Received: 23 September 2023 | Revised: 4 March 2024 | Accepted: 15 March 2024 | Published online: 29 March 2024

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 3(1) 72–82
DOI: 10.47852/bonviewAIA42021781

BON VIEW PUBLISHING

FormulAI: Designing Rule-Based Datasets for Interpretable and Challenging Machine Learning Tasks

Hegler Tissot[1,*]

[1]*Department of Information Science, Drexel University, United States*

Abstract: In an era marked by the transformative impact of machine learning (ML) algorithms across various disciplines, challenges in achieving model interpretability persist. Existing evaluation datasets often lack transparency, thereby obscuring the decision-making process of ML models, particularly in complex deep learning architectures. This opacity raises concerns across sectors like healthcare, emphasizing the pivotal role of explainability in fostering trust and adhering to non-supervisory norms. While progress has been made through the development of interpretable models, the absence of formalized, interpretable datasets hampers the validation and comparison of techniques. Rule-based datasets, distinct from general synthetic datasets, provide an avenue to simulate real-world challenges while maintaining interpretability. This paper introduces FormulAI, a framework for generating comprehensive rule-grounded datasets encompassing categorical and continuous features, calibrated noise, and imbalanced class distribution. Emphasizing scalability and reproducibility, these datasets serve as a robust standard, fostering exploration in interpretability and robustness.

Keywords: synthetic datasets, rule-based datasets, pattern recognition, interpretability and explainability, class imbalanced

1. Introduction

Machine learning (ML) algorithms have achieved impressive results in various fields, revolutionizing industries and solving complex problems. Despite these achievements, researchers and practitioners still face ongoing challenges, including model interpretability and dealing with imbalanced class distributions. Therefore, benchmark datasets used in ML development should be designed to address these challenges by incorporating well-annotated instances and realistic class imbalances.

Explainability is crucial for establishing trust in artificial intelligence (AI) systems and represents a regulatory requirement in critical areas such as healthcare [1]. Nevertheless, existing evaluation datasets often lack interpretability, posing challenges in understanding the decision-making process of resulting ML models. These models, particularly deep learning architectures, frequently function as black boxes, impeding insights into their predictive mechanisms. Efforts to create more interpretable models, including those leveraging attention mechanisms, have demonstrated promise. However, the absence of standardized, interpretable datasets limits our capacity to validate and compare various interpretability techniques [2, 3].

The presence of interpretable ground truth labels plays a crucial role in evaluating the interpretability of ML models [4]. Such labels serve as benchmarks for validating the explanations derived from these models, allowing comparison against expected model behavior. Interpretable ground truth labels can be generated based on domain-expert knowledge or synthetic rules.

Distinct from general synthetic datasets designed to simulate real-world data characteristics with controlled aspects like data distribution, noise levels, and feature interactions [5, 6], rule-based datasets constitute a specific type of synthetic data. These datasets generate data instances and class labels based on explicit rules, criteria, or formulas. Rule-based datasets offer a strategic solution for replicating imbalanced scenarios that mirror real-world challenges while upholding interpretability. This capability facilitates the development and enhancement of ML algorithms and techniques. The establishment and use of rule-based benchmarks play a pivotal role in comprehending and refining the decision-making processes of ML models, ultimately fostering more reliable AI systems. Rule-based datasets prove invaluable in challenging ML applications, serving as essential benchmarks to assess ML model performance under controlled conditions, including model interpretability and explainability [3], robustness [7], adversarial testing [8], incremental complexity [9], and domain-specific challenges [10].

Rule-based datasets can cover various scenarios, contributing to bolstering the robustness and generalization of AI models, thereby enhancing their performance on unseen data and expanding their capacity to handle a broader spectrum of inputs [11]. Additionally, diverse and representative samples can aid research endeavors focused on fairness and bias reduction [12]. Nevertheless, creating such benchmarks presents several challenges that can impact the dataset's quality, applicability, and representativeness. These challenges encompass factors such as data complexity [13], scalability [11], rule significance [14, 15], and the balancing noise and uncertainty [16].

*Corresponding author: Hegler Tissot, Department of Information Science, Drexel University, United States. Email: hegler.tissot@drexel.edu

© The Author(s) 2024. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

This paper introduces FormulAI as an extensive framework for generating rule-based datasets, primarily aimed at addressing the challenge of explainability in ML applications. The resulting datasets are formulated based on explicit rules that govern the relationships between input features and output labels, including a blend of categorical and continuous features to mirror the diverse data encountered in real-world applications. Deliberately unbalanced labels simulate scenarios where certain outcomes occur infrequently and yet prioritize model interpretability. To infuse realism, controlled noise is introduced to emulate the complexities found in the real world. Emphasizing scalability and reproducibility, the FormulAI datasets aim to serve as benchmarks for assessing ML model performance under challenging conditions. This initiative intends to foster research and development in key areas, including ML interpretability, managing imbalanced classes, and enhancing robustness. Furthermore, these datasets will enable the evaluation and comparison of various algorithms, assess the efficacy of imbalanced learning techniques, and facilitate the development of innovative approaches to enhance model prediction explanations.

2. Literature Review

Datasets constitute the fundamental building blocks upon which ML models are trained. As a primary source of information and context, datasets allow models to learn, generalize, and infer from the underlying patterns used to make informed predictions. The training process is similar to creating a cognitive map and its effectiveness is intrinsically linked to the quality of training data.

Recent advancements in deep reinforcement learning highlight the significant impact of training ML models on extensive datasets. This emphasizes the pivotal role of datasets in facilitating models to comprehend intricate patterns and complexities, empowering them to accomplish tasks previously deemed difficult or outside the realm of ML algorithms [17]. Nevertheless, contemporary accurate decision-support systems often operate as black boxes, concealing their internal logic from users. This absence of explanation poses both practical and ethical concerns [18].

While rich datasets enhance models' ability to generalize to unseen data, improving robustness and accuracy, dataset diversity embraces a broad spectrum of scenarios, variations, and edge cases, enhancing model generalization [19]. However, specific considerations in dataset design crucially impact model performance: (a) biases within datasets can be learned and perpetuated, leading to biased predictions and unfair outcomes [20]; (b) outliers, anomalies, and noisy data in a dataset can adversely affect model training and performance [21]; finally, (c) larger, more complex datasets significantly contribute to model performance by capturing intricate patterns often missed in smaller datasets, enhancing model scalability [22].

2.1. Dataset resources

The UCI (University of California, Irvine) Machine Learning Repository is a widely recognized benchmark resource in the ML community[1]. It offers freely available standardized datasets for comparing and evaluating the effectiveness of ML methods across various application domains, including healthcare, finance, and social sciences. The most frequently used UCI datasets are Iris [23], Wine [24], Breast Cancer Wisconsin [25], Boston Housing [26], among many others. However, some datasets in the repository might be older or possess simpler characteristics compared to real-world data.

[1]https://archive.ics.uci.edu

Kaggle is another well-known platform that promotes a dynamic environment for data science competitions, favoring the development of cutting-edge models to address real-world challenges. Kaggle also hosts large datasets, often taken from industry contexts, which provide useful and up-to-date insights for ML applications. These datasets serve as valuable benchmarks for testing novel algorithms and exploring advanced techniques, demonstrating their potential to solve complex problems facing modern industry.

In addition to UCI and Kaggle, other benchmark dataset resources that are used for evaluating ML algorithms include Physionet [27], ImageNet [28], OpenML [29], and SNAP [30].

2.2. Interpretability

Interpretable ML often involves understanding how a model's predictions are influenced by input features. To evaluate interpretability in ML applications, it is recommended to use datasets carefully designed to assess interpretability challenges and curated to reflect the problem's characteristics. This might involve selecting datasets with features challenging interpretability techniques, such as intricate feature interactions [31], non-linearity [32], and conditional dependencies [33]. While UCI and Kaggle datasets are valuable for various ML research tasks, they might not always be optimal for evaluating ML applications in critical domains requiring interpretable models, such as healthcare [1], autonomous systems [34], and finance [35].

For interpretability, comparing model-agnostic explanations to a "ground truth" understanding of the data is crucial. Unfortunately, most available datasets lack established ground truth explanations, posing challenges in assessing interpretability methods effectively. Moreover, evaluating interpretability solely on simple datasets might not adequately mirror the diversity and complexity encountered in real applications, such as high dimensionality, missing values, imbalanced classes, and noisy data. Models performing well on simpler datasets might not translate to effective explanations in complex scenarios, where relationships are less linear and classes overlap more, reducing the model's interpretability [36].

For example, consider a dataset that primarily represents a subset of a complex domain. In such cases, interpretability methods trained on this limited dataset might struggle to handle deviations beyond the domain's boundaries. This limitation can impact the method's ability to offer accurate explanations, especially when the data distribution significantly differs from what was available during training [1]. Additionally, ensuring that the rules used to generate the data capture the underlying distribution of the target domain is crucial. Biased synthetic data can impede the generalization of ML models in real-world scenarios, underscoring the importance of well-designed and realistic synthetic datasets for robust testing and evaluation [13].

The field of interpretability in ML has garnered considerable attention, leading researchers to explore a myriad of approaches. Past studies have investigated techniques spanning from rule-based interpretation to model-specific feature mapping methods. The literature showcases a diverse array of efforts dedicated to enhancing comprehension of model decisions and bridging the gap between AI system outputs and human understanding. Some of these approaches are outlined below.

Letham et al. [37] generated interpretable predictive models using Bayesian Rule Lists. These models are constructed through a series of if-then statements designed to simplify complex multivariate feature spaces into understandable decision rules. Experimental results demonstrate that Bayesian Rule Lists achieve predictive accuracy

comparable to leading ML algorithms. They showcased high accuracy and interpretability in medical scoring systems, suggesting potential replacement of the CHADS score, commonly used in clinical practice to estimate stroke risk in atrial fibrillation patients.

Ribeiro et al. [38] introduced a novel model-agnostic system utilizing anchors, which act as localized and sufficient conditions to efficiently compute explanations for any black-box model. The versatility of these anchors was demonstrated across various models in different domains and tasks. A user study revealed that anchors notably enhance users' ability to predict a model's behavior on unseen instances with greater precision and reduced effort compared to existing linear explanations or scenarios lacking explanations.

Chen et al. [39] introduced a methodology for instance-wise feature selection aimed at model interpretation. This method involves training a feature selector to identify the most informative subset of features for each specific example. The optimization objective for this selector is to maximize the mutual information between the selected features and the response variable. The authors claim the method's utility lies in explaining the behavior of models requiring interpretation of the conditional distribution of the response variable given the input. Additionally, the study introduces an efficient variational approximation for computing mutual information and demonstrates the methodology's effectiveness across diverse datasets, both synthetic and real, using quantitative metrics and human evaluations.

Hooker et al. [40] introduced an empirical measure for assessing the approximate accuracy of feature importance estimates within deep neural networks. Their experiments, conducted across multiple large-scale image classification datasets, revealed that several widely adopted interpretability methods produce feature importance estimates that do not outperform randomly assigned feature importance values.

Various techniques, including gradient methods and surrogate models, have been proposed to analyze the behavior of complex models. However, the development of datasets tailored for evaluating interpretability is not as common as creating interpretable ML models. This scarcity is due to several factors, such as the complexity of data collection [41], subjectivity in interpretability [1], and a lack of standardization [39].

2.3. Benchmark design

Synthetic rule-based datasets enable the design of interpretable models with clearly defined rules that are easier to explain [3]. The development of rule-based datasets represents a crucial step toward evaluating and improving the decision-making mechanisms of ML models, leading to the further development of reliable and robust AI systems. These datasets are critical for rigorous testing of ML applications and can serve as an indispensable benchmark for evaluating model performance in tightly controlled and well-defined scenarios. However, designing rule-based datasets presents several challenges, which can impact the dataset's quality, applicability, and representativeness:

1) Designing large and complex synthetic datasets requires striking a balance between computational effort and meaningfulness. As dataset size increases, the computational demand for training models also rises. Handling high-dimensional feature spaces or intricate structures can be particularly computationally taxing when generating sizable synthetic datasets. Scaling the dataset generation process to accommodate big data requirements presents significant computational challenges. To ensure feasibility and practicality, it is crucial to manage the size and complexity of datasets effectively. Employing efficient generative techniques becomes essential, maintaining a representation of real-world complexity within the dataset for robust evaluation [11].
2) The choice and definition of rules significantly impact a dataset's utility. Selecting meaningful and relevant rules is crucial to ensure that the dataset accurately mirrors the target application's characteristics. However, erroneous rule choices can introduce bias or unrealistic patterns, adversely affecting the dataset's usefulness in training and testing ML models. Balancing rule complexity with interpretability is a delicate trade-off. Complex rules might hinder model interpretability and obscure feature-label relationships, while overly simple rules may oversimplify the problem domain, resulting in inadequate datasets. Achieving the right balance in rule complexity is crucial to creating synthetic datasets that are both realistic and interpretable [15].
3) Accurately capturing complex interactions within real-world data poses a challenge in designing synthetic datasets. Precisely modeling intricate data relationships, especially those based on overlapping or nesting rules, is non-trivial. Ensuring rule consistency and reproducibility is crucial for creating reliable benchmark datasets where different data generation runs produce consistent results [42]. Additionally, there is a risk of overfitting to the specific rules used in dataset creation. When a dataset closely mirrors its generating rules, ML models may perform exceptionally well on synthetic data but struggle with real-world data due to differing distributions [14].
4) Noise and uncertainty play vital roles in replicating the inherent randomness and variability present in real datasets. Introducing noise and uncertainty into datasets is crucial for capturing the inherent stochastic nature of real-world data. Integrating these elements allows synthetic datasets to better emulate the diverse and unpredictable characteristics of real data, although achieving the right balance is challenging. Excessive noise might obscure underlying patterns, reducing the dataset's significance, while insufficient noise may fail to accurately represent real-world scenarios [16]. As more rules and noise are incorporated, they systematically increase complexity, offering greater control and incremental evaluation of their impact on model performance [9].
5) Robustness and Adversarial Testing. Synthetic rule-based datasets can be crafted to incorporate specific edge cases and corner scenarios, challenging models to excel under adverse conditions and constraints where traditional algorithms might falter [7]. Adversarial samples, more difficult to detect, underscore the importance of assessing robustness to gauge a model's consistency and generalization ability. Analyzing a model's response to these demanding scenarios helps identify potential pitfalls and assess how effectively a model generalizes beyond simple patterns [43].

A recent survey on neural network interpretability [44] provides a comprehensive overview of the intricate concept of interpretability, emphasizing its pivotal role in fostering trust. Within this research domain, several related studies offer insights relevant to the creation of rule-based datasets. These encompass discussions regarding: (a) the efficacy of rule-based datasets in addressing challenges and augmenting the importance of model interpretability [2]; (b) strategies to facilitate explainable classifier predictions [3]; (c) the influence of adversarial samples [8] and handling imbalanced class distributions [45] on model generalization; and (d) assessing neural network robustness against diverse corruptions and perturbations [7].

3. Methodology

Rule-based datasets serve as standardized benchmarks, facilitating the comparative evaluation of models across varying complexity levels and noise considerations. FormulAI stands out as a comprehensive framework designed to create rule-based datasets to address diverse challenges encountered in ML applications. This method presents a systematic approach to generating synthetic data instances, employing explicit rules governing the relationships between input features and output labels. Its primary aim is to challenge model interpretability and enhance prediction explainability within ML applications.

FormulAI can function as a foundational framework for establishing tailored benchmarks that replicate real-world scenarios. This approach encourages systematic investigations into model behavior, interpretability, and overall performance. The creation of the proposed rule-based dataset involves a systematic process that includes selecting features, formulating explicit rules, and implementing them algorithmically. These datasets not only challenge ML models but also enhance their capability to navigate complex decision-making environments.

The crux of this methodology revolves around crafting transparent rules that govern class assignments, considering both categorical and continuous features. These explicit rules bridge the gap between opaque "black box" models and human comprehension by providing insight into how features influence predictions. Additionally, intentionally introducing imbalance in the dataset mirrors real-world scenarios, offering a strategic approach to addressing label imbalance challenges while preserving model interpretability.

The resulting datasets can serve as controlled benchmarks, validating the performance and robustness of ML models under diverse conditions. Furthermore, they foster enhancements in model interpretability by design.

3.1. Feature selection

The proposed method starts with the selection of categorical and continuous features. Categorical features cover different classes or names and capture the variation inherent in real data, whereas continuous features encapsulate quantitative properties that facilitate decision-making.

Each categorical feature represents a distinct aspect of the underlying data distribution. For instance, in a financial transaction record, the classification function for transaction categories may include labels such as "retail," "entertainment," "groceries," "healthcare," and "travel." Likewise, continuous features span a range of quantitative measures. For instance, in a climate modeling scenario, the temperature – a continuous function – might span from −10 °C to 40°C. In a financial risk assessment dataset, continuous values for income could range from $20,000 to $200,000 per year, showcasing variations in income levels among individuals.

In FormulAI, categorical and continuous value ranges are selected to emulate the diversity and intricacy found in real data distributions. This enables the creation of intricate and meaningful rules that define the association between features and labels.

Each categorical feature, denoted as f_c, delimits a range of possible categorical values $c_i \in C$, where $C = \{C_0, C_1, \ldots, C_{m-1}\}$, $0 \leq i < m$. Here, m denotes the maximum number of distinct values that can be attributed to f_c.

The default configuration employs $m = 10$ for all categorical features: (a) C_0 represent "no information", (b) values C_1, C_2, and C_3 form the rules for assigning labels to each instance, and (c) values from C_4 to C_9 introduce random noise into the generated dataset. Hence, only C_1, C_2, and C_3 are utilized in creating the labeling rules for instances, and none of the other values (C_0 and $C_4 \ldots C_9$) should be considered as criteria for interpretability.

The parameter m is adjustable to simulate various complexity requirements. In practical terms, increasing the value of m will result in more challenging interpretability scenarios, as it is correlated with the complexity of finding the explanations leading to each target label. However, we observed that when using the default setup proposed – wherein 3 out of 10 possible categorical values per feature are linked to assigned labels – only 30% of the valid categorical values contribute to compiling rules used for labeling instances, thereby presenting initial interpretability challenges.

Next, a feature set, denoted as f_v, is devised to represent a range of continuous values $v \in \mathrm{R}$, where $v_{min} \leq v < v_{max}$ and (v_{min}, v_{max}) are the lower and upper bound values for each feature f_v. By default, the configuration setup uses (v_{min}, v_{max}) = (0, 10). Akin to categorical features, the integer portion of each continuous value is utilized: (a) $v = 0.0$ denotes "no information", (b) values 1.0, 2.0, and 3.0 are employed in constructing rules that allocate labels to individual instances, and (c) values $v \geq 4.0$ introduces random noise into the generated dataset. Lastly, the decimal portion of each continuous value contributes to adding noise to the resultant dataset.

In each experimental setup presented in the evaluation protocol, the number of features varies depending on the complexity of the simulation.

3.2. Synthetic rules

The rationale behind crafting rules that capture real-world challenges while maintaining interpretability is two-fold. Firstly, it empowers ML models to navigate intricate challenges, enhancing their adaptability and learning capabilities. Secondly, it serves as a conduit between model functionality and human comprehension, a vital aspect in fostering trust and accountability in AI systems.

Balancing rule complexity with the need for interpretability stands as a crucial aspect in synthetic datasets, offering robust benchmarks for evaluating ML models. While complex rules hold potential in encapsulating intricate data patterns, their lack of transparency might prevent human understanding. Conversely, overly simplistic rules may fail to capture the subtleties present in real-world data.

Motivated by the goal of capturing latent data patterns that might elude simpler rules, FormulAI establishes various connections among features. This attribute is crucial for generating more intricate patterns that cannot be detected using basic rule structures. The strategic design and iterative construction of rules are tailored to generate tuples based on four distinct criteria: (a) the number of features comprising each rule, (b) the potential valid values assignable to each feature, (c) the intended target label, and (d) the level of noise introduced into each resulting tuple.

By default, each generation rule allows a maximum combination of 3 features. This means a rule can be defined by a single feature or a combination of two or three categorical and/or continuous features. While this parameter impacts the size of the resulting dataset, it offers versatility. Once a controlled value is assigned to each categorical or continuous feature selected for a generation rule, all resulting tuples governed by the same rule are assigned the same target label. Finally, noise is added.

3.3. Labeling class imbalance

Deliberately incorporating label imbalance into the rule-based datasets serves as a strategic mechanism for emulating the

imbalanced class distributions often encountered in practical scenarios. FormulAI aims to replicate challenges posed by rare outcomes while maintaining interpretability within the generated datasets. By mimicking situations where certain classes naturally occur less frequently, this approach elevates the dataset's complexity. Importantly, the calibrated imbalance introduced does not compromise the inherent interpretability of the data instances. This balanced fusion of realism and transparency allows for assessing model performance in scenarios akin to real-world situations, ensuring authenticity without unnecessary complexity.

For each new rule considered during the dataset generation process, a subsequent corresponding class label is selected from a predefined list of labels. The frequency of each target label within this reference list determines the imbalance factor for each classification class. For instance, within the default label set, there are six distinct labels (denoted as $l \in L$), where $L = \{P, Q, R, S, T, V\}$, distributed in the following imbalanced sequence:

$$\text{target labels} = [\text{P}, \text{Q}, \text{R}, \text{S}, \text{T}, \text{V}, P, Q, R, S, T, P, Q, R, S, P, Q, R, P, Q, P, V]$$

According to the proposed sequence of target labels above, "P" emerges as the predominant label, being selected by 6 out of 22 rules, while "T" and "V" represent the minority classes, each chosen only by 2 out of 22 rules. The target labels can assume various configurations concerning the number of classes and imbalanced distributions. Therefore, as the rules generate samples, the initial rule allocates all resulting samples the label "P," and subsequently, the second rule assigns the label "Q" to the generated samples. Note that the list of *target labels* functions in a circular manner; once the last element in the list is reached, the subsequent generation rule will consider the first element as the next target class. This list of target labels can be arranged in numerous ways to create (a) balanced distributions, (b) binary classifications, or even (c) a greater number of target classes.

3.4. Random noise

Incorporating unnecessary noise can result in datasets that are excessively complex and difficult to interpret, thereby obscuring discernible underlying patterns. Conversely, insufficient noise may not capture the intricate nuances inherent in real-world data. The introduction of controlled noise inputs an element of authenticity by emulating the uncertainties, while preserving the quality of the generated dataset, assuring that the established rules remain comprehensible and interpretable.

FormulAI introduces noise to both continuous and categorical features that are still designated as $c = C_0$ or $v = 0.0$ after their respective controlled values for categorical features are assigned. The noise ratio parameter (nr) controls the degree of random noise incorporated into the resulting synthetic dataset. When $nr = 0.0$, no noise is added, meaning that all categorical or continuous features not used as part of a generation rule retain values as $c = C_0$ or $v = 0.0$, respectively. Conversely, when $nr = 1.0$, the maximum allowable noise is applied, resulting in random values assigned to all categorical and continuous features not utilized as part of a generation rule.

For example, when $nr = 0.1$: (a) 10% of categorical features that still retain the value C_0 will have a random value selected between C_4 and C_9 assigned; (b) 10% of continuous features that still hold the value 0.0 will have a random value between 4 and 9 (integer) assigned; and finally, (c) for each of continuous feature assigned a controlled value $\in \{1.0, 2.0, 3.0\}$ additional random noise $\pm r \square nr$ is added, where r is randomly selected from $0 \leq r < 1$.

Although the resulting rule-based datasets aim for interpretability, the proposed evaluation protocol was designed to ensure an appropriate level of complexity and noise, making the benchmark dataset challenging yet interpretable. Each instance is uniquely identified by a name that represents the specific rule used to generate that particular row. For instance, in the primary proposed dataset, a test record is denoted as "Fc8C1Fv17V3Fv19V2LRS4," indicating: (a) categorical feature $Fc8$ is set as C_1, (b) continuous feature $Fv17$ is set as 3.0, (c) continuous feature $Fv19$ is set as 2.0, (d) the label assigned to this row is "R," and (e) this sample is the fourth one among a total of 25 samples generated with the same rule. Features not explicitly mentioned in the sample ID are either set as C_0 (categorical) or 0.0 (continuous) or are assigned random values such as C_4 to C_9 (categorical) or 4.0 to 9.0 (continuous) with a random ratio of 0.35. For instance, categorical feature is $Fc7 = C_6$, while continuous feature $Fv17$ has some added noise resulting in its final value being 2.9492 (instead of 3.0). Other categorical and continuous features also possess randomly assigned values.

3.5. Evaluation protocol

The experimental design aims to ensure that the resulting benchmark datasets present challenges for both model performance and the ability to explain predictions. To comprehensively evaluate the effectiveness and versatility of datasets generated through FormulAI, the following evaluation protocol was designed. It measures how different parameters used during rule-based dataset generation impact the performance of four baseline ML models: (a) logistic regression (LR) [46], (b) support vector machines (SVMs) [47, 48], (c) random forest (RF) [49], and (c) eXtreme gradient boosting (XGBoost) [50]. Evaluated parameters include the dataset size (determined by the number of features and rules) and the noise ratio.

Each model is trained to predict target labels in three distinct ways. Firstly, the models are trained for exclusive classification, where a single unique label chosen from non-overlapping groups or classes is assigned to each instance. Next, a binary classification model is trained to distinguish the majority class "P" from the others. Lastly, another binary classification model is trained to distinguish the minority class "V" from the others. Datasets are generated by splitting the data into training and test sets. However, during evaluation, 20% of instances are randomly extracted from the training set to form a tuning set. This subset aids in fine-tuning specific parameters that might impact the effectiveness of each approach, such as setting thresholds for binary classification concerning imbalanced labels or determining the maximum depth for tree-based models (RF and XGBoost).

Models are trained using the training set. The maximum number of iterations is set to 25,000 for LR and 100,000 for SVM. In RF and XGBoost, models are tuned to determine the best maximum depth, selecting values between 5 and 20. All parameter optimizations are based solely on the tuning set. Finally, the test set is utilized to calculate the final model performance, reported as $F1$ score. Additionally, AUPRC is presented to showcase how varying levels of added noise affect model performance.

4. Experimental Results and Discussion

Our experimental results are based on evaluating different dataset generation parameters using four baseline approaches. Initially, we showcase how the size and complexity of datasets, in terms of rules, impact the performance of baseline models. Subsequently, while keeping all parameters constant except for

the noise level, which varies from 0.0 to 0.35, we analyze its effect. Finally, we introduce a benchmark rule-based dataset that we consider challenging in terms of predictions and explainability.

4.1. Dataset complexity

The complexity of a generated dataset, determined by its number of features and the inherent rules used to assign target classes to each sample, can significantly impact the performance of a baseline model. This effect is assessed in the initial experiment. Table 1 outlines the characteristics of five distinct datasets created to augment complexity concerning the number of features, instances, and rules.

Some parameters used to generate the datasets were fixed. No noise ratio was applied ($nr = 0.0$), and eight resulting samples were generated per rule, of which two samples were allocated to the test set, maintaining a 60:20 ratio between the training and test sets. The resulting $F1$ scores are presented in Table 2 and compared in Figure 1.

Table 1
Five datasets generated to assess how the number of features, instances, and rules impacts baseline model performance

F	*Tr*	*Ts*	*R*	*Rp*	*Rv*
2	90	30	15	3	1
4	1044	348	174	47	15
8	10728	3576	1788	487	163
16	97488	32496	16248	4431	1477
32	830880	276960	138480	37767	12589

where F = number of features (equally split between categorical and continuous) Tr = number of instances (rows) in the training set; Ts = number of instances (rows) in the test set; R = number of different rules assigning target labels; Rp = number of different rules assigning the majority class P; Rv = number of different rules assigning the minority class V

XGBoost not only outperforms other baselines but also perfectly identifies all correct answers in less complex dataset formulations. However, the performance of LR and SVM tends to decline as the complexity of the benchmark dataset increases. In the first three datasets, XGBoost achieved a maximum tuned $F1$ score of 1.0 using a decision tree depth between 5 and 9. Conversely, for the more complex dataset, the best $F1$ score was obtained using a depth of 19.

LR assumes a linear relationship between features and the target variable. While it can perform well in certain scenarios, it may fail to capture complex nonlinear relationships in the data, particularly when the decision boundary is not well approximated by a linear function [14]. Although LR models have been used as baselines in several experiments, we found they can struggle to resolve challenging datasets, as those proposed in this work. Although SVM remains effective as a baseline for predicting minority classes, its time-consuming training process might render it unfeasible for more realistic tasks (see Table 3). In our evaluation, the decision tree-based ensemble models, RF and XGBoost, are the top-performing candidates. Notably, while both models exhibited strong performance, XGBoost still demonstrated a slight superiority over RF, as revealed by our experimental results.

4.2. Noise level

The second experiment aimed to assess the impact of introducing random noise into the rule-based generated data. Table 4 illustrates how categorical features are affected by random noise. Controlled categorical values C_1 to C_3 assigned to each categorical feature remain unchanged. However, random noise replaces the original C_0

Table 2
Resulting $F1$ scores were obtained from evaluations using XGBoost (XGB), random forest (RF), SVM, and logistic regression (LR). These evaluations were conducted on five datasets featuring varying levels of complexity. Each dataset differed in the number of categorical and continuous features, instances, and rules used to assign target labels (refer to Table 1 for detailed information)

	Multilabel (macro $F1$)				Predominant class (label P)				Minority class (label V)			
Feat	LR	SVM	RF	XGB	LR	SVM	RF	XGB	LR	SVM	RF	XGB
2	0.1476	1.0000	1.0000	1.0000	0.6000	1.0000	1.0000	1.0000	0.6667	1.0000	1.0000	1.0000
4	0.1858	0.5438	1.0000	1.0000	0.4332	0.7170	1.0000	1.0000	0.2414	1.0000	1.0000	1.0000
8	0.1781	0.4193	1.0000	1.0000	0.4337	0.6726	1.0000	1.0000	0.1796	0.8344	1.0000	1.0000
16	0.1818	0.2472	0.8537	1.0000	0.4287	0.6517	0.9528	0.9982	0.1674	0.7695	0.9633	1.0000
32	0.1818	0.2373	0.2709	0.9706	0.4286	0.6226	0.7199	0.9230	0.1669	0.7357	0.8853	0.9527

Figure 1
Resulting $F1$ scores from XGBoost (XGB), Random Forest (RF), SVM, and Logistic Regression (LR) models. The evaluation was conducted across five datasets characterized by varying levels of complexity. These datasets comprise different numbers of categorical and continuous features, samples, and rules utilized for assigning target labels (refer to Table 1 for detailed information). The evaluation covers three classification scenarios: (a) mutually exclusive label classification (6 labels); (b) binary classification, distinguishing the predominant class 'P' from other target classes; and (c) binary classification, distinguishing the minority class 'V' from other target classes

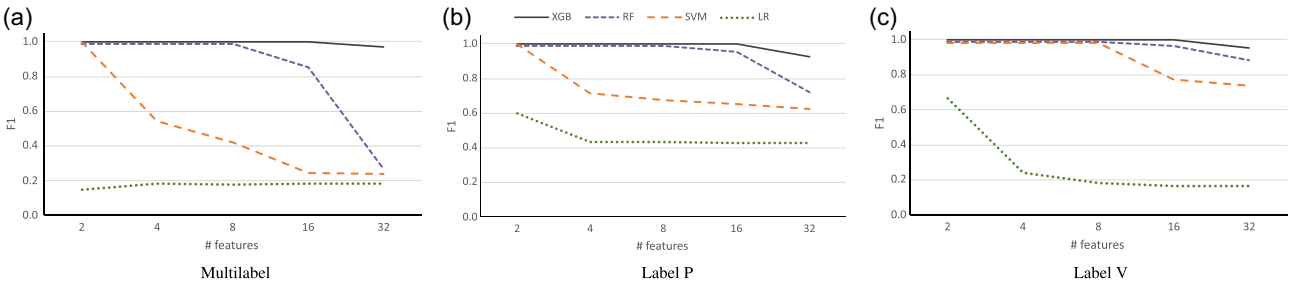

Table 3
Time to train each model – All evaluation protocol runs used an Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz processor, with 32.0 GB RAM – Values can be used as a reference to compare training resource requirements between different approaches

Model	Multilabel	Label P	Label V
LR	31 s	17 s	18 s
SVM	≈ 5 days	≈ 5 days	≈ 5 days
RF	≈ 20 h	≈ 10 h	≈ 10 h
XGB	≈ 8 h	≈ 1.5 h	≈ 1.5 h

values with random values ranging from C_4 to C_9. Similarly, continuous features undergo the same effect. Each continuous feature, replaced with a random integer value between 4.0 and 9.0, also receives a positive or negative random increment of $\pm r * nr$, where r represents a random value in the range $0 \leq r < 1$, and nr is the noise ratio. If a continuous feature remains set as 0.0 (not chosen to receive a random value), no $\pm r * nr$ noise is added.

Several parameters were standardized for generating the datasets: (a) four categorical and four continuous features were used; (b) the training set consists of 21,456 samples, while the test set comprises 5,364 samples; (c) 1,788 distinct rules were employed to assign one of six target labels to each sample; finally, (d) 25 samples were generated per rule, of which 5 were allocated to the test set. The resulting $F1$ scores are detailed in Table 5 and compared in Figure 2.

In contrast to the complexity introduced by features and rules, the addition of noise significantly impacts the performance of all baseline models. Although XGBoost shows strong performance in varied dataset complexities, increasing random noise negatively affects its classification performance. Nonetheless, even with added noise, XGBoost remains more efficient compared to SVM and LR as a baseline model.

4.3. Challenging benchmark dataset

In this final experiment, a benchmark dataset is proposed, and Table 6 illustrates all parameters used in its generation.

These datasets include: (a) mutually exclusive label classification (6 labels); (b) binary classification differentiating the predominant class P from other target classes; and (c) binary classification differentiating the minority class V from other target classes. Further details can be found in Table 4.

Table 4
Frequency of categorical values *C*0 to *C*9 in the first categorical feature when the noise ratio increases from 0.0 to 0.35 – Once controlled values *C*1 to *C*3 are assigned to each feature, they are never replaced by random values

		Controlled values			Random values					
Noise	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
0.000	13860	2532	2532	2532	0	0	0	0	0	0
0.025	13514	2532	2532	2532	54	49	67	58	67	51
0.050	13191	2532	2532	2532	102	108	115	128	85	131
0.075	12821	2532	2532	2532	176	168	186	174	175	160
0.100	12488	2532	2532	2532	232	206	239	215	222	258
0.150	11699	2532	2532	2532	360	351	352	382	355	361
0.200	11081	2532	2532	2532	455	455	460	483	458	468
0.250	10456	2532	2532	2532	570	551	562	566	592	563
0.300	9646	2532	2532	2532	719	672	713	723	684	703
0.350	9088	2532	2532	2532	809	799	783	819	792	770

Figure 2
***F*1 scores resulting from the evaluation of XGBoost (XGB), Random Forest (RF), SVM, and Logistic Regression (LR) in 10 datasets with varying levels of noise are presented. These datasets include: (a) mutually exclusive label classification (6 labels); (b) binary classification differentiating the predominant class P from other target classes; and (c) binary classification differentiating the minority class V from other target classes. Further details can be found in Table 4**

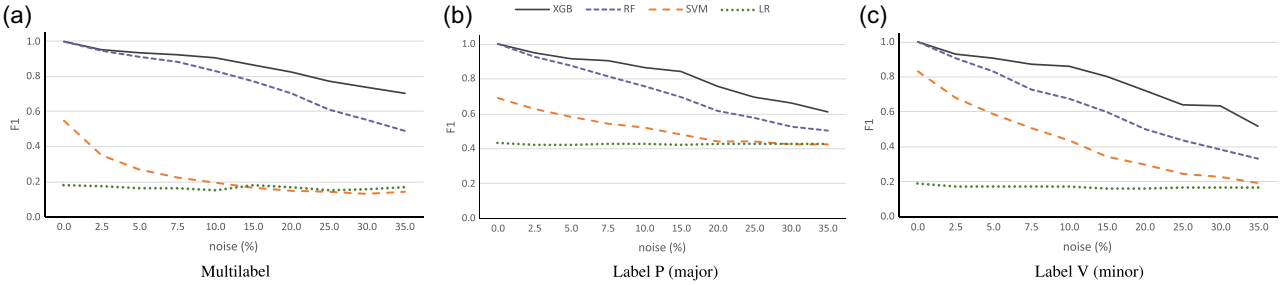

Table 5
Resulting *F*1 scores, precision, and recall from XGBoost, SVM, and logistic regression, evaluated in ten datasets with distinct levels of noise

	Multilabel			Predominant class (label P)									Minority class (label V)					
	Macro *F*1			*F*1			Precision			Recall			*F*1			Precision			Recall		
Noise	LR	SVM	XGB	LR	SVM	XGB	LR	SVM	XGB	LR	SVM	XGB	LR	SVM	XGB	LR	SVM	XGB	LR	SVM	XGB
0.000	0.1781	0.5464	1.0000	0.4349	0.6913	1.0000	0.2788	0.6286	1.0000	0.9877	0.7680	1.0000	0.1907	0.8306	1.0000	0.1129	0.9058	1.0000	0.6135	0.7669	1.0000
0.025	0.1721	0.3495	0.9532	0.4245	0.6304	0.9521	0.2722	0.5461	0.9649	0.9637	0.7454	0.9398	0.1729	0.6818	0.9329	0.0958	0.6137	0.9570	0.8875	0.7669	0.9100
0.050	0.1634	0.2676	0.9350	0.4235	0.5829	0.9172	0.2719	0.4954	0.9248	0.9569	0.7077	0.9097	0.1743	0.5857	0.9049	0.0968	0.5007	0.9365	0.8712	0.7055	0.8753
0.075	0.1649	0.2238	0.9238	0.4273	0.5455	0.9052	0.2720	0.4493	0.9189	0.9959	0.6940	0.8919	0.1736	0.5045	0.8736	0.0963	0.4180	0.9347	0.8773	0.6360	0.8200
0.100	0.1519	0.1944	0.9026	0.4273	0.5233	0.8632	0.2724	0.4228	0.8998	0.9904	0.6865	0.8296	0.1744	0.4357	0.8600	0.0968	0.3348	0.9330	0.8814	0.6237	0.7975
0.150	0.1786	0.1639	0.8641	0.4224	0.4837	0.8444	0.2719	0.3883	0.8245	0.9459	0.6413	0.8652	0.1623	0.3415	0.8054	0.0980	0.2590	0.8889	0.4724	0.5010	0.7362
0.200	0.1656	0.1454	0.8218	0.4283	0.4429	0.7557	0.2725	0.3441	0.7308	1.0000	0.6215	0.7823	0.1635	0.2963	0.7226	0.0907	0.2142	0.7975	0.8282	0.4806	0.6605
0.250	0.1519	0.1446	0.7744	0.4273	0.4425	0.6982	0.2722	0.3352	0.6445	0.9938	0.6509	0.7618	0.1663	0.2467	0.6393	0.0917	0.1614	0.7794	0.8916	0.5235	0.5419
0.300	0.1575	0.1301	0.7376	0.4282	0.4279	0.6621	0.2724	0.2723	0.6088	1.0000	0.9986	0.7255	0.1696	0.2284	0.6323	0.0937	0.1494	0.7188	0.8896	0.4847	0.5644
0.350	0.1674	0.1431	0.7049	0.4278	0.4274	0.6129	0.2723	0.2762	0.5863	0.9973	0.9446	0.6420	0.1683	0.1908	0.5191	0.0919	0.1312	0.5228	0.9918	0.3497	0.5153

Table 6
Parameters used to generate the final proposed benchmark dataset – The resulting dataset is the one evaluated in Figure 3 by the full red line

Parameter	Value/Description
Features	24
Categorical	12
Continuous	12
Labels	$L = \{P, Q, R, S, T, V\}$
Rules (total)	57,204
P	15,601
Q	13,001
R	10,401
S	7,801
T	5,200
V	5,200
Samples	25 (per rule)
Noise ratio	0.35
Training set	80% (1,144,080 rows)
Test set	20% (286,020 rows)

The chosen parameters were carefully selected to ensure that the resulting dataset presents significant challenges and is inherently non-trivial to address.

The noise ratio was set to 0.35. However, due to the larger number of features and samples per rule, the proposed dataset exhibits notably higher levels of complexity and difficulty, particularly in terms of class predictions. This complexity is compared to previously evaluated experimental datasets with different levels of noise, as illustrated in Figure 3. The blue AUPRCs represent the incremental noise added in the initial experimental setup (see Section 4.2.), while the red AUPRC is the curve derived from the final proposed benchmark dataset. All curves were extracted from the tuning set when attempting to predict the majority class "P."

Finally, Table 7 presents the final precision, recall, and *F*1 scores resulting from each of the baseline models considered in the evaluation protocol. LR, as a simpler linear model, may struggle to capture complex, nonlinear relationships within rule-based datasets. While generally considered more interpretable, this simplicity can be a limitation when dealing with intricate data patterns. SVM can perform well when the data are separable or linearly/non-linearly separable with an appropriate kernel. However, it does not prove to be suitable for the proposed task and is time-consuming in terms of the training process, as demonstrated in Table 3. On the other hand, XGBoost, an ensemble method combining predictions from multiple decision trees, is known for effectively handling complex, nonlinear relationships within data, demonstrating superior performance as a baseline model when dealing with rule-based generated data.

The presented results indicate that the proposed benchmark dataset is an invaluable resource for challenging ML applications, especially in the realm of interpretability. These findings affirm the dataset's pivotal role in fostering advancements in the field and addressing the longstanding challenges associated with model understanding and prediction explainability in AI systems.

Figure 3

Resulting XGBoost AUPRC (average precision) in the tuning set comparing performance between the 10 datasets (blue) used to evaluate different noise ratios and the final proposed benchmark dataset (red), all given when predicting the majority class P, In which: (a) curves in blue result from incremental noise ratio from 0.00 to 0.35 (top to bottom, respectively) as in the experimental setup described in Table 5; (b) curve in red results from the main proposed dataset, which also has a noise ratio of 0.35, however designed with a larger number of features, as described in Table 6

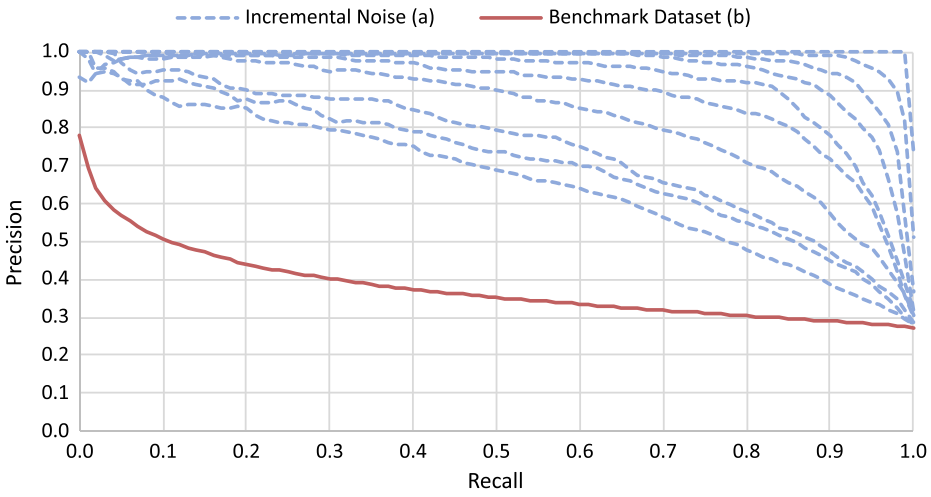

Table 7

Resulting *F*1 scores, precision, and recall from XGBoost (XGB), random forest (RF), SVM, and logistic regression (LR), evaluated with the main proposed dataset

	Multilabel	Predominant class (label P)			Minority class (label V)		
ML	Macro *F*1	*F*1	Precision	Recall	*F*1	Precision	Recall
LR	0.1572	0.4286	0.2727	0.9997	0.1667	0.0909	0.9999
SVM	0.1831	0.4285	0.2728	0.9987	0.1665	0.0908	0.9902
RF	0.1799	0.4295	0.2763	0.9640	0.1797	0.1079	0.5358
XGB	0.4195	0.4427	0.3049	0.8071	0.2892	0.2450	0.3529

5. Conclusion

This paper introduces FormulAI, a novel framework for generating rule-based benchmark datasets designed to challenge ML applications, particularly in the domain of interpretability. The resulting generated datasets consist of complex rules, providing transparent explanations for their predictions. Through the controlled introduction of noise, label imbalance, and intricate rules within the dataset, it mirrors real-world complexities without sacrificing interpretability. The evaluation protocol demonstrates that the resulting datasets are not easy to resolve, making them a valuable asset for the machine learning community.

Looking ahead, there are several avenues for future work stemming from this research. Firstly, we plan to expand the dataset generation process to encompass a broader range of domain-specific real-world application challenges, such as those found in healthcare and finance, including multi-relational data generation. Additionally, we will continue to refine the benchmark dataset by introducing more complex rules based on domain-specific knowledge, such as feature correlations, aiming not only to enhance model interpretability but also synthetic data generation. Finally, we plan to improve the proposed approach by generating synthetic datasets to mimic those problems in which recurrent and convolutional neural networks are applicable.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/hextrato/FormulAI.

Author Contribution Statement

Hegler Tissot: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613

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

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint: 1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

[5] Abufadda, M., & Mansour, K. (2021). A survey of synthetic data generation for machine learning. In *22nd International Arab Conference on Information Technology*, 1–7. https://doi.org/10.1109/ACIT53391.2021.9677302

[6] Nikolenko, S. I. (2021). *Synthetic data for deep learning*. Switzerland: Springer. https://doi.org/10.1007/978-3-030-75178-4

[7] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv Preprint: 1903.12261*. https://doi.org/10.48550/arXiv.1903.12261

[8] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 1–11.

[9] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning,* 1060–1069.

[10] Yang, J., & Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 3. https://doi.org/10.1145/2350190.2350193

[11] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115. https://doi.org/10.1145/3446776

[12] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. USA: The MIT Press.

[13] Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

[14] Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Canada: Springer.

[15] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[16] Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., & Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition,* 6575–6583. https://doi.org/10.1109/cvpr.2017.696

[17] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

[18] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93. https://doi.org/10.1145/3236009

[19] Zhou, Z., Zhao, G., & Pietikäinen, M. (2011). Towards a practical lipreading system. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, 137–144. https://doi.org/10.1109/cvpr.2011.5995345

[20] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., …, & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

[21] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15. https://doi.org/10.1145/1541880.1541882

[22] Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

[23] Unwin, A., & Kleinman, K. (2021). The iris data set: In search of the source of *virginica*. *Significance*, *18*(6), 26–29. https://doi.org/10.1111/1740-9713.01589

[24] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

[25] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, *43*(4), 570–577. https://doi.org/10.1287/opre.43.4.570

[26] Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

[27] Moody, G. B., Mark, R. G., & Goldberger, A. L. (2001). PhysioNet: A web-based resource for the study of physiologic signals. *IEEE Engineering in Medicine and Biology Magazine*, *20*(3), 70–75. https://doi.org/10.1109/51.932728

[28] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Li, F. F. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/cvpr.2009.5206848

[29] Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, *15*(2), 49–60. https://doi.org/10.1145/2641190.2641198

[30] Leskovec, J., & Sosič, R. (2017). SNAP: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology*, *8*(1), 1. https://doi.org/10.1145/2898361

[31] Kha, Q. H., Le, V. H., Hung, T. N. K., Nguyen, N. T. K., & Le, N. Q. K. (2023). Development and validation of an explainable machine learning-based prediction model for drug–food interactions from chemical structures. *Sensors*, *23*(8), 3962. https://doi.org/10.3390/s23083962

[32] de Haro Pizarroso, G., & van Kampen, E. J. (2023). Explainable artificial intelligence techniques for the analysis of reinforcement learning in non-linear flight regimes. In *AIAA SCITECH 2023 Forum*. https://doi.org/10.2514/6.2023-2534

[33] Cortiñas-Lorenzo, K., & Lacey, G. (2023). Toward explainable affective computing: A review. *IEEE Transactions on Neural*

Networks and Learning Systems, *35*(10), 13101–13121. https://doi.org/10.1109/tnnls.2023.3270027

[34] Flammini, F., Alcaraz, C., Bellini, E., Marrone, S., Lopez, J., & Bondavalli, A. (2022). Towards trustworthy autonomous systems: Taxonomies and future perspectives. *IEEE Transactions on Emerging Topics in Computing*, *12*(2), 601–614. https://doi.org/10.1109/tetc.2022.3227113

[35] Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, *74*, 867–907. https://doi.org/10.1007/s11301-023-00320-0

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

[37] Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371. https://doi.org/10.1214/15-aoas848

[38] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11491

[39] Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, 883–892.

[40] Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In *33rd Conference on Neural Information Processing Systems,* 1–12.

[41] Rudin, C. (2014). Algorithms for interpretable machine learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1519–1519. https://doi.org/10.1145/2623330.2630823

[42] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. USA: MIT Press.

[43] Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3–14. https://doi.org/10.1145/3128572.3140444

[44] Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(5), 726–742. https://doi.org/10.1109/tetci.2021.3100641

[45] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

[46] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. USA: Wiley.

[47] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1023/a:1022627411411

[48] Steinwart, I., & Christmann, A. (2008). *Support vector machines*. USA: Springer.

[49] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

[50] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

How to Cite: Tissot, H. (2025). FormulAI: Designing Rule-Based Datasets for Interpretable and Challenging Machine Learning Tasks. *Artificial Intelligence and Applications*, *3*(1), 72–82. https://doi.org/10.47852/bonviewAIA42021781