will cause Z”. According to [31], a few explanation methods that fall under Example-Based Explanations are described as follows:

4.3.1 Counterfactual

The counterfactual method indicates the required changes in the input side that will have significant changes (e.g., reverse the prediction) in the prediction/output. Counterfactual explanations can explain individual predictions. For instance, it can provide an explanation that describes causal situations such as “If A had not occurred, B would not have occurred”. Although counterfactual explanations are human-friendly, it suffers from the “Rashomon effect”, where each counterfactual explanation tells a different story to reach a prediction. In other words, there are multiple true explanations (counterfactual) for each instance level prediction, and the challenge is how to choose the best one. The counterfactual methods do not require access to data or models and could work with a system that does not use machine learning at all. In addition, this method does not work well for categorical variables with many values. For instance, if the credit score of customer 5 (from Table 3) can be increased to 749 (similar to the credit score of customer 6) from 748, given other features values remain unchanged, the customer will not default on a payment. In short, there can be multiple different ways to tune feature values to make customers move from non-default to default, or vice versa.

Traditional explanation methods are mostly based on explaining correlation rather than causation. Moraffah et al. [41] focus on the causal interpretable model that explains the possible decision under different situations such as being trained with different inputs or hyperparameters. This causal interpretable approach share concept of counterfactual analysis as both work on causal inference. Their work also suggests possible use in fairness criteria evaluation of decisions.

4.3.2 Adversarial

An adversarial technique is capable of flipping the decision using counterfactual examples to fool the machine learner

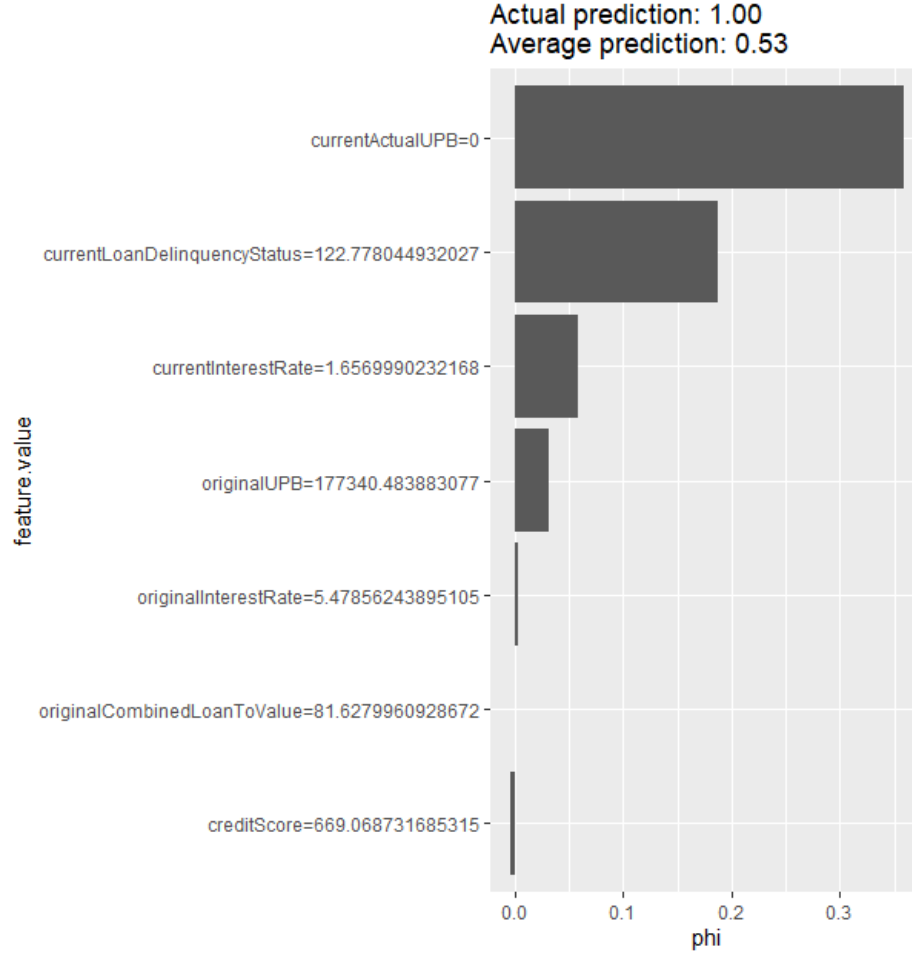


Fig. 10. Shapely values

TABLE 3
Example-Based Explanations

Customer	Delinquency	Credit score	Defaulted
1	162	680	yes
2	149	691	yes
3	6	728	yes
4	6	744	yes
5	0	748	yes
6	0	749	no
7	0	763	no
8	0	790	no
9	0	794	no
10	0	806	no

(i.e., small intentional perturbations in input to make a false prediction). However, adversarial examples could help to discover hidden vulnerabilities as well as to improve the model. For instance, an attacker can intentionally design adversarial examples to cause the AI system to make a mistake (i.e., fooling the machine), which poses greater threats to cyber-security and autonomous vehicles. As an example, the credit default prediction system can be fooled for customer 5, just by increasing the credit score by 1 (see Table 3), leading to a reversed prediction.

Hartl et al. [42] emphasize on understanding the implica-

tions of adversarial samples on Recurrent Neural Network (RNNs) based IDS because RNNs are good for sequential data analysis, and network traffic exhibits some sequential patterns. They find that adversarial the adversarial training procedure can significantly reduce the attack surface. Furthermore, [43] apply an adversarial approach to finding minimum modification of the input features of an intrusion detection system needed to reverse the classification of the misclassified instance. Besides satisfactory explanations of the reason for misclassification, their approach work provide further diagnosis capabilities.

4.3.3 Prototypes

Prototypes consist of a selected set of instances that represent the data very well. Conversely, the set of instances that do not represent data well are called criticisms [44]. Determining the optimal number of prototypes and criticisms are challenging. For example, customers 1 and 10 from Table 3 can be treated as prototypes as those are strong representatives of the corresponding target. On the other hand, customers 5 and 6 (from Table 3) can be treated as a criticism as the distance between the data points is minimal, and they might be classified under either class from run to run of the same or different models.

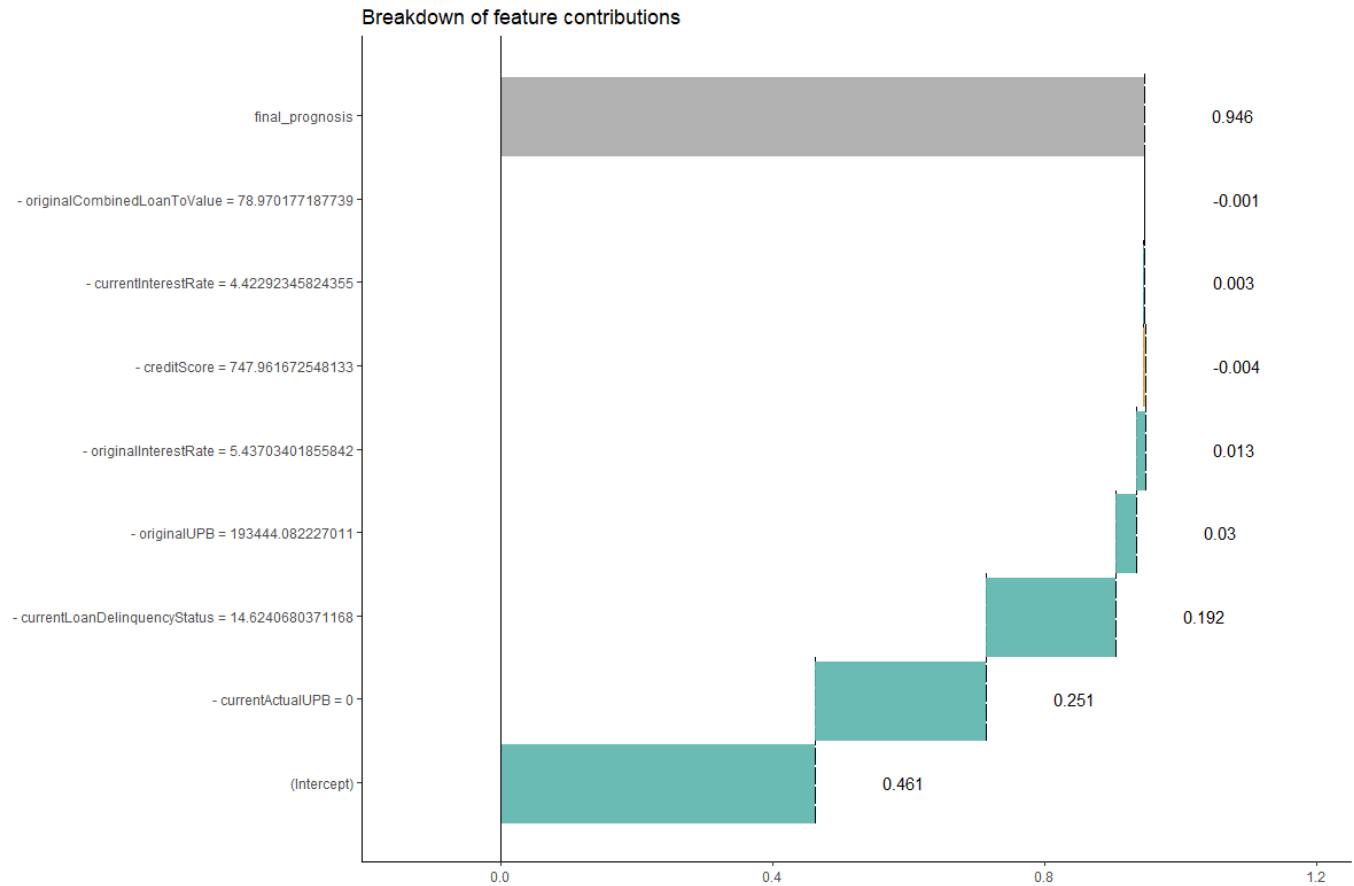


Fig. 11. Breakdown

4.3.4 Influential Instances

Influential instances are data points from the training set that are influential for prediction and parameter determination of the model. While it helps to debug the model and understand the behavior of the model better, determining the right cutoff point to separate influential or non-influential instances is challenging. For example, based on the values of feature credit score and delinquency, customers 1, 2, 9, and 10 from Table 3 can be treated as influential instances as those are strong representatives of the corresponding target. On the other hand, customers 5 and 6 are not influential instances, as those would be in the margin of the classification decision boundary.

4.3.5 k-nearest Neighbors Model

The prediction of the k-nearest neighbor model can be explained with the k-neighbor data points (neighbors those were averaged to make the prediction). A visualization of the individual cluster containing similar instances provides an interpretation of why an instance is a member of a particular group or cluster. For example, in Figure 12, the new sample (black circle) is classified according to the other three (3-nearest neighbor) nearby samples (one gray, two white). This visualization gives an interpretation of why a particular sample is part of a particular class.

Table 4 summarizes the explainability methods from the perspective of (A) whether the method approximates the

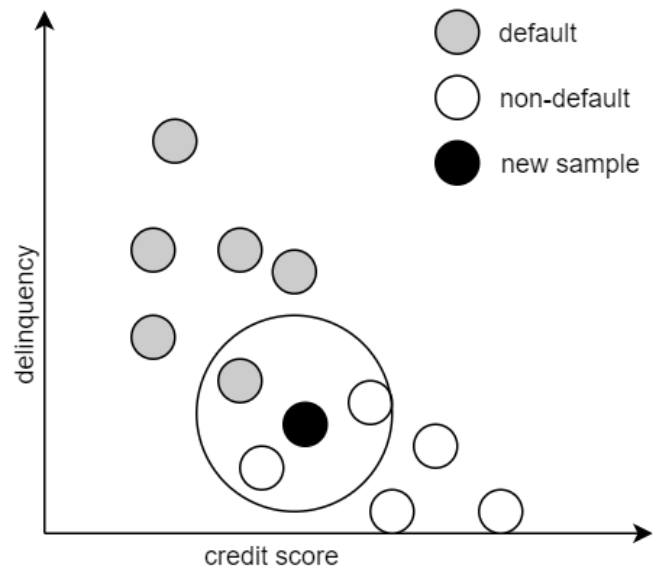


Fig. 12. KNN

model behavior (i.e., creates an illusion of interpretability) or finds actual behavior, (B) whether the method alone is inherently interpretable or not, (C) whether the interpretation method is ante-hoc, that is, it incorporates explain-

ability into a model from the beginning, or post-hoc, where explainability is incorporated after the regular training of the actual model (i.e., testing time), (D) whether the method is model agnostic (i.e., works for any ML model) or specific to an algorithm, and (E) whether the model is local, providing instance-level explanations, or global, providing overall model behavior.

Our analysis says there is a lack of an explainability method (i.e., a gap in the literature), which is, at the same time actual and direct (i.e., does not create an illusion of explainability by approximating the model), model agnostic, and local, such that it utilizes the full potential of the explainability method in different applications. There are some recent works that bring external knowledge and infuse that into the model for better interpretation. These XAI methods have the potential to fill the gap to some extent by incorporating domain knowledge into the model in a model agnostic and transparent way (i.e., not by illusion).

4.4 Other Techniques

Chen et al. [45] introduce an instance-wise feature selection as a methodology for model interpretation where the model learns a function to extract a subset of most informative features for a particular instance. The feature selector attempt to maximize the mutual information between selected features and response variables. However, their approach is mostly limited to posthoc approaches.

In a more recent work, [46] study explainable ML using information theory where they quantify the effect of an explanation by the conditional mutual information between the explanation and prediction considering user background. Their approach provides personalized explanation based on the background of the recipient, for instance, a different explanation for those who know linear algebra and those who don't. However, this work is yet to be considered as a comprehensive approach which considers a variety of user and their explanation needs. To understand the flow of information in a Deep Neural Network (DNN), [47] analyzed different gradient-based attribution methods that assign an attribution value (i.e., contribution or relevance) to each input feature (i.e., neuron) of a network for each output neurons. They use a heatmap for better visualizations where a particular color represents features that contribute positively to the activation of target output, and another color for features that suppress the effect on it.

A survey on the visual representation of Convolutional Neural Networks (CNNs), by [20], categorizes works based on a) visualization of CNN representations in intermediate network layers, b) diagnosis of CNN representation for feature space of different feature categories or potential representation flaws, c) disentanglement of "the mixture of patterns" encoded in each filter of CNNs, d) interpretable CNNs, and e) semantic disentanglement of CNN representations.

In the industrial control system, an alarm from the intrusion/anomaly detection system has a very limited role unless the alarm can be explained with more information. [5] design a layer-wise relevance propagation method for DNN to map the abnormalities between the calculation process and features. This process helps to compare the normal

samples with abnormal samples for better understanding with detailed information.

4.5 Knowledge Infusion Techniques

[48] propose a concept attribution-based approach (i.e., sensitivity to the concept) that provides an interpretation of the neural network's internal state in terms of human-friendly concepts. Their approach, *Testing with CAV (TCAV)*, quantifies the prediction's sensitivity to a high dimensional concept. For example, a user-defined set of examples that defines the concept 'striped', TCAV can quantify the influence of 'striped' in the prediction of 'zebra' as a single number. However, their work is only for image classification and falls under the post-modeling notion (i.e., post-hoc) of explanation.

[49] propose a knowledge-infused learning that measures information loss in latent features learned by the neural networks through Knowledge Graphs (KGs). This external knowledge incorporation (via KGs) aids in supervising the learning of features for the model. Although much work remains, they believe that (KGs) will play a crucial role in developing explainable AI systems.

[50] and [51] infuse popular domain principles from the domain in the model and represent the output in terms of the domain principle for explainable decisions. In [50], for a bankruptcy prediction problem they use the 5C's of credit as the domain principle which is commonly used to analyze key factors: character (reputation of the borrower/firm), capital (leverage), capacity (volatility of the borrower's earnings), collateral (pledged asset) and cycle (macroeconomic conditions) [52], [53]. In [51], for an intrusion detection and response problem, they incorporate the CIA principles into the model; C stands for *confidentiality*—concealment of information or resources, I stands for *integrity*—trustworthiness of data or resources, and A stands for *availability*—ability to use the information or resource desired [54]. In both cases, the infusion of domain knowledge leads to better explainability of the prediction with negligible compromises in performance. It also comes with better execution time and a more generalized model that works better with unknown samples.

Although these works [50], [51] come with unique combinations of merits such as model agnosticism, the capability of both local and global explanation, and authenticity of explanation—simulation or emulation free, they are still not fully off-the-shelf systems due to some domain-specific configuration requirements. Much work still remains and needs further attention.

5 QUANTIFYING EXPLAINABILITY AND FUTURE RESEARCH DIRECTIONS

5.1 Quantifying Explainability

The quantification or evaluation of explainability is an open challenge. There are two primary directions of research towards the evaluation of explainability of an AI/ML model: (1) model complexity-based, and (2) human study-based.

TABLE 4

Comparison of different explainability methods from a set of key perspectives (approximation or actual; inherent or not; post-hoc or ante-hoc; model-agnostic or model specific; and global or local)

Method	Approx.	Inherent	Post/Ante	Agnos./Spec.	Global/Local
Linear/Logistic Regression	No	Yes	Ante	Specific	Both
Decision Trees	No	Yes	Ante	Specific	Both
Decision Rules	No	Yes	Ante	Specific	Both
k-Nearest Neighbors	No	Yes	Ante	Specific	Both
Partial Dependence Plot (PDP)	Yes	No	Post	Agnostic	Global
Individual Conditional Expectation (ICE)	Yes	No	Post	Agnostic	Both
Accumulated Local Effects (ALE) Plot	Yes	No	Post	Agnostic	Global
Feature Interaction	No	Yes	Both	Agnostic	Global
Feature Importance	No	Yes	Both	Agnostic	Global
Global Surrogate	Yes	No	Post	Agnostic	Global
Local Surrogate (LIME)	Yes	No	Post	Agnostic	Local
Shapley Values (SHAP)	Yes	No	Post	Agnostic	Local
Break Down	Yes	No	Post	Agnostic	Local
Counterfactual explanations	Yes	No	Post	Agnostic	Local
Adversarial examples	Yes	No	Post	Agnostic	Local
Prototypes	Yes	No	Post	Agnostic	Local
Influential instances	Yes	No	Post	Agnostic	Local

5.1.1 Model Complexity-based Explainability Evaluation

In the literature, model complexity and (lack of) model interpretability are often treated as the same [10]. For instance, in [55], [56], model size is often used as a measure of interpretability (e.g., number of decision rules, depth of the tree, number of non-zero coefficients).

[56] propose a scalable Bayesian Rule List (i.e., probabilistic rule list) consisting of a sequence of IF-THEN rules, identical to a decision list or one-sided decision tree. Unlike the decision tree that uses greedy splitting and pruning, their approach produces a highly sparse and accurate rule list with a balance between interpretability, accuracy, and computation speed. Similarly, the work of [55] is also rule-based. They attempt to evaluate the quality of the rules using a rule learning algorithm by: the observed coverage, which is the number of positive examples covered by the rule, which should be maximized to explain the training data well; and consistency, which is the number of negative examples covered by the rule, which should be minimized to generalize well to unseen data.

According to [57], while the number of features and the size of the decision tree are directly related to interpretability, the optimization of the tree size or features (i.e., feature selection) is costly as it requires the generation of a large set of models and their elimination in subsequent steps. However, reducing the tree size (i.e., reducing complexity) increases error, as they could not find a way to formulate the relation in a simple functional form. More recently, [10] attempts to quantify the complexity of the arbitrary machine learning model with a model agnostic measure. In that work, the author demonstrates that when the feature interaction (i.e., the correlation among features) increases, the quality of representations of explainability tools degrades. For instance, the explainability tool ALE Plot (see Figure 5 starts to show harsh lines (i.e., zigzag lines) as feature interaction increases. In other words, with more interaction comes a more combined influence in the prediction, induced from different correlated subsets of features (at least two),

which ultimately makes it hard to understand the causal relationship between input and output, compared to an individual feature influence in the prediction. In fact, from our study of different explainability tools (e.g., LIME, SHAP, PDP), we have found that the correlation among features is a key stumbling block to represent feature contribution in a model agnostic way. Keeping the issue of feature interactions in mind, [10] propose a technique that uses three measures: number of features, interaction strength among features, and the main effect (excluding the interaction part) of features, to measure the complexity of a post-hoc model for explanation.

Although, [10] mainly focuses on model complexity for post-hoc models, their work was a foundation for the approach by [58] for the quantification of explainability. Their approach to quantify explainability is model agnostic and is for a model of any notion (e.g., pre-modeling, post-hoc) using proxy tasks that do not involve a human. Instead, they use known truth as a metric (e.g., the less number of features, the more explainable the model). Their proposed formula for explainability gives a score in between 0 and 1 for explainability based on the number of cognitive chunks (i.e., individual pieces of information) used on the input side and output side, and the extent of interaction among those cognitive chunks.

5.1.2 Human Study-based Explainability Evaluation

The following works deal with the application-level and human-level evaluation of explainability involving human studies.

[26] investigate the suitability of different alternative representation formats (e.g., decision tables, (binary) decision trees, propositional rules, and oblique rules) for classification tasks primarily focusing on the explainability of results rather than accuracy or precision. They discover that decision tables are the best in terms of accuracy, response time, the confidence of answer, and ease of use.

[24] argue that interpretability is not an absolute concept; instead, it is relative to the target model, and may or

may not be relative to the human. Their finding suggests that a model is readily interpretable to a human when it uses no more than seven pieces of information [59]. Although, this might vary from task to task and person to person. For instance, a domain expert might consume a lot more detailed information depending on their experience.

The work of [27] is a human-centered approach, focusing on previous work on human trust in a model from psychology, social science, machine learning, and human-computer interaction communities. In their experiment with human subjects, they vary factors (e.g., number of features, whether the model internals are transparent or a black box) that make a model more or less interpretable and measures how the variation impacts the prediction of human subjects. Their results suggest that participants who were shown a transparent model with a small number of features were more successful in simulating the model's predictions and trusted the model's predictions.

[25] investigate interpretability of a model based on two of its definitions: simulatability, which is a user's ability to predict the output of a model on a given input; and "what if" local explainability, which is a user's ability to predict changes in prediction in response to changes in input, given the user has the knowledge of a model's original prediction for the original input. They introduce a simple metric called *runtime operation count* that measures the interpretability, that is, the number of operations (e.g., the arithmetic operation for regression, the boolean operation for trees) needed in a user's mind to interpret something. Their findings suggest that interpretability decreases with an increase in the number of operations.

Despite some progress, there are still some open challenges surrounding explainability such as an agreement of what an explanation is and to whom; a formalism for the explanation; and quantifying the human comprehensibility of the explanation. Other challenges include addressing more comprehensive human studies requirements and investigating the effectiveness among different approaches (e.g., supervised, unsupervised, semi-supervised) for various application areas (e.g., natural language processing, image recognition).

5.2 Future Research Directions

The long term goal for current AI initiatives is to contribute to the design, development, and deployment of human-centered artificial intelligent systems, where the agents collaborate with the human in an interpretable and explainable manner, with the intent on ensuring fairness, transparency, and accountability. To accomplish that goal, we propose a set of research plans/directions towards achieving responsible or human-centered AI using XAI as a medium.

5.2.1 A Generic Framework to Formalize Explainable Artificial Intelligence

The work in [50] and [51], demonstrates a way to collect and leverage domain knowledge from two different domains, finance and cybersecurity, and further infused that knowledge into black-box models for better explainability. In both of these works, competitive performance with enhanced explainability is achieved. However, there are some

open challenges such as (A) a lack of formalism of the explanation, (B) a customized explanation for different types of explanation recipients (e.g., layperson, domain expert, another machine), (C) a way to quantify the explanation, and (D) quantifying the level of comprehensibility with human studies. Therefore, leveraging the knowledge from multiple domains, a generic framework could be useful considering the mentioned challenges. As a result, mission-critical applications from different domains will be able to leverage the black-box model with greater confidence and regulatory compliance.

5.2.2 Towards Fair, Accountable, and Transparent AI-based Models

Responsible use of AI is crucial for avoiding risks stemming from a lack of fairness, accountability, and transparency in the model. Remediation of data, algorithmic, and societal biases is vital to promote fairness; the AI system/adopter should be held accountable to affected parties for its decision; and finally, an AI system should be analyzable, where the degree of transparency should be comprehensible to have trust in the model and its prediction for mission-critical applications. Interestingly, XAI enhances understating directly, increasing trust as a side-effect. In addition, the explanation techniques can help in uncovering potential risks (e.g., what are possible fairness risks). So it is crucial to adhere to fairness, accountability, and transparency principles in the design and development of explainable models.

5.2.3 Human-Machine Teaming

To ensure the responsible use of AI, the design, development, and deployment of human-centered AI, that collaborates with the humans in an explainable manner, is essential. Therefore, the explanation from the model needs to be comprehensible by the user, and there might be some supplementary questions that need to be answered for a clear explanation. So, the interaction (e.g., follow-ups after the initial explanation) between humans and machines is important. The interaction is more crucial for adaptive explainable models that provide context-aware explanations based on user profiles such as expertise, domain knowledge, interests, and cultural backgrounds. The social sciences and human behavioral studies have the potential to impact XAI and human-centered AI research. Unfortunately, the Human-Computer Interaction (HCI) community is kind of isolated. The combination of HCI empirical studies and human science theories could be a compelling force for the design of human-centered AI models as well as furthering XAI research. Therefore, efforts to bring a human into the loop, enabling the model to receive input (repeated feedback) from the provided visualization/explanations to the human, and improving itself with the repeated interactions, has the potential to further human-centered AI. Besides adherence to fairness, accountability, and transparency, the effort will also help in developing models that adhere to our ethics, judgment, and social norms.

5.2.4 Collective Intelligence from Multiple Disciplines

From the explanation perspective, there is plenty of research in philosophy, psychology, and cognitive science on how

people generate, select, evaluate, and represent explanations and associate cognitive biases and social expectations in the explanation process. In addition, from the interaction perspective, human-computer teaming involving social science, the HCI community, and social-behavioral studies could combine for further breakthroughs. Furthermore, from the application perspective, the collectively learned knowledge from different domains (e.g., Health-care, Finance, Medicine, Security, Defense) can contribute to furthering human-centric AI and XAI research. Thus, there is a need for a growing interest in multidisciplinary research to promote human-centric AI as well as XAI in mission-critical applications from different domains.

6 CONCLUSION

We demonstrate and analyze mutual XAI methods using a mutual test case to explain competitive advantages and elucidate the challenges and further research directions. Most of the available works on XAI are on the post-hoc notion of explainability. However, the post-hoc notion of explainability is not purely transparent and can be misleading, as it explains the decision after it has been made. The explanation algorithm can be optimized to placate subjective demand, primarily stemming from the emulation effort of the actual prediction, and the explanation can be misleading, even when it seems plausible [60], [61]. Thus, many suggest not to explain black-box models using post-hoc notions, instead, they suggest adhering to simple and intrinsically explainable models for high stakes decisions [17]. Furthermore, from the literature review, we find that explainability in pre-modeling is a viable option to avoid the transparency related issues, albeit, under-focused. In addition, knowledge infusion techniques have the potential to enhance explainability greatly, although, also an under-focused challenge. Therefore, we need more focus on the explainability of “black box” models using domain knowledge. At the same time, we need to focus on the evaluation or quantification of explainability using both human and non-human studies. We believe this review provides a good insight into the current progress on XAI approaches, evaluation and quantification of explainability, open challenges, and a path towards responsible or human-centered AI using XAI as a medium.

ACKNOWLEDGMENTS

Our sincere thanks to Christoph Molnar for his open E-book on Interpretable Machine Learning and contribution to the open-source R package “iml”. Both were very useful in conducting this survey.

REFERENCES

- [1] G. Ras, M. van Gerven, and P. Haselager, “Explanation methods in deep learning: Users, values, concerns and challenges,” in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 19–36.
- [2] B. Goodman and S. Flaxman, “Eu regulations on algorithmic decision-making and a “right to explanation”,” in *ICML workshop on human interpretability in machine learning (WHI 2016)*, New York, NY. <http://arxiv.org/abs/1606.08813> v1, 2016.
- [3] B. Wyden, “Algorithmic accountability,” <https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf>, (Accessed on 11/21/2019).
- [4] M. T. Esper, “Ai ethical principles,” <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>, February 2020, (Accessed on 03/07/2020).
- [5] Z. Wang, Y. Lai, Z. Liu, and J. Liu, “Explaining the attributes of a deep learning based intrusion detection system for industrial control networks,” *Sensors*, vol. 20, no. 14, p. 3817, 2020.
- [6] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.
- [7] A. Fernandez, F. Herrera, O. Cordón, M. J. del Jesus, and F. Marcelloni, “Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to?” *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, 2019.
- [8] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (xai),” *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.
- [9] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, “Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai,” *arXiv preprint arXiv:1902.01876*, 2019.
- [10] C. Molnar, G. Casalicchio, and B. Bischl, “Quantifying model complexity via functional decomposition for better post-hoc interpretability,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 193–204.
- [11] M. Staniak and P. Biecek, “Explanations of model predictions with live and breakdown packages,” *arXiv preprint arXiv:1804.01955*, 2018.
- [12] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.
- [13] D. Collaris, L. M. Vink, and J. J. van Wijk, “Instance-level explanations for fraud detection: A case study,” *arXiv preprint arXiv:1806.07129*, 2018.
- [14] F. K. Došilović, M. Brčić, and N. Hlupić, “Explainable artificial intelligence: A survey,” in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.
- [15] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (xai): towards medical xai,” *arXiv preprint arXiv:1907.07374*, 2019.
- [16] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.
- [17] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
- [18] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bénéttot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [19] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, 2018.
- [20] Q.-s. Zhang and S.-C. Zhu, “Visual interpretability for deep learning: a survey,” *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.
- [21] B. Chandrasekaran, M. C. Tanner, and J. R. Josephson, “Explaining control strategies in problem solving,” *IEEE Intelligent Systems*, no. 1, pp. 9–15, 1989.
- [22] W. R. Swartout and J. D. Moore, “Explanation in second generation expert systems,” in *Second generation expert systems*. Springer, 1993, pp. 543–585.
- [23] W. R. Swartout, “Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: Bg buchanan and eh shortliffe,(addison-wesley, reading, ma, 1984); 702 pages,” 1985.
- [24] A. Dhurandhar, V. Iyengar, R. Luss, and K. Shanmugam, “Tip: Typifying the interpretability of procedures,” *arXiv preprint arXiv:1706.02952*, 2017.
- [25] S. A. Friedler, C. D. Roy, C. Scheidegger, and D. Slack, “Assessing the local interpretability of machine learning models,” *arXiv preprint arXiv:1902.03501*, 2019.

- [26] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, no. 1, pp. 141–154, 2011.
- [27] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," *arXiv preprint arXiv:1802.07810*, 2018.
- [28] Q. Zhou, F. Liao, C. Mou, and P. Wang, "Measuring interpretability for different types of machine learning models," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 295–308.
- [29] "Single family loan level dataset - freddie mac." [Online]. Available: http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page
- [30] "Iml-cran package." [Online]. Available: <https://cran.r-project.org/web/packages/iml/index.html>
- [31] C. Molnar et al., "Interpretable machine learning: A guide for making black box models explainable," E-book at <https://christophm.github.io/interpretable-ml-book/>, version dated, vol. 10, 2018.
- [32] J. H. Friedman, B. E. Popescu et al., "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.
- [33] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
- [34] A. Goldstein, A. Kapelner, J. Bleich, and M. A. Kapelner, "Package 'icebox'," 2017.
- [35] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] A. Fisher, C. Rudin, and F. Dominici, "Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective," *arXiv preprint arXiv:1801.01489*, 2018.
- [37] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.
- [38] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.
- [39] S. Lundberg and S.-I. Lee, "An unexpected unity among methods for interpreting model predictions," *arXiv preprint arXiv:1611.07478*, 2016.
- [40] B. B. . B. Greenwell, "Chapter 16 interpretable machine learning — hands-on machine learning with r," <https://bradleyboehmke.github.io/HOML/iml.html>, (Accessed on 11/28/2019).
- [41] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 1, pp. 18–33, 2020.
- [42] A. Hartl, M. Bachl, J. Fabini, and T. Zseby, "Explainability and adversarial robustness for rnns," *arXiv preprint arXiv:1912.09855*, 2019.
- [43] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 3237–3243.
- [44] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems*, 2016, pp. 2280–2288.
- [45] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," *arXiv preprint arXiv:1802.07814*, 2018.
- [46] A. Jung and P. H. J. Nardelli, "An information-theoretic approach to personalized explainable machine learning," *IEEE Signal Processing Letters*, 2020.
- [47] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.
- [48] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.
- [49] U. Kursuncu, M. Gaur, and A. Sheth, "Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning," *arXiv preprint arXiv:1912.00512*, 2019.
- [50] S. R. Islam, W. Eberle, S. Bundy, and S. K. Ghafoor, "Infusing domain knowledge in ai-based "black box" models for better explainability with application in bankruptcy prediction," *ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2019, Anomaly Detection in Finance Workshop*, 2019.
- [51] S. R. Islam, W. Eberle, S. K. Ghafoor, A. Siraj, and M. Rogers, "Domain knowledge aided explainable artificial intelligence for intrusion detection and response," *arXiv preprint arXiv:1911.09853*, 2019.
- [52] E. Angelini, G. di Tollo, and A. Roli, "A neural network approach for credit risk evaluation," *The quarterly review of economics and finance*, vol. 48, no. 4, pp. 733–755, 2008.
- [53] J. Segal, "Five cs of credit." [Online]. Available: <https://www.investopedia.com/terms/f/five-c-credit.asp>
- [54] B. Matt et al., *Introduction to computer security*. Pearson Education India, 2006.
- [55] J. Fürnkranz, D. Gamberger, and N. Lavrač, "Rule learning in a nutshell," in *Foundations of Rule Learning*. Springer, 2012, pp. 19–55.
- [56] H. Yang, C. Rudin, and M. Seltzer, "Scalable bayesian rule lists," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR.org, 2017, pp. 3921–3930.
- [57] S. Rüping et al., "Learning interpretable models," 2006.
- [58] S. R. Islam, W. Eberle, and S. K. Ghafoor, "Towards quantification of explainability in explainable artificial intelligence methods," *arXiv preprint arXiv:1911.10104*, 2019.
- [59] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.
- [60] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.
- [61] P. Gandhi, "Explainable artificial intelligence." [Online]. Available: <https://www.kdnuggets.com/2019/01/explainable-ai.html>