Evaluating Logic Learning Machine model with SAFE-AI Metrics

Vasily Kolesnikov[1,*], Damiano Verda[2], Nicolò Pinna[2], Danilo Franco[2] and Paolo Giudici[1]

[1]University of Pavia, Corso Strada Nuova, 65 - 27100 Pavia – Italy
[2]Rulex Innovation Laboratories, 16122 Genova, Italy

Abstract

Explainable Artificial Intelligence is increasingly required in high-stakes domains where models must be accurate, robust, and transparent. In this work, we evaluate the Rulex Logic Learning Machine (LLM), an inherently interpretable rule-based classifier, using both traditional predictive metrics and SAFE-AI metrics based on the Rank Graduation framework derived from the Cramér–von Mises divergence. The evaluation is conducted on three heterogeneous datasets: HMDA credit lending data, a brain cancer MRI dataset, and the Financial PhraseBank text dataset. The results show that the SAFE-AI metrics provide complementary insights beyond the standard evaluation. In particular, the rule-based model performs competitively on structured tabular data, while classical models dominate in image and text settings.

Keywords

Explainable Artificial Intelligence, SAFE-AI, Logic Learning Machine, Rule-based Learning, Robustness, Explainability, Credit Lending, Medical Imaging, Financial Text Classification

1. Introduction

Explainable Artificial Intelligence (XAI) has gained increasing relevance in high-stakes domains such as credit risk assessment, medical diagnosis, and financial decision-making [1, 2]. In these settings, predictive accuracy alone is insufficient: models must also provide transparent reasoning, robustness to perturbations, and consistent behavior across subgroups. This multidimensional evaluation challenge motivates the integration of interpretable modeling approaches with principled risk and compliance metrics.

Logic Learning Machine (Rulex LLM), developed by Rulex Innovation Laboratories, is a rule-based learning methodology designed to develop predictive models in which the underlying inferential mechanisms are inherently interpretable and transparent, facilitating structured justification and explanation of the resulting outcomes. Despite this structural interpretability, the Rulex LLM model is usually evaluated using traditional predictive metrics such as Accuracy, F1-score, or AUC. The Secure, Accurate, Fair and Explainable (SAFE) framework [3] proposes a unified statistical perspective for a deeper evaluation beyond these metrics using Cramér–von Mises divergence [4]. Within this framework, explainability and robustness are assessed through model perturbations, such as removing variables or adding Gaussian noise, allowing quantification of their impact on model behavior and enabling evaluation of both predictive performance and model stability.

In this work, we investigate how a logic-based model such as Rulex LLM behaves when evaluated under both classical predictive metrics and SAFE-AI metrics. We perform a cross-domain empirical analysis on three datasets: the HMDA mortgage lending dataset, the brain cancer image classification dataset, and the financial text classification dataset. By comparing Rulex LLM against classical machine learning models under a unified evaluation protocol, we aim to understand whether conclusions drawn

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.
vasily.kolesnikov01@universitadipavia.it (V. Kolesnikov); damiano.verda@rulex.ai (D. Verda); nicolo.pinna@rulex.ai (N. Pinna); danilo.franco@rulex.ai (D. Franco); paolo.giudici@unipv.it (P. Giudici)
0009-0000-7415-4721 (V. Kolesnikov); 0000-0001-9912-3454 (D. Verda); 0000-0001-5947-8728 (N. Pinna); 0000-0003-0611-8600 (D. Franco); 0000-0002-4198-0127 (P. Giudici)

from standard metrics remain consistent when accuracy resilience, robustness and explainability are explicitly quantified. It is worth noting that Rulex LLM exhibits a qualitative difference with respect to the other models considered in the comparison, being the only one to produce a set of rules representing the considered problem and, as a consequence, a motivation for each proposed decision in terms of an IF-THEN statement. Instead of digging deeper in the nuances of this differentiation, in the present paper we aim to evaluate this model more quantitatively to investigate how big the price of this naively available interpretability is, in terms of performance. In addition, we empirically assess the behavior and stability of the SAFE-AI evaluation framework, testing its generality as a domain-independent risk assessment methodology in the sense of providing a unified evaluation protocol across heterogeneous data modalities rather than enabling direct comparability of absolute metric values across datasets.

2. Methodology

2.1. Logic Learning Machine Model

The Logic Learning Machine (LLM) is a rule-based method alternative to decision trees [5] [6]. In plain words, it transforms the data into a Boolean domain where some Boolean functions (one for each output value) are reconstructed starting from a portion of their truth table with a method described in the paper [7]. The process creates a set of intelligible rules through Boolean function synthesis, evaluates their quality, applies them, and assigns an output. This happens following these steps: (i) Discretization; (ii) Latticization or Binarization; (iii) Positive Boolean function; (iv) Rule generation; (v) Standard Applied Procedure.

In a classification problem d-dimensional examples $x \in X \subset \Re^d$ are to be assigned to one of q possible classes, labeled by the values of a categorical output y. Starting from a training set S including N pairs (x_i, y_i), with $i = 1, \ldots, N$, deriving from previous observation, techniques for solving classification problems have the aim of generating a model $g(x)$, called classifier, that provides the correct answer $y = g(x)$, for most input patterns x. In the *discretization* step, a mapping converts each continuous variable domain into a discrete domain. The simplest strategy is *Equal Width discretization*, where intervals of equal length are created. This method is computationally efficient, although output-guided discretization may yield more compact representations. In the *binarization* step, each discretized domain is transformed into a binary domain and, as a result, the new training set is $S^{'} = \{(z_i, y_i)\}_{i=1}^{N}$, with $z_i \in \{0,1\}^B$, where $B = \sum_{j=1}^{d} M_j$ and M_j is the number of intervals generated for feature j. The training set $S^{'}$, obtained after binarization, can be divided into two different subsets according to the output class: T is the set containing (z_i, y_i) with $y_i = O_1$, whereas F is the set containing the example for which $y_i = O_0$. T and F can be viewed as a portion of the truth table of a *Boolean function* f that must be reconstructed. A method for finding f must retrieve the set of minimum true points to be used from T and F, to represent f in its irredundant PDNF. It follows that the set of all bottom points for (T, F) is an antichain, which elements are candidate minimum true points. The algorithm LLM employs to produce implicants is called *Shadow Clustering* [8].

In the *rule generation* step, each implicant of the positive Boolean function f is transformed into an intelligible rule, where, as said before, a function is generated for each output value. Each rule r_k (representing the k-th rule) can be generally formalised as:

$$r_k : \text{IF} \bigwedge_{j=1}^{J_i} X_j \text{ THEN } O_i \tag{1}$$

where $\bigwedge$ denotes logical AND, $X_j \in V_i$ represents the condition of the feature X_j that corresponds to a value (or range of values) of rule r_k, O_i denotes the output class associated with rule r_k, and J_i represents the number of premises (features) in rule r_k. For multiclass problems it is enough to decompose the problem into several bi-class issues, for each sub-problem, the target class is labelled with 1, and all the remaining classes with 0.

It is usual that an element of the training set covers the rule of different classes. When it happens, the output class is established according to the relevance of the rules satisfied by it. This is the so-called *Standard Applied Procedure (SAP)*. In order to present relevance, the following quantities are defined for a rule r_k in the form *IF <premise> THEN <consequence>*: $TP(r_k)$ is the number of training set examples that satisfy both the premise and the consequence of the rule r_k, $FP(r_k)$ is the number of examples that satisfy the premise but not the consequence, $TN(r_k)$ is the number of examples that satisfy neither the premise nor the consequence, and $FN(r_k)$ is the number of examples that do not satisfy the premise but satisfy the consequence.

Notice that a sample x_i satisfies the premise of the rule r_k if it satisfies all its premise conditions, whereas x_i does not satisfy the premise of the rule r_k if it does not satisfy at least one among its premise conditions. Combining these quantities, it is possible to compute quality measures for a rule r_k:

$$\text{Covering: } C(r_k) = \frac{TP(r_k)}{TP(r_k) + FN(r_k)}, \quad \text{Error: } E(r_k) = \frac{FP(r_k)}{TN(r_k) + FP(r_k)} \tag{2}$$

It is evident that the greater the covering, the more general the rule is; on the other hand, the smaller the error, the more precise the rule is. Then, the relevance of a rule r_k is obtained by combining $C(r_k)$ and $E(r_k)$:

$$R(r_k) = C(r_k)(1 - E(r_k)) \tag{3}$$

Once the relevance of the rule is defined, it is possible to compute a score $S(x_i, o)$ for each class o that measures how likely it is that $y_i = o$, and from these scores to define the probability $P(o \mid x_i)$:

$$S(x_i, o) = \sum_{r \in \mathfrak{R}_o^i} R(r), \qquad P(o \mid x_i) = \frac{S(x_i, o)}{\sum_{k \in O} S(x_i, k)} \tag{4}$$

with $\mathfrak{R}_o^i = \{r \mid r \in \mathfrak{R},\, r \leq x_i,\, O(r) = o\}$, where $\mathfrak{R}$ is the complete ruleset, $r \leq x_i$ denotes that x_i satisfies the premise of a rule and $O(r)$ denotes the consequence of the rule. Then the selected output is the one that maximizes the output probability $\tilde{y}_i = \underset{o}{max} P(o \mid x_i)$.

2.2. SAFE-AI Metrics

To evaluate the proposed machine learning models in a coherent and comparable manner, we adopt a set of SAFE-AI metrics grounded in the Cramér–von Mises (CvM) divergence. Let Y and Y' be two random variables with cumulative distribution functions $F_Y, F_{Y'} : \mathbb{R} \to [0, 1]$. The p-order Cramér–von Mises divergence between their distributions is defined as

$$CvM_p(F_Y, F'_Y) = \int_{-\infty}^{\infty} \left|F_Y(u) - F'_Y(u)\right|^p dF_Y(u) \tag{5}$$

In practice, F_Y and $F_{Y'}$ may represent the empirical cumulative distributions of the true outcomes and the corresponding model predictions on a test sample. More generally, the divergence can also compare two predictive distributions obtained from distinct datasets or model architectures. Building on the first-order CvM divergence, we introduce the Rank Graduation (RG) index as

$$\mathrm{RG}_1(Y, Y') = 1 - \frac{CvM_1(F_Y, F'_Y)}{G(Y)}, \tag{6}$$

where $G(Y)$ denotes the Gini coefficient computed from the empirical distribution of the true outcome [9]. An equivalent formulation of the CvM divergence can be expressed as

$$\mathrm{CvM}_p(\mu, \mu^*) = \int_{\mathbb{R}} |F_\mu(t) - F_{\mu^*}(t)|^p \; dF_\mu(t) = \int_0^1 |L_{Y,Z}(t) - L_Y(t)|^p \; d\mu(t), \tag{7}$$

where $L_Y(t)$ denotes the Lorenz curve associated with Y, and $L_{Y,Z}(t)$ denotes the concordance curve between Y and Z.

In the binary setting, when Y represents the true labels and $Y' = \hat{Y}$ the predicted scores, the RG measure coincides with the classical Area Under the ROC Curve (AUC). Unlike AUC, RG naturally extends to ordinal and multiclass outcomes, providing a consistent ranking-based performance measure. For multiclass problems, RG is computed using a one-vs-rest strategy: a score is evaluated for each class by comparing the class indicator with the predicted class probability. The overall multiclass RG is obtained as a convex combination of the per-class scores weighted by the empirical class frequencies. The metric can also quantify different SAFE-AI dimensions depending on the variables (Y, Y') under comparison. When Y is the ground truth and $Y' = \hat{Y}$ the model prediction, it measures accuracy resilience (RGA). When predictions on original data $Y = \hat{Y}$ are compared with predictions on perturbed inputs $Y' = \hat{Y}^{(p)}$, it evaluates robustness (RGR). Finally, when predictions are compared before $Y = \hat{Y}$ and after $Y' = \hat{Y}^{(p)}$ controlled feature modifications, it measures the explainability (RGE). While absolute metric values are not intended for direct comparison across datasets with different task complexity, the framework enables consistent within-dataset model comparison because the RG family operates on predicted probabilities and ground truth labels common to all modalities, with only the perturbation strategy adapting to the data type.

A key strength of the SAFE-AI framework is that RGA, RGR, and RGE are not separate metrics but specific realizations of the same RG family. This common foundation ensures conceptual consistency and enables their aggregation into a unified evaluation framework. Furthermore, it naturally extends to the AUC–based approach, where the integral of each RG curve provides a global summary measure [3]. The resulting areas under the corresponding curves are defined as

$$\mathrm{AURGA} = \int_0^1 \mathrm{RGA}(x)\,dx, \qquad \mathrm{AURGR} = \int_0^1 \mathrm{RGR}(\sigma)\,d\sigma, \qquad \mathrm{AURGE} = \int_0^1 \mathrm{RGE}(x)\,dx \quad (8)$$

AURGA (Area Under the RGA Curve) evaluates ranking stability when progressively removing the most confident predictions. For each removal fraction $x \in [0, 1]$, samples are sorted by prediction confidence, the top x fraction is removed, and RGA is recomputed. *AURGR (Area Under the Robustness Curve)* measures prediction stability under increasing perturbation levels σ obtained through Gaussian noise injection. The predictions are recomputed at each perturbation level and compared with the original output using RGR. *AURGE (Area Under the Explainability Curve)* evaluates structural dependence on input components. Features are progressively removed according to a predefined strategy (greedy importance-based feature removal for tabular data, spatial or token masking for images and text), predictions are recomputed, and the resulting $\mathrm{RGE}(x)$ curve is integrated accordingly. Higher AURGE values indicate that predictions remain stable even after removing a large fraction of features, suggesting that model decisions rely on a relatively small subset of informative variables rather than on many weak contributions. In contrast to local attribution methods such as SHAP or LIME, which explain individual predictions, AURGE provides a global indicator of how strongly model behavior depends on the available input variables.

3. Application

3.1. Data and Preprocessing

We evaluate the proposed framework on three datasets covering tabular, image, and text modalities:

- The first dataset is derived from the 2017 Home Mortgage Disclosure Act (HMDA) records for the state of New York [10]. The data contain detailed information on mortgage loan applications and is used in studies of credit risk and fairness [11]. The target variable is *action_taken*, representing loan denial or approval. During preprocessing, irrelevant variables are removed while categorical attributes (e.g. applicant sex, race, ethnicity, loan type, lien status) are preserved as nominal features, which can be directly handled by the Rulex LLM model.
- The second dataset is the Brain Cancer MRI dataset [12], containing 6,056 MRI scans across three tumor classes: brain glioma, brain meningioma, and brain tumor. Images are provided

at a resolution of 512×512 pixels. Background regions are removed using contour-based thresholding and morphological operations [13], after which images are resized to 224×224 pixels and normalized. Feature extraction is performed using a pretrained ResNet-18 model [14], producing 512-dimensional embeddings used as inputs for the classification models.

- The third dataset is the Financial PhraseBank [15], a benchmark for financial sentiment analysis consisting of short news sentences labeled as positive, neutral, or negative. We use the *sentences_allagree* subset containing only samples with 100% annotator agreement. Text is represented using TF–IDF vectorization [16] with lowercased unigrams and bigrams, limiting the vocabulary to 2,000 features and removing rare terms. Binary weighting is applied so that each feature indicates only the presence or absence of the corresponding term. The resulting TF–IDF matrix is converted to a dense floating-point representation for SAFE-AI metric evaluation.

3.2. Models and Evaluation

Across all datasets, we benchmark the Rulex LLM against a set of widely used classifiers representing linear, kernel-based, tree-based, and neural approaches: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and a Multilayer Perceptron (MLP). In addition, two ensemble strategies are considered: a soft Voting Ensemble Model (VEM) combining RF and XGB, and a Stacking Ensemble Model (SEM) using RF and XGB as base learners with a Logistic Regression meta-learner. Hyperparameter optimization for LR, RF, SVM, and XGB is performed using Optuna [17], maximizing the mean macro-F1 score under an internal stratified cross-validation. The resulting optimal parameters are reused in the outer folds to prevent data leakage. The MLP baseline uses a simple feed-forward architecture with one hidden layer and GELU activation, trained using the Adam optimizer and cross-entropy loss. The best checkpoint within each fold is selected according to validation loss. Model evaluation follows a stratified 5-fold cross-validation protocol. Models are trained on the training split and evaluated on the held-out fold, with results averaged across folds. Standard predictive performance is reported using Accuracy, macro-F1 and Mean Squared Error (MSE) computed from the predicted class probabilities. It is important to note that, since image embeddings and TF–IDF features are not directly semantically interpretable, SAFE-AI metrics in these settings are interpreted primarily as measures of prediction stability under perturbations of the learned representations rather than as indicators of semantic explainability, whereas the tabular HMDA dataset allows a more direct interpretation of feature-removal effects.

4. Results

4.1. HMDA Credit Lending

The results in Table 1 reveal differences between predictive performance and SAFE-AI metrics on the HMDA dataset. Two configurations of the Rulex LLM are reported, corresponding to different rule generalization levels. More specific rules improve predictive fitting, while more general rules reduce predictive performance but increase robustness and explainability. This trade-off is particularly evident for this dataset and appears less pronounced in the image and text datasets considered later.

Considering the standard predictive metrics, MLP achieves the highest accuracy (0.820) and the lowest MSE (0.136), while RF obtains the best F1-macro (0.630), indicating a stronger balance under class imbalance. Other models such as SVM, VEM and XGBoost also reach high accuracy, although SVM shows the lowest F1-macro (0.460). SAFE-AI metrics provide complementary insights into model behavior. RF achieves the highest RGA and AURGA values, while MLP and VEM obtain comparable results in terms of these metrics, indicating stability of predictive rankings under feature removal. In terms of robustness to input perturbations, Logistic Regression shows strong AURGR despite lower predictive performance, highlighting that simpler models can remain stable under noise. Similarly, explainability robustness measured by AURGE is highest for Random Forest and MLP, while SVM appears more sensitive to the removal of informative features. It is important to note that XGBoost

Table 1
Performance comparison across the HMDA, Brain Cancer MRI, and Financial Text datasets using standard and SAFE-AI metrics across 5 folds.

Dataset	Model	Accuracy	F1-score	MSE	RGA	AURGA	AURGR	AURGE
HMDA	LR	0.637	0.568	0.225	0.683	0.249	0.650	0.661
	RF	0.756	0.630	0.169	0.722	0.281	0.521	0.681
	SVM	0.816	0.460	0.146	0.614	0.196	0.475	0.455
	XGB	0.794	0.583	0.156	0.675	0.238	0.430	0.590
	VEM	0.795	0.607	0.147	0.712	0.273	0.491	0.653
	SEM	0.776	0.595	0.197	0.665	0.233	0.417	0.576
	MLP	0.820	0.514	0.136	0.713	0.274	0.551	0.673
	Rulex LLM (Specific)	0.800	0.583	0.152	0.668	0.235	0.658	0.609
	Rulex LLM (General)	0.806	0.527	0.235	0.756	0.157	0.739	0.715
Brain MRI	LR	0.946	0.946	0.027	0.997	0.978	0.886	0.688
	RF	0.938	0.938	0.053	0.999	0.992	0.886	0.721
	SVM	0.983	0.983	0.008	0.999	0.999	0.939	0.717
	XGB	0.957	0.957	0.023	0.999	0.995	0.918	0.740
	SEM	0.956	0.956	0.023	0.999	0.995	0.923	0.741
	VEM	0.957	0.957	0.031	0.999	0.995	0.916	0.742
	MLP	0.959	0.959	0.022	0.998	0.984	0.936	0.676
	Rulex LLM	0.832	0.830	0.084	0.927	0.706	0.733	0.618
Financial Text	LR	0.872	0.826	0.069	0.944	0.720	0.816	0.813
	RF	0.857	0.804	0.086	0.924	0.652	0.904	0.786
	SVM	0.887	0.839	0.059	0.949	0.736	0.882	0.834
	XGB	0.869	0.815	0.068	0.933	0.678	0.847	0.815
	SEM	0.865	0.810	0.071	0.930	0.669	0.913	0.808
	VEM	0.870	0.817	0.074	0.932	0.672	0.908	0.809
	MLP	0.863	0.803	0.068	0.944	0.719	0.805	0.828
	Rulex LLM	0.818	0.760	0.089	0.913	0.632	0.878	0.777

and Rulex LLM (Specific) achieve similar accuracy (0.794 and 0.800), F1-score (0.583 and 0.583), and MSE (0.156 and 0.152), yet Rulex LLM shows substantially higher AURGR (0.658 and 0.430), suggesting that SAFE-AI metrics may favor a different model choice despite comparable standard performance. Focusing on the Rulex LLM, the configuration with more specific rules achieves competitive predictive performance, while the more general configuration improves robustness and explainability, reaching the highest AURGR and AURGE values. Overall, these results highlight how SAFE-AI metrics reveal strong robustness and explainability properties depending on the rule generalization level.

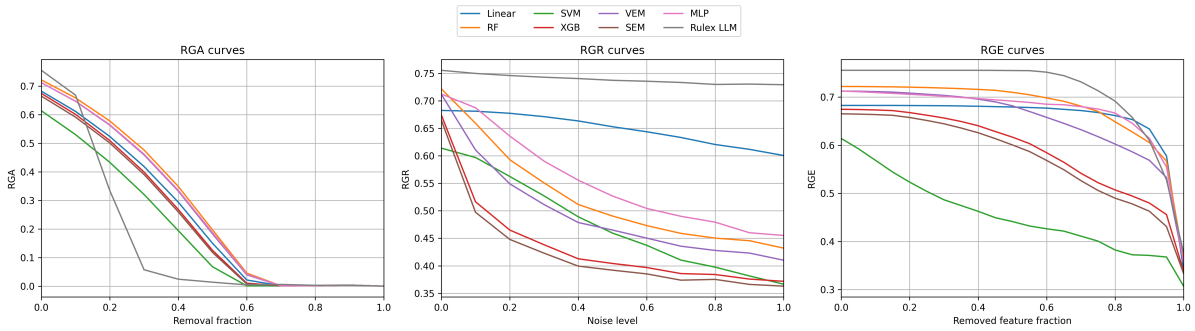

Figure 1: Example of SAFE-AI curves on the HMDA dataset. The figure shows the mean RGA, RGR, and RGE curves computed across the 5-fold evaluation for all models

Figure 1 illustrates example of the SAFE-AI curves, which highlight differences in model stability

under the SAFE-AI evaluation framework. Due to space limitations, SAFE-AI curves are reported only for the HMDA dataset as a representative example. Consistent with the aggregated metrics reported in Table 1, the Rulex LLM configuration with more general rules exhibits stronger robustness patterns, reflected in the slower decrease of the RGR and RGE curves.

4.2. Brain Cancer MRI Images

The results on the Brain Cancer image dataset in Table 1 show very strong predictive performance across all classical models. SVM achieves the highest accuracy and F1-score (0.983), followed by MLP, XGB, SEM, and VEM, all exceeding 0.95. MSE values are extremely low (down to 0.008 for SVM), indicating well-calibrated probability estimates. SAFE-AI metrics confirm this behavior. All classical models obtain high RGA values ($\approx$0.997–0.999) and AURGA is also close to 1.0 for most models, showing that predictions remain stable under data removal. In terms of noise robustness (AURGR), SVM (0.939) and MLP (0.936) perform best, while explainability (AURGE) is highest for VEM (0.742) and SEM (0.741). The Rulex LLM achieves moderate predictive performance (Accuracy = 0.832, F1 = 0.830) and a higher MSE. SAFE-AI metrics further highlight this gap: RGA (0.927) and AURGA (0.706) indicate lower predictive stability, while AURGR (0.733) and AURGE (0.618) show reduced robustness to noise and pixel blurring. This gap can be explained by the nature of the image representation used in the experiment. After feature extraction, images are represented by normalized continuous embeddings that lack direct semantic meaning. Since Rulex LLM relies on rule induction over interpretable variables, these latent features are less suitable for rule extraction, which may limit its performance in this setting.

4.3. Financial Text

On the Financial Text dataset, classical models achieve strong predictive performance with low MSE values in Table 1. SVM obtains the highest accuracy (0.887) and F1-score (0.839), followed by LR (Accuracy = 0.872, F1 = 0.826). Considering SAFE-AI metrics, SVM achieves the strongest overall performance, with the highest RGA (0.949), AURGE (0.834) and AURGA (0.736) scores. LR and MLP show similar high RGA (0.944) and AURGA (0.720 and 0.719), while SEM and VEM obtain the strongest robustness scores (AURGR = 0.913 and 0.908). RF also demonstrates high perturbation stability (AURGR = 0.904). The Rulex LLM shows lower predictive performance (Accuracy = 0.818, F1 = 0.760) and generally lower SAFE-AI scores, although its robustness (AURGR = 0.878) remains comparable to several classical models. This behavior is likely related to the TF–IDF representation used for textual data, which produces high-dimensional token-based features that are well suited for statistical classifiers but less favorable for rule induction. Consequently, rule-based models such as Rulex LLM may be disadvantaged in this setting, similarly to the case of latent image embeddings.

5. Conclusions

The experimental results across tabular (HMDA), image (Brain Cancer MRI), and text (Financial PhraseBank) datasets highlight several consistent observations. First, predictive performance alone does not fully capture model reliability. Models with high accuracy do not necessarily achieve the strongest SAFE-AI scores. This is particularly evident on the HMDA dataset, where robustness and explainability metrics reveal differences not visible from traditional measures such as Accuracy, F1-score, and MSE. Second, model behavior varies across data modalities. On the image dataset, models built on deep feature representations achieve very high predictive stability, while tabular and text datasets reveal clearer trade-offs between predictive performance, robustness, and explainability. Third, the Rulex LLM exhibits domain-dependent behavior. It performs competitively on structured tabular data, where interpretable features enable effective rule induction, but is generally outperformed on image and text datasets with latent or high-dimensional representations. Overall, the results confirm the usefulness of SAFE-AI metrics as a complementary evaluation framework and highlight the importance of considering both model architecture and data representation when assessing interpretable machine learning

approaches. The implementation of the SAFE-AI metrics used in this work is publicly available at https://github.com/koleso500/rulex-llm-safe.

Acknowledgments

We acknowledge funding from the Italian PRIN2020 project FIN4GREEN - Finance for a sustainable, green and resilient society.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull for Overleaf in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Sahoh, A. Choksuriwong, The role of explainable artificial intelligence in high-stakes decision-making systems: a systematic review, Journal of Ambient Intelligence and Humanized Computing 14 (2023) 7827–7843. doi:10.1007/s12652-023-04594-w.

[2] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbrock, Explainable ai in fintech risk management, Frontiers in Artificial Intelligence 3 (2020). doi:10.3389/frai.2020.00026.

[3] P. Giudici, V. Kolesnikov, Safe ai metrics: An integrated approach, Machine Learning with Applications 23 (2026) 100821. doi:https://doi.org/10.1016/j.mlwa.2025.100821.

[4] M. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, R. Munos, The cramer distance as a solution to biased wasserstein gradients (2017). doi:10.48550/arXiv.1705.10743.

[5] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1986) 81–106.

[6] J. R. Quinlan, C4.5: programs for machine learning, Elsevier, 2014.

[7] M. Muselli, Switching neural networks: A new connectionist model for classification, in: Italian Workshop on Neural Nets, Springer, 2005, pp. 23–30.

[8] M. Muselli, E. Ferrari, Coupling logical analysis of data and shadow clustering for partially defined positive boolean function reconstruction, IEEE Transactions on Knowledge and Data Engineering 23 (2011) 37–50. doi:10.1109/TKDE.2009.206.

[9] G. Auricchio, A. E. Bernardelli, P. Giudici, G. Toscani, On rank graduation metrics for high dimensional ordinal data, Mathematical Models and Methods in Applied Sciences (2026).

[10] Consumer Financial Protection Bureau, Home mortgage disclosure act (hmda) data, https://www.consumerfinance.gov/data-research/hmda/, 2017.

[11] G. Babaei, P. Giudici, L. Wu, Explainable fairness in mortgage lending, in: R. Guidotti, U. Schmid, L. Longo (Eds.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2026, pp. 378–398.

[12] M. M. Rahman, Brain cancer – mri dataset, 2024. doi:10.17632/mk56jw9rns.1.

[13] Z. Li, O. Dib, Empowering brain tumor diagnosis through explainable deep learning, Machine Learning and Knowledge Extraction 6 (2024) 2248–2281. doi:10.3390/make6040111.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[15] P. Malo, A. Sinha, P. Takala, P. Korhonen, J. Wallenius, Good debt or bad debt: Detecting semantic orientations in economic texts, Journal of the American Society for Information Science and Technology (2014). doi:10.1002/asi.23062.

[16] J. Ramos, Using tf-idf to determine word relevance in document queries (2003).

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.