

Literature review report

Knowledge discovery with fuzzy rule-based systems

FIFATA project

Raquel Fernandez-Peralta^a

^aMathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 814 73 Bratislava, Slovakia

January, 2026

1 Introduction

Data mining, or knowledge discovery in databases (KDD), is defined as the automatic extraction of patterns representing knowledge implicitly stored or captured in data [26]. Up to now, a wide variety of data mining techniques have been introduced and developed [40]. These techniques are usually divided into two types: exploratory and predictive. Exploratory data analysis focuses on searching relations between objects of a dataset (clustering, association rules, linguistic summaries...), whereas predictive data analysis aims to extract knowledge from discovered data with the aim of predicting or classifying unseen examples (classification, regression...). Within this context, rule based algorithms have been one of the top choices for knowledge extraction because of their human understandable output. Furthermore, with the emergence of the research area of explainable artificial intelligence (XAI) [8], they have become even more popular since they are highly interpretable, which is an appealing solution to provide an easily understandable representation of complex black box models [63, 31].

This report reviews the literature underlying six lines of work in fuzzy rule based knowledge discovery within the FIFATA project, one per section. At the end of each section we point out the open problem identified. Section 2 reviews the meaning given to a rule across different knowledge discovery tasks and the alternatives proposed to the usual co-occurrence reading. Section 3 reviews the optimization of rule based classifiers, fuzzy and crisp. Section 4 reviews tree induction for rule based models. Section 5 reviews the comparison between crisp and fuzzy rule representations and their interpretability. Section 6 reviews independence testing and causal discovery based on fuzzy relations. Section 7 reviews dynamic feature selection and the predictive uncertainty it raises. We assume the reader is familiar with basic concepts of fuzzy sets, fuzzy linguistic variables and fuzzy operators, we refer the reader to the following references for more information [51, 5, 52, 4, 110].

2 Rule semantics: co-occurrence and implication

In rule based models, the output is given in terms of rules which are usually represented as $A \Rightarrow C$, in which A is called the antecedent and C is called the consequent. Although in these techniques the output has the same representation, depending on the knowledge to be captured we can find several rule mining techniques, just to mention some of them: (i) association rules, which aim to capture situations in which if A is satisfied, then C is likely to occur also [1]; (ii) classification rules,

which aim to obtain a set of rules that effectively classify items into predefined classes [68]; (iii) sequential rules, which search for rules indicating that if some event(s) occur, some other event(s) are likely to occur [24]; (iv) subgroup discovery, which aims to obtain rules with the most unusual statistical characteristics with respect to a property of interest [3].

Despite the representation of the rules as logical conditionals, they are usually interpreted and evaluated as the co-occurrence of antecedent and consequent. Although this assumption is quite conventional in the literature, it has also received some criticism [14, 43, 44]. Indeed, the discrepancy between the linguistic representation of the rules and their underlying mathematical modeling can be confusing for the expert. For instance, if one considers the rule $\text{Smoking} \Rightarrow \text{Respiratory Problems}$, it may seem unnatural that the support of this rule is unaffected by its direction, because there are many respiratory problems that are not associated with smoking. An alternative to overcome this drawback is to consider a more logic oriented approach in which the conditional in the rule is interpreted as a logical implication, in this case normally the term “implicative rules” is used. In the crisp setting, the authors in [14] proposed to interpret the logical conditional through a measure called “conviction” as an alternative to confidence, which is based on conditional probability. However, in the fuzzy setting, fuzzy implications can be directly used. Indeed, in [43] the author introduced implication based fuzzy association rules by using a t-norm for modeling conjunction, a fuzzy implication to model the logical conditional, and the generalized modus ponens as inference schema. However, no further development of this framework or a thorough theoretical study of the proposal was made.

Directionality is also a fundamental property of all major causality frameworks, as causal relations are inherently asymmetric and non-exchangeable. Recent works combine fuzzy rule mining with causal discovery techniques [58, 107, 106]; however, causality is typically introduced as an external step rather than being formalized within the fuzzy rule model itself, and the directional semantics of fuzzy implications are not explicitly exploited. More generally, the relevance of directionality has also been discussed for other classes of operators, where asymmetric modeling has been shown to provide additional explanatory power in applied settings, for instance, in the case of copulas [72, 88].

Furthermore, it is well known that the theoretical development of fuzzy operators has experienced significant growth in the last two decades. Indeed, the introduction of new fuzzy operators and the study of their properties is nowadays so vast that there exist many different families among which one can choose. For instance, only recently has it been documented that more than 100 families of fuzzy implications have been introduced in the literature [64]. This fact, although clearly positive as it indicates great flexibility, also raises the inherent problem of determining which operator is most suitable for each application. In this sense, Zimmermann defines different criteria for an adequate selection of the involved aggregation operators in a specific model [111]. However, authors like Mendel point out that these criteria are rather subjective and may not be successfully implemented in engineering applications [67]. Thus, the adequate selection of the set of operators remains an important topic of discussion today.

Open problem. The fuzzy setting lacks a systematic study of which properties the pair of conjunction and implication operators used to model a rule should satisfy for its quality measures to behave coherently, and the directional semantics that fuzzy implications provide have not been exploited to give rules a logic based, rather than purely co-occurrence based, meaning.

3 Optimization of fuzzy rule based classifiers

Deep neural networks (DNNs) have been used to solve complex problems in machine learning where unstructured data are available in large volumes, such as image and video [56]. However, the use of these models is not always possible in cases where human liability is still relevant in the decision making process, like medicine and finance [8]. Rule based algorithms are considered one of the most trustworthy for users, as they explicitly show the patterns found and their relevance in each decision. They are also considered more faithful than other post hoc explainable artificial intelligence (XAI) methods, which are not always reliable [69, 80, 87].

By studying the rules themselves, practitioners can find additional clues about a problem and they can also disregard patterns found by the classifier that are contrary to existing knowledge [57]. It is also a popular use case for rule based inference to use the rule base as a proxy for a more complicated model, such as a deep learning model, to perform post hoc explanations [105, 57, 89]. One of the main research topics in rule based classification is the trade off between interpretability and accuracy, as the larger the number of rules, the less interpretable the model becomes. For example, ensembles of tree based models usually perform very well, at the cost of losing the interpretation capabilities that a single tree has [12, 29]. Bayesian reasoning has also been used, but it usually requires Markov chain models to train them, which is very time consuming [93].

Fuzzy logic has also been considered as a tool to balance performance and rule interpretability. For example, genetic optimization has been used to balance accuracy and complexity, but it suffers from significant scaling problems [2]. It is also possible to use fuzzy rules with gradient based optimization using the popular Takagi Sugeno Kang (TSK) architecture, but this creates fuzzy sets with no intrinsic meaning, and the consequents are linear models instead of classes, which hinders interpretability [109]. Pure gradient based approaches have also been used to obtain rules in classification, regression [94, 108, 98, 74, 90] and subgroup discovery [96]. However, they can result in an arbitrarily high number of additive rules, which can only be controlled by a dense search of hyperparameters. In addition, how these methods partition the state can also be hard for a user to understand, especially when the number of rules is high, as the cut points can be arbitrary and very numerous. This limits the interpretability of the model concerning a human user, which is instrumental in its deployment in real world systems [19].

Rule based learning has also been mostly applied to tabular data, where the symbolic pieces of information that are used to form the rule antecedents can be easily understood by human users [60]. In order to use rule reasoning on non structured data, it is possible to use concept bottlenecking, which consists of using a deep learning network architecture to detect a set of concepts which are then used to make the final decision [53]. However, concept bottlenecking faces several limitations in practical applications. First, the quality of concept detection remains difficult to evaluate objectively, as there is no ground truth for many high level concepts [7, 32]. Second, the interpretability of concept bottlenecks depends critically on the quality of the intermediate concepts, but poor concept detection can lead to misleading explanations that appear plausible but are factually incorrect [92]. Because of this, rule based models are still applied mostly to tabular data [69].

Classical rule based models are typically trained using heuristic or greedy algorithms due to their discrete and non differentiable nature [95]. While these methods are inherently interpretable, they often suffer from suboptimal convergence [81] and scalability limitations, particularly in high dimensional spaces [97]. Modern adaptations attempt to mitigate these issues by extracting rules post hoc from ensemble models like gradient boosting or random forests [30] or by mining frequent patterns [101]. However, these approaches still struggle with large scale data and often sacrifice interpretability in the process.

Recent work has combined rule based methods with deep neural networks to leverage the strengths of both approaches. For example, concept bottleneck networks extract relevant concepts while symbolic rules perform reasoning. Some studies fix the logic structure and focus on identifying the right concepts [53, 7], while others optimize the connections among predefined concepts [92]. However, the interpretability of such systems depends heavily on the quality of concept detection, which is difficult to evaluate. It is also possible to distill black box models into rule based ones with good results [57], at the expense of having to train both models.

Due to the scalability of gradient based optimization, several notable approaches have been proposed to extract rules from neural network classifiers. These methods often rely on the popular straight through estimator (STE) [10], which is a method to approximate the derivative of non differentiable functions. For instance, in [74] the authors propose a neural network architecture that uses binarization for real valued variables and encodes the rules themselves in the neurons. Similarly, [108] learns “soft intervals” for binarization, regularized with an additional loss to avoid excessive partitioning of the feature space. However, these models often generate a large number of rules unless trained extensively (e.g., over 1,000 epochs), compromising interpretability. This is also the problem of logic gate based networks [9, 102]: while they are good for automated formal verification of inference, their complexity makes it difficult for humans to interpret the results effectively.

Another gradient based approach, gradient grafting [94], combines discrete and continuous models: the continuous model guides gradient flow, while the discrete model produces layer wise outputs. However, this approach tends to create large numbers of additive rules, so the reasoning mechanism is very difficult for a human being to understand.

For the case of fuzzy rule based inference, it is possible to optimize fuzzy sets using gradient descent [23]. However, traditional neuro fuzzy TSK models suffer from two key limitations when trained via gradient optimization: exponential rule growth with input dimensionality [68], and interpretability loss in hierarchical architectures and their fuzzy set optimization process [109].

Open problem. Existing gradient based rule learners either bound rule complexity only through soft, implicit penalties or lose the linguistic meaning of their conditions during optimization, so a rule learner that enforces a hard, architectural bound on the number of rules and conditions while keeping fuzzy, linguistically labeled antecedents remains largely unexplored.

4 Tree induction for crisp and fuzzy rule models

Fuzzy rule based systems have been popular for explainable artificial intelligence (XAI) as they rely on logic reasoning that humans easily understand [80]. However, explainability is limited by the complexity of the system. For example, a system composed of hundreds of rules is much harder to understand than one with only a handful [34]. Therefore, besides the use of highly performant classifiers such as FURIA [42], there has also been particular research interest in a subset of fuzzy rule based classifiers (FRBCs) that constrain the resulting rule base according to the end user’s desired complexity [32].

Two main approaches have been proposed to address this challenge. Architecturally constrained methods, such as adaptive network based fuzzy inference system (ANFIS) [47], use neural architectures where the number of rules and conditions is pre-specified. Evolutionary optimization approaches can incorporate complexity constraints into the optimization process [2], just as gradient based methods enforce sparsity through loss functions [108]. However, this approach has limitations: constraints are often implemented as soft preferences that may still result in complex models, and even when successful, the computational cost typically exceeds that of standard classifiers, discouraging practitioners from using FRBCs.

The most popular method for performing crisp rule based classification has been to use tree structures with Gini conditioned splits. This has provided a computationally efficient and accurate way to perform rule learning, establishing CART [13] as a widely adopted classification algorithm. However, CART has significant limitations from an interpretability perspective. First, the number of leaf nodes can be quite high, making the overall decision structure difficult to understand. Second, CART uses hard splits that create arbitrary boundaries (e.g., “age < 45.7”), which do not reflect the gradual transitions often present in human reasoning. Finally, CART provides no mechanism for incorporating existing expert knowledge into the learning process, potentially producing rules that conflict with established domain knowledge or are difficult for practitioners to trust. This leaves room for opportunity for fuzzy systems, where conditions can be defined to take into account the intuitions of practitioners through meaningful linguistic partitions [104], and can potentially recover complex decision surfaces using fewer rules. However, the task of finding optimal fuzzy partitions while keeping interpretability is not trivial. Fuzzy systems that rely on predefined fuzzy partitions are computationally efficient, but this strategy may produce suboptimal decision boundaries when the class distribution does not align well with these pre-existing semantics.

Rule based learning methods need to balance model complexity and predictive accuracy [80]. As the number of rules and conditions increases, the model can capture more specific patterns in the data, typically leading to improved classification performance. However, this comes at the cost of interpretability: a rule base with hundreds of rules becomes essentially a black box, defeating the primary purpose of using rule based systems for explainable AI [79]. This trade off is particularly critical in domains such as healthcare [20, 25], finance, and legal decision making [60], where stakeholders must not only trust the model’s predictions but also understand the reasoning behind them. Consequently, finding the optimal balance between accuracy and complexity remains a central challenge in rule based machine learning [69].

Classical crisp tree based methods include CART [13] and C4.5 [75]. More recent crisp approaches leverage gradient based learning to improve scalability. In parallel, Bayesian crisp rule learning methods have been proposed to control model complexity by minimizing the size of the induced rule base while providing theoretical guarantees [93].

For fuzzy systems, gradient based approaches have likewise been widely adopted. ANFIS [47] is perhaps the most well known example, combining neural network learning capabilities with fuzzy reasoning. ANFIS employs a hybrid learning algorithm that uses backpropagation to tune membership function parameters in the premise part and least squares estimation for the consequent parameters. However, this fine tuning comes at the cost of interpretability: the learned membership functions often lose their linguistic meaning, becoming overlapping or irregular in shape. Evolutionary approaches can simultaneously optimize rule structure, membership function parameters, and the number of rules. However, the computational cost remains prohibitive for many applications, as evolutionary algorithms typically require thousands of fitness evaluations.

Open problem. Fuzzy tree and rule induction methods that reach competitive accuracy rely on evolutionary search over rule structure and partitions, which is accurate but computationally expensive, while faster, greedy tree induction has mostly been developed for crisp splits. A greedy counterpart that induces fuzzy trees over linguistic partitions at comparable computational cost is comparatively unexplored.

5 Crisp and fuzzy rule complexity and interpretability

There is a growing need for transparency in modern machine learning systems, as they are deployed in many real life scenarios where human and legal liabilities are present [15, 8]. One of the most pop-

ular methods for explainable classification are rule based and decision tree systems, which explicitly match patterns with outcomes [13], and this helps the user trust the decisions of the system.

Fuzzy rules have been one of the most popular methods in the fuzzy literature, and research in this area remains active [32, 82, 27]. Its principal lines of research are devoted to finding the best rules possible and to obtaining better performance and explainability trade offs [68, 33]. Nonetheless, there exists a historical and ongoing reluctance surrounding the adoption of fuzzy logic techniques outside the fuzzy logic community [45]. Part of this mistrust relies in the fact that non fuzzy users are not familiar with the theory of fuzzy logic, and they may prefer crisp sets or methods purely based on probability theory and statistics. The interpretability of linguistic variables can also be debated, especially when the number of them is high or when the users are not particularly familiar with the vocabulary used [34].

In rule based classifiers, the decision function is constructed in terms of a collection of if-then rules called the rule base [104]. In general, the problem of multiclass classification consists of finding a function f such that $f(X_i) = Y_i$ for as many instances as possible [85]. In this case, f is called a decision function and it divides the input space into regions, one per class, such that each point in a certain region is assigned the corresponding class. For comparing different classifications, there exists plenty of metrics in the literature [37]. Various structures of membership functions have been considered (triangular, trapezoidal, Gaussian...); the selection of these functions depends on the specific problem and/or application [52].

Open problem. Fuzzy and crisp rule based classifiers are usually compared through predictive accuracy or raw rule counts, without a common formal description of the decision regions each one induces. It remains unclear how to compare their complexity, or translate one representation into the other, on equal terms.

6 Independence and causal discovery with fuzzy relations

Causality has also been introduced directly into fuzzy rule based systems. Recent works combine fuzzy rule mining with causal discovery techniques [58, 107, 106]; however, causality is typically introduced as an external step rather than being formalized within the fuzzy rule model itself, and the directional semantics of the fuzzy operators involved are not explicitly exploited to support it.

Independence testing is central to many statistical and machine learning tasks, including causal discovery, where it enables algorithms such as PC to identify dependencies and remove spurious edges in causal graphs [84]. Existing work combining fuzzy data with causal methods typically either applies classical probabilistic techniques to fuzzy valued data or extends causal estimands (e.g., treatment effects) to fuzzy settings. However, a formal notion of independence for data without an assumed probability distribution remains largely unexplored.

Fuzzy entropy has been defined for fuzzy T -equivalence relations [100], where each relation represents similarity between instances with respect to a feature. Entropy is defined analogously to Shannon entropy by replacing probabilities with normalized fuzzy cardinalities of equivalence classes, i.e., the total membership degree of all instances to the class of a given instance divided by the number of instances. However, this formulation is restricted to finite universes and does not ensure subadditivity, which may yield negative fuzzy mutual information and limits its interpretation as a dependence measure between fuzzy relations. Established mutual information based independence testing procedures [54] give a natural point of comparison for extending this notion of dependence to the fuzzy setting.

Open problem. A formal notion of independence for data without an assumed probability distribution remains largely unexplored, and the existing definition of fuzzy entropy on fuzzy T -

equivalence relations does not guarantee subadditivity, so the fuzzy mutual information derived from it cannot be relied upon as a dependence measure for conditional independence testing.

7 Dynamic feature selection and predictive uncertainty

Many prediction problems are constrained not by model capacity but by the cost of acquiring new information. In clinical diagnosis, industrial inspection, or interactive data collection, every additional laboratory test, sensor reading, or user query has a monetary, temporal, and human cost. The relevant variables are often instance specific: two patients may require different tests, and two images may require different regions to be inspected. Dynamic feature selection (DFS) addresses this setting by sequentially acquiring features for each instance under a budget [66].

The existing literature on DFS primarily explores two main approaches: reinforcement learning (RL) [48] and greedy optimization [18]. RL based approaches offer the potential to discover optimal policies for feature acquisition, but they come with the inherent challenges of RL training, including significant computational demands and potential convergence issues [25]. Greedy optimization methods maximize a measure that relates the known inputs with the response variable [16]. Usually, this measure is the conditional mutual information (CMI) between input variables and the response variable, which has some desirable qualities in terms of uncertainty minimization and prediction power [35], although it has non differentiability issues [22]. These approaches can focus on the CMI estimation task directly [35], circumventing the issue of selecting a discrete variable in a neural network, or by using generative models [62]. Nevertheless, both approaches rely on predictors that operate across a large number of feature subsets. This requirement is closely related to the core problem studied in invariant risk minimization [55], where a single model must generalize across multiple environments. These subset adaptive methods allow specialization while retaining shared knowledge across subsets. Examples of this approach are hierarchical or modular parametrizations [78], mixture of experts architectures [83], or meta-learning strategies [28].

Uncertainty quantification (UQ) measures how much a model’s prediction should be trusted. Following the taxonomy given in [46], predictive uncertainty is commonly decomposed into aleatoric and epistemic components. Aleatoric uncertainty refers to irreducible ambiguity in the data generating process, such as class overlap or measurement noise. Epistemic uncertainty refers to uncertainty caused by limited knowledge about the predictor, its parameters, or the region of the input space being evaluated; in principle, it can be reduced by better data, models, or adaptation.

The dominant view of DFS frames each acquisition as the feature that maximally reduces uncertainty about the label, formalized through the conditional mutual information [66, 35]. Because CMI is intractable for arbitrary subsets, recent work learns to amortize it. Covert et al. [22] train a predictor network jointly with a policy network under random feature masking, so that a single greedy estimator approximates the CMI maximizing feature at every step; Chattopadhyay et al. [16] learn query chains that maximize the information gain. More recent work has criticized the use of CMI for DFS [73], and has opted for using an ad hoc metric instead. A second family models the unobserved features explicitly. EDDI [62] fits a partial variational autoencoder to the conditional distribution of the unobserved features and acquires the feature with the largest expected information reward, while Li and Oliva [59] pair a generative surrogate with model based RL to amortize the same lookahead. Model free RL formulations such as [48] and [99] instead learn an acquisition and stop policy end to end by treating cost as a negative reward. Earlier cost sensitive formulations framed acquisition as budgeted classification [50, 49] or as maximizing the value of information at each step [17], and the idea has been extended to images through probabilistic hard attention policies that acquire informative glimpses [77].

Classical feature selection chooses a single global subset for all instances through filter, wrapper, or embedded criteria [39], including differentiable relaxations such as concrete autoencoders [6]. Such methods do not adapt the subset per instance, but provide strong, low cost baselines that DFS must beat to justify its overhead. A separate motivation for keeping a single fixed predictor is interpretability: rule and tree based models are valued because their decision logic is inspectable [79, 69], and pooling them into ensembles for uncertainty estimation would obscure that logic, complementing concept based routes to interpretable acquisition [53].

The aleatoric and epistemic decomposition [46, 91] splits the predictive uncertainty of a model into an irreducible part, originating from noise and class overlap in the data, and a reducible part, originating from the model’s limited knowledge of the parameters or the input region being evaluated. In practice, total uncertainty is computed as the entropy of the prediction, while the epistemic component is estimated from the disagreement among several predictors obtained through Monte Carlo dropout or deep ensembles. Other methods for uncertainty quantification include post hoc calibration, such as temperature scaling, and deterministic single pass estimators [36, 76, 38, 71, 61, 70]. These estimators are almost always studied for a fixed input, where the aleatoric uncertainty is constant and only the epistemic term is expected to reduce with more data [41]. DFS violates that premise: the conditioning set changes at every acquisition step, so both components become subset dependent, and a predictor that is well calibrated on the full feature set need not remain calibrated on partial subsets. This subset dependence of both uncertainty components during sequential acquisition is, to date, not addressed in the uncertainty quantification or DFS literature.

Open problem. Existing dynamic feature selection methods evaluate acquisition mainly through accuracy-budget curves, without separating aleatoric and epistemic uncertainty across acquisition steps, and require predictors trained specifically for arbitrary feature subsets. It remains open whether an already validated, fixed predictor (a rule-based model as a particular case) can be reused for sequential acquisition under this subset dependent uncertainty.

Acknowledgements

Raquel Fernandez-Peralta was funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00557.

References

- [1] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. American Association for Artificial Intelligence, 1996.
- [2] Jesús Alcalá-Fdez, Rafael Alcalá, and Francisco Herrera. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. *IEEE Transactions on Fuzzy Systems*, 19(5):857–872, 2011.
- [3] Martin Atzmueller. Subgroup discovery. *WIREs Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.
- [4] Michał Baczyński, Przemysław Grzegorzewski, Piotr Helbin, and Wanda Niemyska. Properties of the probabilistic implications and s-implications. *Information Sciences*, 331:2–14, 2016.
- [5] Michał Baczyński and Balasubramaniam Jayaram. *Fuzzy Implications*, volume 231 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, 2008.

- [6] Muhammed Fatih Balin, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *Proceedings of the 36th International Conference on Machine Learning*, pages 444–453, 2019.
- [7] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lió, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1801–1825, 2023.
- [8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.
- [9] Adrien Benamira, Thomas Peyrin, Trevor Yap, Tristan Guerand, and Bryan Hooi. Truth table net: Scalable, compact and verifiable neural networks with a dual convolutional small Boolean circuit networks form. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 13–21, 2024.
- [10] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- [11] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.
- [12] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [13] Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [14] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 255–264, 1997.
- [15] Ruth M. J. Byrne. Good explanations in explainable artificial intelligence (XAI): Evidence from human explanatory reasoning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6536–6544, 2023.
- [16] Aditya Chattopadhyay, Kwan Ho Ryan Chan, Benjamin David Haeffele, Donald Geman, and Rene Vidal. Variational information pursuit for interpretable predictions. In *The Eleventh International Conference on Learning Representations*, 2023.
- [17] Suming Chen, Arthur Choi, and Adnan Darwiche. Value of information based on decision robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- [18] Yuxin Chen, S. Hamed Hassani, Amin Karbasi, and Andreas Krause. Sequential information maximization: When is greedy near-optimal? In *Conference on Learning Theory*, pages 338–363, 2015.
- [19] Zhisong Chen and Ching Y. Suen. Complexity metrics for rule-based expert systems. In *Proceedings of the International Conference on Software Maintenance*, pages 382–391. IEEE, 1994.
- [20] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G. Lee, Vikram Deshpande, Joseph Schwab, Michael H. Lev, Ramon G. Gonzalez, Michael S. Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, 7(6):711–718, 2023.
- [21] William W. Cohen. Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123, 1995.
- [22] Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J. White, and Su-In Lee. Learning to maximize mutual information for dynamic feature selection. In *Proceedings of the 40th International Conference on Machine Learning*, pages 6424–6447, 2023.

- [23] Yuqi Cui, Dongrui Wu, and Jian Huang. Optimize TSK fuzzy systems for classification problems: Mini-batch gradient descent with uniform regularization and batch normalization. *IEEE Transactions on Fuzzy Systems*, 28(12):3065–3075, 2020.
- [24] Gautam Das, King-IP Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth. Rule discovery from time series. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 16–22. AAAI Press, 1998.
- [25] Gabriel Erion, Joseph D. Janizek, Carly Hudelson, Richard B. Utarnachitt, Andrew M. McCoy, Michael R. Sayre, Nathan J. White, and Su-In Lee. A cost-aware framework for the development of AI models for healthcare applications. *Nature Biomedical Engineering*, 6(12):1384–1398, 2022.
- [26] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 82–88. AAAI Press, 1996.
- [27] Raquel Fernandez-Peralta, Sebastia Massanet, Megha Gupta, Kavitha Nanavati, and Balasubramaniam Jayaram. Subgroup discovery through sharp transitions using implicative type rules. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6. IEEE, 2023.
- [28] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.
- [29] Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.
- [30] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.
- [31] Javier Fumanal-Idocin and Javier Andreu-Perez. Ex-Fuzzy: a library for symbolic explainable AI through fuzzy logic programming. *Neurocomputing*, 599:128048, 2024.
- [32] Javier Fumanal-Idocin, Javier Andreu-Perez, Oscar Cordón, Hani Hagra, and Humberto Bustince. ARTxAI: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *IEEE Transactions on Fuzzy Systems*, 32(4):1915–1926, 2024.
- [33] Javier Fumanal-Idocin, Mohammadreza Jamalifard, and Javier Andreu-Perez. Interpreting contrastive embeddings in specific domains with fuzzy rules. In *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2024.
- [34] María José Gacto, Rafael Alcalá, and Francisco Herrera. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360, 2011.
- [35] Soham Gadgil, Ian Connick Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *The Twelfth International Conference on Learning Representations*, 2024.
- [36] Yarín Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.
- [37] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: An overview. *arXiv preprint arXiv:2008.05756*, 2020.
- [38] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.
- [39] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications*, volume 207. Springer, 2006.
- [40] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier, 3rd edition, 2011.

- [41] Arthur Hoarau, Benjamin Quost, Sébastien Destercke, and Willem Waegeman. Reducing aleatoric and epistemic uncertainty through multi-modal data acquisition. *Machine Learning*, 115(5):121, 2026.
- [42] Jens Hühn and Eyke Hüllermeier. FURIA: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, 19:293–319, 2009.
- [43] Eyke Hüllermeier. Implication-based fuzzy association rules. In *Principles of Data Mining and Knowledge Discovery*, volume 2168 of *Lecture Notes in Computer Science*, pages 241–252. Springer Berlin Heidelberg, 2001.
- [44] Eyke Hüllermeier. Fuzzy rules in data mining: From fuzzy associations to gradual dependencies. In *Combining Experimentation and Theory: A Hommage to Abe Mamdani*, pages 123–135. Springer Berlin Heidelberg, 2012.
- [45] Eyke Hüllermeier. Does machine learning need fuzzy logic? *Fuzzy Sets and Systems*, 281:292–299, 2015.
- [46] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.
- [47] Jyh-Shing Roger Jang. ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3):665–685, 1993.
- [48] Jaromír Janisch, Tomáš Pevný, and Viliam Lisý. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3959–3966, 2019.
- [49] Mohammad Kachuee, Sajad Darabi, Babak Moatamed, and Majid Sarrafzadeh. Dynamic feature acquisition using denoising autoencoders. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2252–2262, 2019.
- [50] Sergey Karayev, Mario J. Fritz, and Trevor Darrell. Dynamic feature selection for classification on a budget. In *International Conference on Machine Learning: Workshop on Prediction with Sequential Models*, volume 95, 2013.
- [51] Erich Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*, volume 8 of *Trends in Logic*. Kluwer Academic Publishers, 2000.
- [52] George J. Klir and Bo Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, 1995.
- [53] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348, 2020.
- [54] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.
- [55] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 2021.
- [56] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [57] Ruoyu Li, Qing Li, Yu Zhang, Dan Zhao, Yong Jiang, and Yong Yang. Interpreting unsupervised anomaly detection in security via rule extraction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [58] Wenrui Li, Wei Zhang, Qinghao Zhang, Xuegong Zhang, and Xiaowo Wang. Weakly supervised causal discovery based on fuzzy knowledge and complex data complementarity. *IEEE Transactions on Fuzzy Systems*, 32(12):7002–7014, 2024.

- [59] Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6450–6459. PMLR, 2021.
- [60] Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
- [61] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512, 2020.
- [62] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4234–4243, 2019.
- [63] Dawid Macha, Michał Kozielski, Łukasz Wróbel, and Marek Sikora. RuleXAI—a package for rule-based explanations of machine learning model. *SoftwareX*, 20:101209, 2022.
- [64] Sebastia Massanet, Raquel Fernandez-Peralta, Michał Baczynski, and Balasubramaniam Jayaram. On valuable and troubling practices in the research on classes of fuzzy implication functions. *Fuzzy Sets and Systems*, 476:108786, 2024.
- [65] Hayden McTavish, Chudi Zhong, Reto Achermann, Ilias Karimalis, Jacques Chen, Cynthia Rudin, and Margo Seltzer. Fast sparse decision tree optimization via reference ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9604–9613, 2022.
- [66] Prem Melville, Maytal Saar-Tsechansky, Foster Provost, and Raymond Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM)*, pages 483–486. IEEE, 2004.
- [67] Jerry M. Mendel. Fuzzy logic systems for engineering: A tutorial. *Proceedings of the IEEE*, 83(3):345–377, 1995.
- [68] Jerry M. Mendel. *Explainable Uncertain Rule-Based Fuzzy Systems*. Springer, 2024.
- [69] Christoph Molnar. *Interpretable Machine Learning*. Lulu, 2020.
- [70] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.
- [71] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- [72] Roger B. Nelsen. Extremes of nonexchangeability. *Statistical Papers*, 48:329–336, 2007.
- [73] Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela van der Schaar, and Pietro Lió. Stochastic encodings for active feature acquisition. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 46784–46814, 2025.
- [74] Litao Qiao, Weijia Wang, and Bill Lin. Learning accurate and interpretable decision rule sets from neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4303–4311, 2021.
- [75] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [76] Rahul Rahaman and Alexandre H. Thiery. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.
- [77] Samrudhdi B. Rangrej and James J. Clark. A probabilistic hard attention model for sequentially observed scenes. *arXiv preprint arXiv:2111.07534*, 2021.

- [78] Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8119–8127, 2018.
- [79] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [80] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.
- [81] Günter Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5(1):96–101, 1994.
- [82] Armin Salimi-Badr and Mohammad Mehdi Ebadzadeh. A novel learning algorithm based on computing the rules’ desired outputs of a TSK fuzzy neural network with non-separable fuzzy rules. *Neurocomputing*, 470:139–153, 2022.
- [83] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.
- [84] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.
- [85] Pang-Ning Tan, Anuj Karpatne, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson, global edition, 2019.
- [86] Roman Timofeev. Classification and regression trees (CART) theory and applications. Master’s thesis, Humboldt University, Berlin, 2004.
- [87] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6021–6029, 2020.
- [88] Raul Torres, Rosa E. Lillo, and Henry Laniado. A directional multivariate value at risk. *Insurance: Mathematics and Economics*, 65:111–123, 2015.
- [89] Ángela R. Troncoso-García, María Martínez-Ballesteros, Francisco Martínez-Álvarez, and Alicia Troncoso. A new metric based on association rules to assess feature-attribution explainability techniques for time series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4140–4155, 2025.
- [90] Haiming Tuo, Zuqiang Meng, Zihao Shi, and Daosheng Zhang. Interpretable neural network classification model using first-order logic rules. *Neurocomputing*, 614:128840, 2025.
- [91] Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022.
- [92] Deepika Vemuri, Gautham Bellamkonda, and Vineeth N. Balasubramanian. Enhancing concept-based learning with logic. In *International Conference on Machine Learning Workshop on Differentiable Almost Everything*, 2024.
- [93] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18(70):1–37, 2017.
- [94] Zhuo Wang, Wei Zhang, Ning Liu, and Jianyong Wang. Learning interpretable rules for scalable data representation and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1121–1133, 2024.

- [95] Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Günlük. Generalized linear rule models. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6687–6696, 2019.
- [96] Sascha Xu, Nils Philipp Walter, Janis Kalofolias, and Jilles Vreeken. Learning exceptional subgroups by end-to-end maximizing KL-divergence. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55267–55285, 2024.
- [97] Fan Yang, Kai He, Linxiao Yang, Hongxia Du, Jingbang Yang, Bo Yang, and Liang Sun. Learning interpretable decision rule sets: A submodular optimization approach. In *Advances in Neural Information Processing Systems*, volume 34, pages 27890–27902, 2021.
- [98] Yang Yang, Wendi Ren, and Shuang Li. HyperLogic: enhancing diversity and accuracy in rule learning with hypernets. In *Advances in Neural Information Processing Systems*, volume 37, pages 3564–3587, 2024.
- [99] Haiyan Yin, Yingzhen Li, Sinno Jialin Pan, Cheng Zhang, and Sebastian Tschiatschek. Reinforcement learning with efficient active feature acquisition. *arXiv preprint arXiv:2011.00825*, 2020.
- [100] Daren Yu, Shuang An, and Qinghua Hu. Fuzzy mutual information based min-redundancy and max-relevance heterogeneous feature selection. *International Journal of Computational Intelligence Systems*, 4(4):619–633, 2011.
- [101] Xiuli Yuan. An improved Apriori algorithm for mining association rules. In *AIP Conference Proceedings*, volume 1808, page 080005. AIP Publishing, 2017.
- [102] Chang Yue and Niraj K. Jha. Learning interpretable differentiable logic networks. *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, 1(1):69–82, 2024.
- [103] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.
- [104] Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8(3):199–249, 1975.
- [105] Quanshi Zhang, Yu Yang, Ying Nian Wu, and Song-Chun Zhu. Interpreting CNNs via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6254–6263, 2019.
- [106] Te Zhang, Christian Wagner, and Jonathan M. Garibaldi. Explain the world—using causality to facilitate better rules for fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 32(12):6671–6683, 2024.
- [107] Te Zhang, Jingda Ying, Christian Wagner, and Jonathan M. Garibaldi. Towards causal fuzzy system rules using causal direction. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6, 2023.
- [108] Wei Zhang, Yongxiang Liu, Zhuo Wang, and Jianyong Wang. Learning to binarize continuous features for neuro-rule networks. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4584–4592, 2023.
- [109] Yuanpeng Zhang, Guanjin Wang, Ta Zhou, Xiuyu Huang, Saikit Lam, Jiabao Sheng, Kup Sze Choi, Jing Cai, and Weiping Ding. Takagi-Sugeno-Kang fuzzy system fusion: A survey at hierarchical, wide and stacked levels. *Information Fusion*, 101:101977, 2024.
- [110] Hongjun Zhou. Characterizations of fuzzy implications generated by continuous multiplicative generators of t-norms. *IEEE Transactions on Fuzzy Systems*, 29(10):2988–3002, 2021.
- [111] Hans-Jürgen Zimmermann. *Fuzzy Set Theory and Its Applications*. Kluwer Academic Publishers, 1991.