S.A.F.E. Artificial Intelligence

Paolo Giudici

Professor of Statistics, University of Pavia
Coordinator of the EU projects FIN-TECH and PERISCOPE
P.I. of PNRR GRINS SUST-AI and of PRIN FIN4GREEN

Proposal

Problem: AI safety is a key priority; however, there are not yet consistent metrics that can measure it

We propose a S.A.F.E. measurement model for the governance of Artificial Intelligence applications, based on four main principles:

- Security: measures the robustness of AI systems
- Accuracy: measures the truthfulness of AI systems
- Fairness: measures the inclusiveness of AI systems
- Explainability: measures the controllability of AI systems

For the measurement, we propose four consistent metrics, that extend the Area Under the ROC Curve (AUC) to all types of response variables, leveraging the properties of the Lorenz curve and of the Gini index.

The Gini index

A measure of inequality for a variable Y:

$$G(Y) = \frac{E|Y_i - Y_j|}{2|E(Y)|}$$

Also a measure of variability, alternative to the coefficient of variation:

$$CV(Y) = \frac{\sqrt{E(Y_i - Y_j)^2}}{|E(Y)|}$$

Also a measure of classification accuracy in machine leraning:

$$G(\hat{Y}) = 1 - 2 * AUC(\hat{Y}),$$

Gini index and Concordance

Let Y be $i = 1, \ldots, n$ response values to be predicted and let:

- Lorenz curve (L_Y): $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, where r_j indicates the increasing ranks of Y and $\bar{y}$ the mean of Y.
- dual Lorenz curve: $(L_Y^{'})$: $(i/n, \sum_{j=1}^{i} y_{r_{n+1-j}}/(n\bar{y}))$, where r_{n+1-j} indicates the decreasing ranks of Y.
- Lorenz Zonoid: area between L_Y and $L_Y^{'}$, corresponding to the Gini index.

Definition

A concordance curve (C) can be defined by: $(i/n, \sum_{j=1}^{i} y_{\hat{r}_j}/(n\bar{y}))$, where $\hat{r}_j$ indicates the increasing ranks of $\hat{Y}$.

Gini index and Concordance

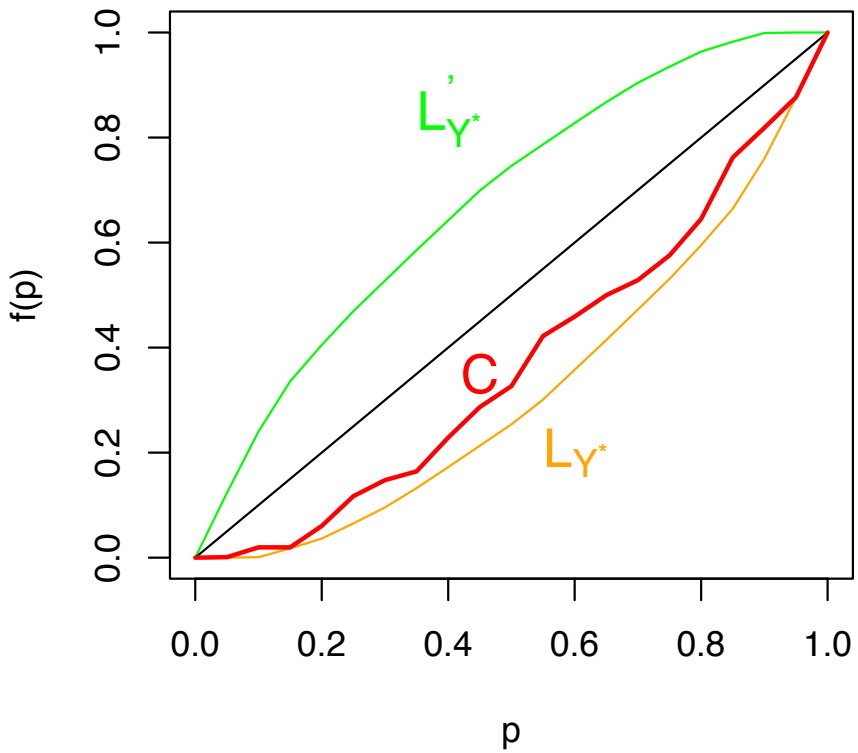


Figure: The L_Y and L'_Y Lorenz and dual Lorenz curves and the C concordance curve, where p and $f(p)$ are the cumulative values of the x and y coordinates of the L_Y, L'_Y and C curves.The area between the Lorenz curves is the Gini index.

The Rank Graduation Accuracy metric (RGA)

- Dividing the area between the concordance curve and the dual by its maximum value (the Lorenz Zonoid) we obtain:

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}}.$$

- It can be shown that: $0 \leq \text{RGA} \leq 1$, with RGA$= 1$ for a perfectly concordant model; RGA$= 0$ for a perfectly discordant model; RGA$= 0.5$ for random predictions;
- When the response Y is binary, RGA=AUC (Giudici and Raffinetti, 2025);
- RGA can however be calculated for all types of response variables: categorical, continuous, and multidimensional.

RGA as a Cramer distance

Definition

Given any two cumulative distribution functions $F, G : \mathbb{R} \to [0,1]$, the Cramer - Von Mises divergence between F and G is

$$\mathrm{CvM}(F,G) = \int_{-\infty}^{\infty} \left|F(u) - G(u)\right| dF(u).$$

Theorem

Let F_Y and $F_{\hat{Y}}$ be the cumulative distributions of Y and $\hat{Y}$. Then:

$$\mathrm{RGA}(Y,\hat{Y}) = 1 - \frac{\mathrm{CvM}(F_Y, F_{\hat{Y}})}{G(Y)}.$$

This allows significance tests via the Cramer-Von Mises statistics.

RGA and R^2

Definition

For any type of response, we can measure accuracy with:

$$\mathrm{RGA} = 1 - \frac{\mathrm{CvM}(F_Y, F_{\hat{Y}})}{G(Y)}.$$

Definition

If the response is continuous we can, in addition, measure accuracy with the predictive R^2:

$$\mathrm{R}^2 = 1 - \frac{MSE(\hat{Y}, Y)}{Var(Y)}.$$

RGX: a Rank Graduation Box

Definition

The mathematical derivation of RGA can be extended to all AI principles by means of a Rank Graduation Box (Babaei, Giudici and Raffinetti, 2025). Define:

$$\mathrm{RGX}(Y,Z) = 1 - \frac{\mathrm{CvM}(F_Y, F_Z)}{G(Y)}.$$

Different pairs (Y,Z) lead to different metrics:

- $(\hat{Y}, \hat{Y}^p)$ leads to RGR, which measures robustness against adversarial perturbations p;
- $(\hat{Y}, \hat{Y^{-k}})$ leads to RGE, which measures the explainability gain attributed to variable(s) k;
- $(\hat{Y^A}, \hat{Y^B})$ leads to RGF, which measures fairness between groups A and B.

Application: accuracy of Large Language Models

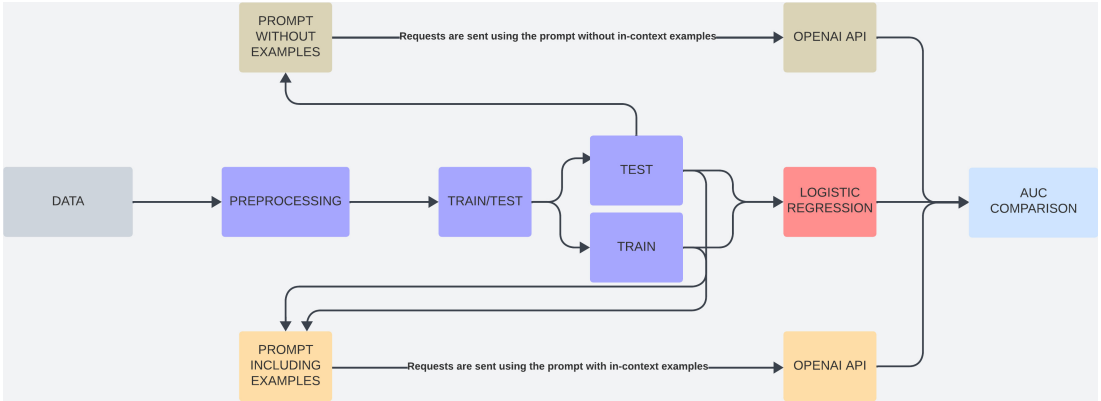


Figure: Generative AI models (Babaei and Giudici, 2024)

Method	Min RGA	Max RGA	Mean RGA	Std RGA
LR (30,000 ex.)	0.7018467852	0.7950752394	0.75318741451	0.0314667993
GPT(0 ex.)	0.5895348837	0.6445964432	0.61264021887	0.0212562962
GPT (100 ex.)	0.6260601915	0.6963064295	0.66655266764	0.0261630545

Application: robustness of NLP models

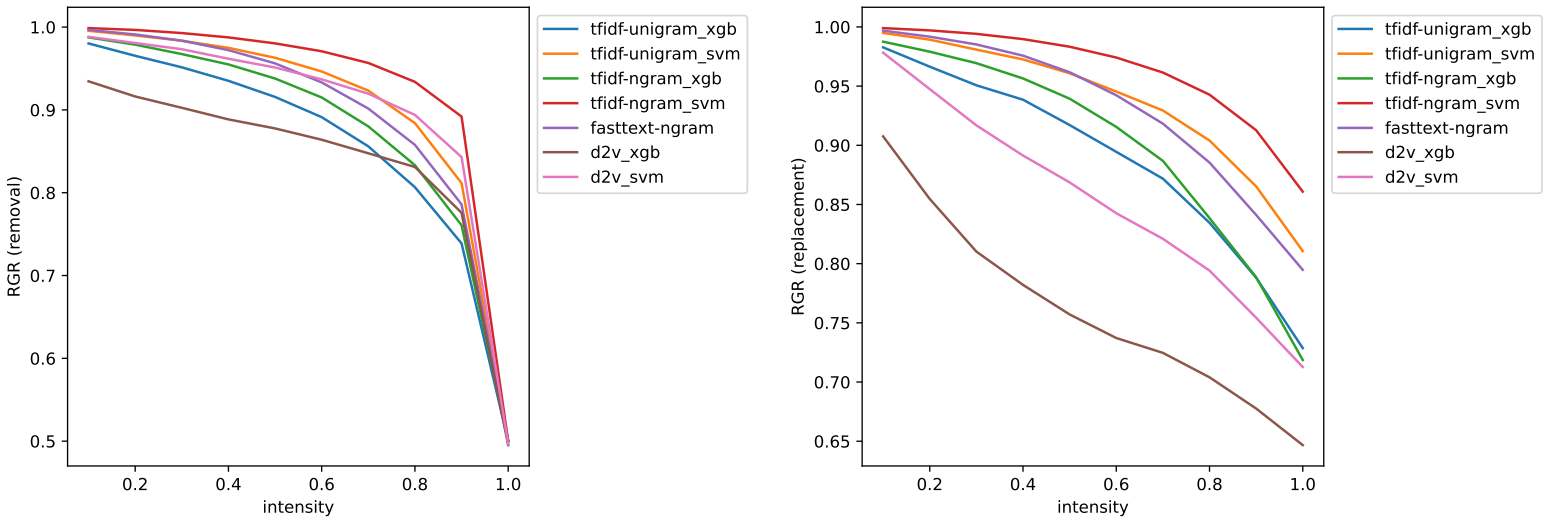


Figure: Robustness of about 68,000 decisions based on a text document, using word removal (left) and word replacement (right), with a growing intensity, for different AI models, based on different Natural Language Processing methods (Babaei, Giudice, Giudici and Maggi, 2025)

Application: explainability of time series

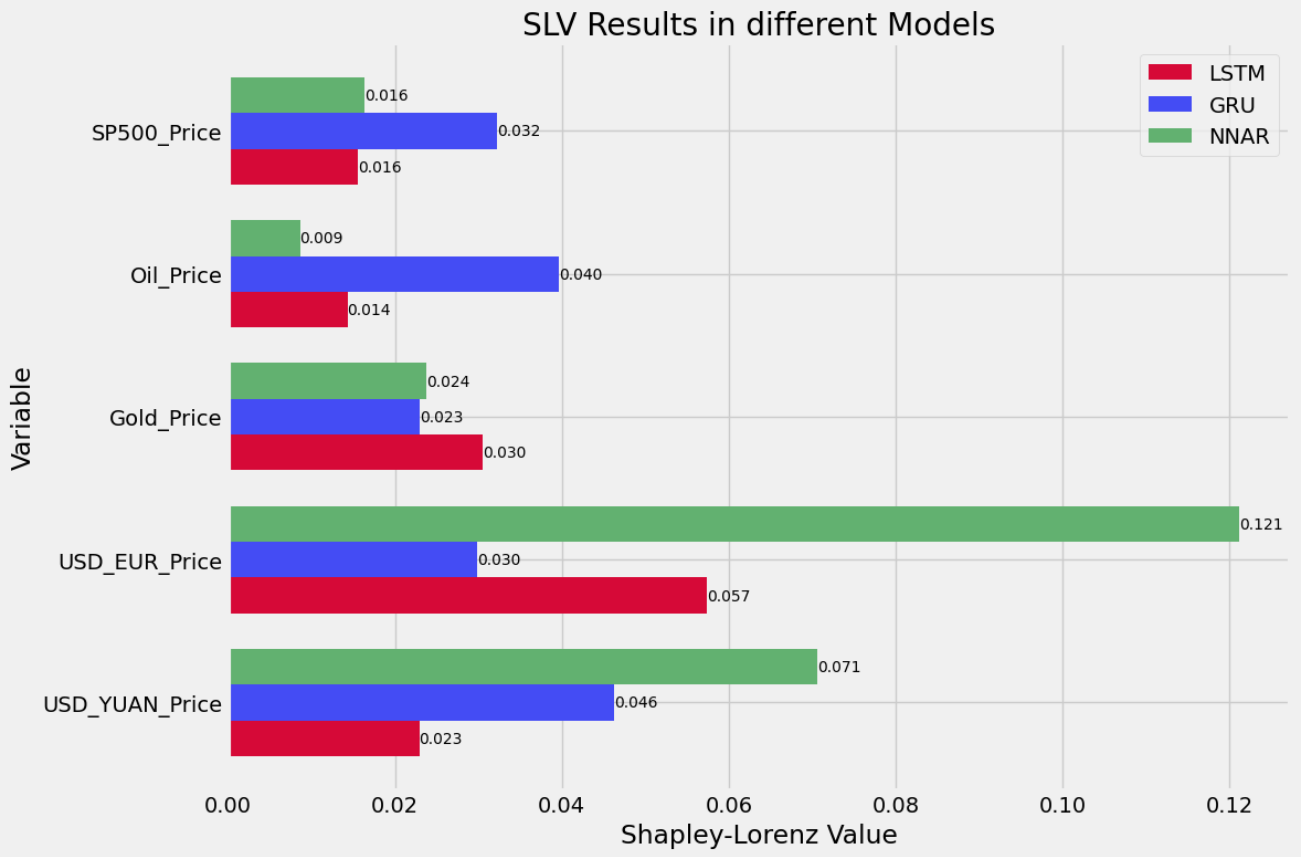


Figure: Model explanations for three neural networks: (NNAR, LSTM, GRU). The Shapley Lorenz values are percentages of explained accuracy.(Giudici, Piergallini, Raffinetti and Recchioni, 2024)

Multivariate extension

Auricchio, Giudici and Toscani (2024) define a multidimensional Gini index which is equivalent to:

$$G(Y) = \sum_{\ell=1}^{d} \lambda_\ell G(Y_\ell^*),$$

where Y_ℓ^* is the ℓ-th component of a whitened version of Y, with mean m_ℓ^*, and $\lambda_\ell = \frac{m_\ell^*}{\sum_{\ell=1}^{d} m_\ell^*}$.

A set of sliced *RGX* metrics can then be defined as:

$$\mathrm{SRGX}(Y, Z) = \sum_{\ell=1}^{d} \lambda_\ell \mathrm{RGX}(Y_\ell, Z_\ell),$$

Application: prediction of E,S,G sustainability

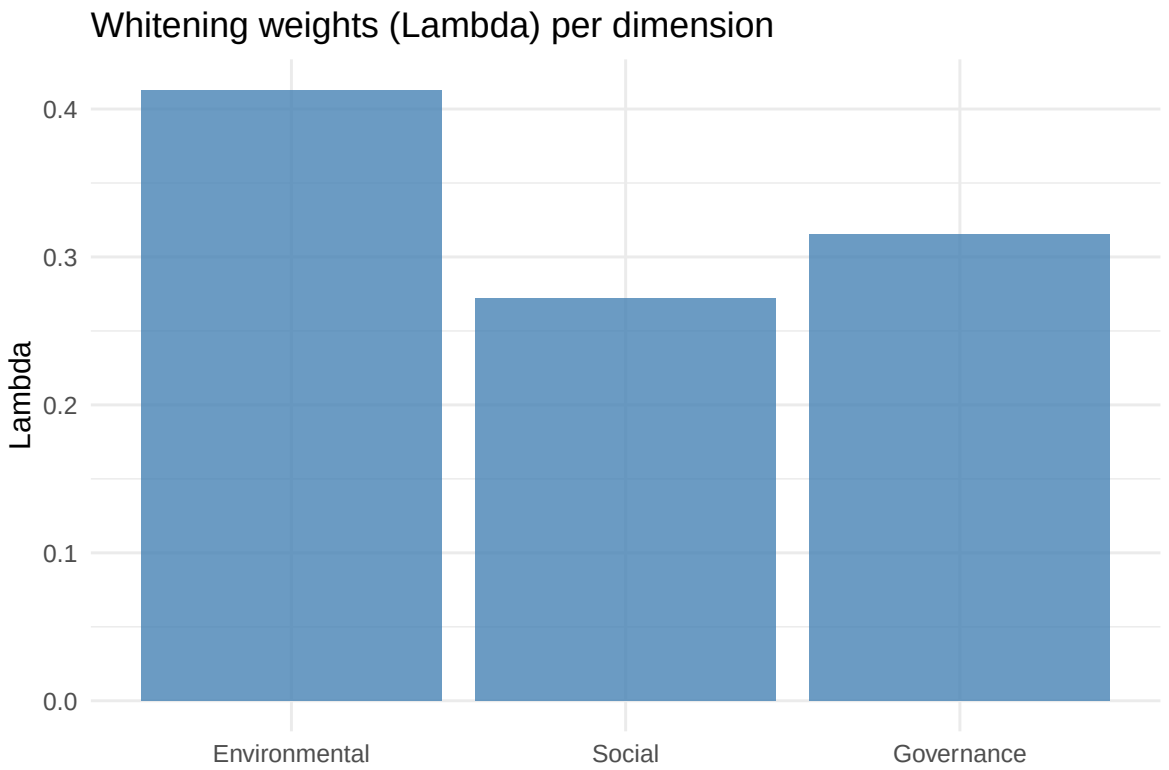


Figure: Weights (λ_j) of whitened ESG dimensions obtained through the whitening transformation

Application: prediction of E,S,G Sustainability

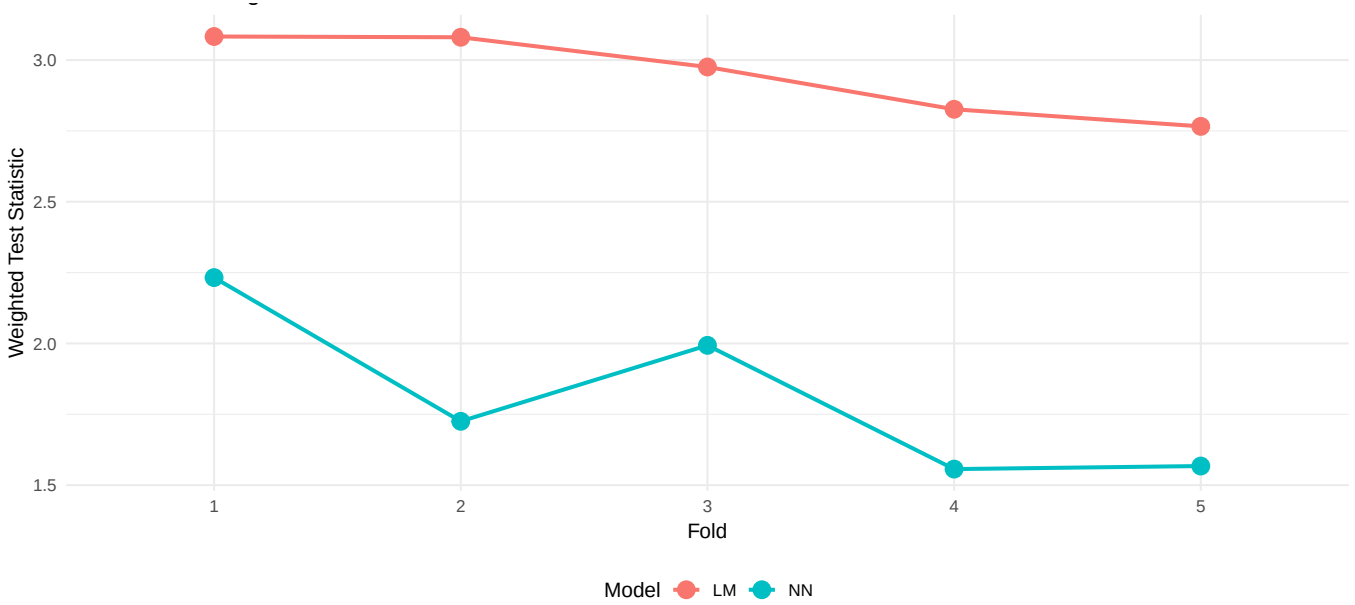


Figure: Cramer-Von-Mises Test Statistics by Fold for multivariate Linear Regression (LM) and Neural Network (NN) models applied to the prediction of Environmental,Social and Governance scores. The results for each fold are weighted by λ_i across the three response dimensions: E,S,G.

SAFE AI assessment (ex-post)

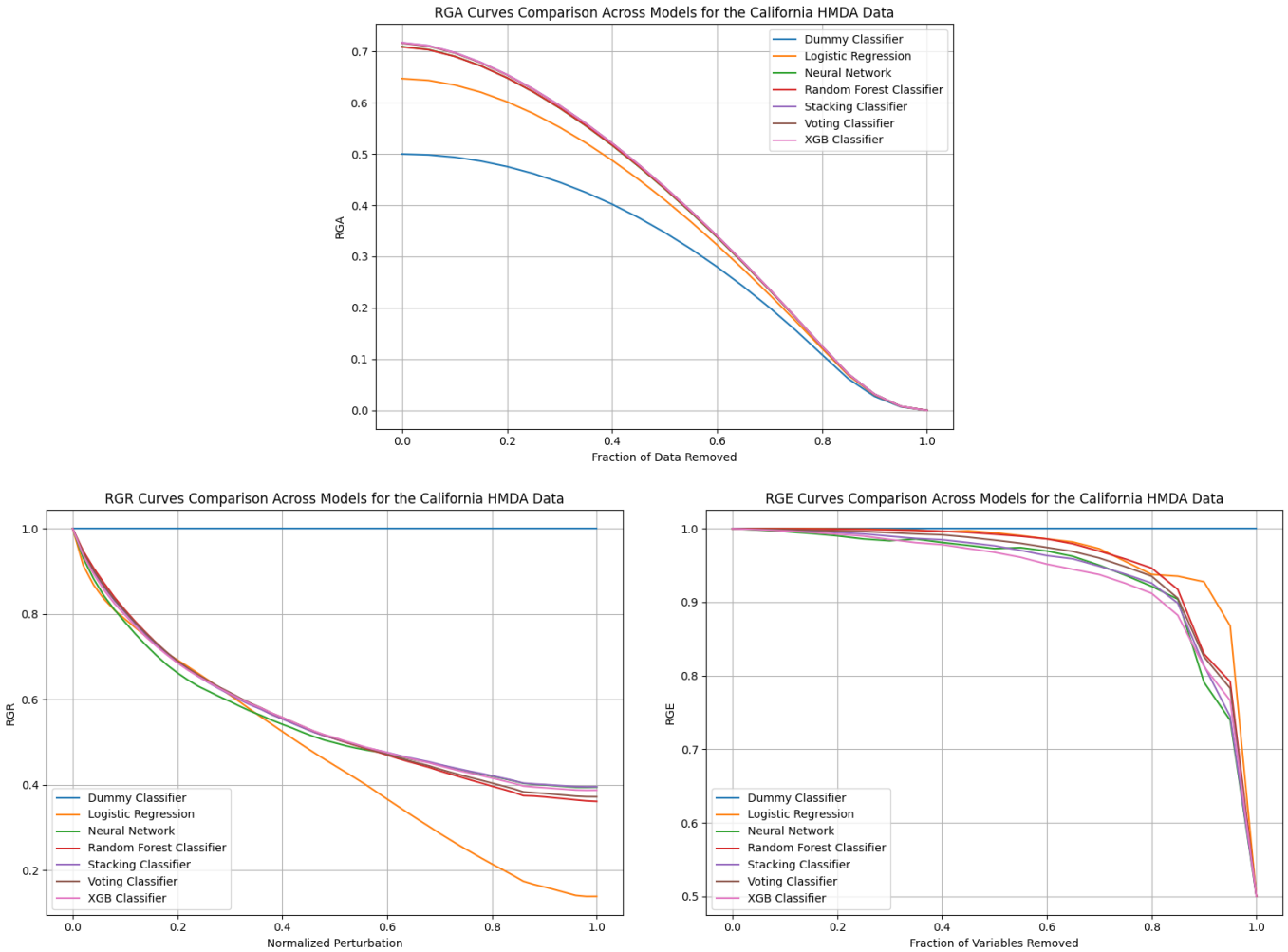


Figure: Comparison of the "resilience" of different machine learning models in a binary classification problem terms of RGA, RGR, RGE (Giudici and Kolesnikov, 2026).

SAFE AI assessment (ex-post)

Model	Sust.	Acc.	Exp.	Prob	RMSE
MLP	0.9661	0.4518	0.5114	0.7768	0.1046
RBF	0.9538	0.4519	0.5443	0.75454	0.0982
NNAR	0.7157	0.3718	0.2405	0.9361	0.1358
LSTM	0.9607	0.8186	0.1122	0.9118	0.0561
GRU	0.9244	0.8865	0.1778	0.8543	0.0439

Table: AI metrics integration: comparison of MLP, RBF, NNAR, LSTM and GRU models to predict a prices, in terms of three S.A.F.E. AI metrics, in comparison with the "classic" Root MSE metric.

SAFE Agentic AI (ex-ante)

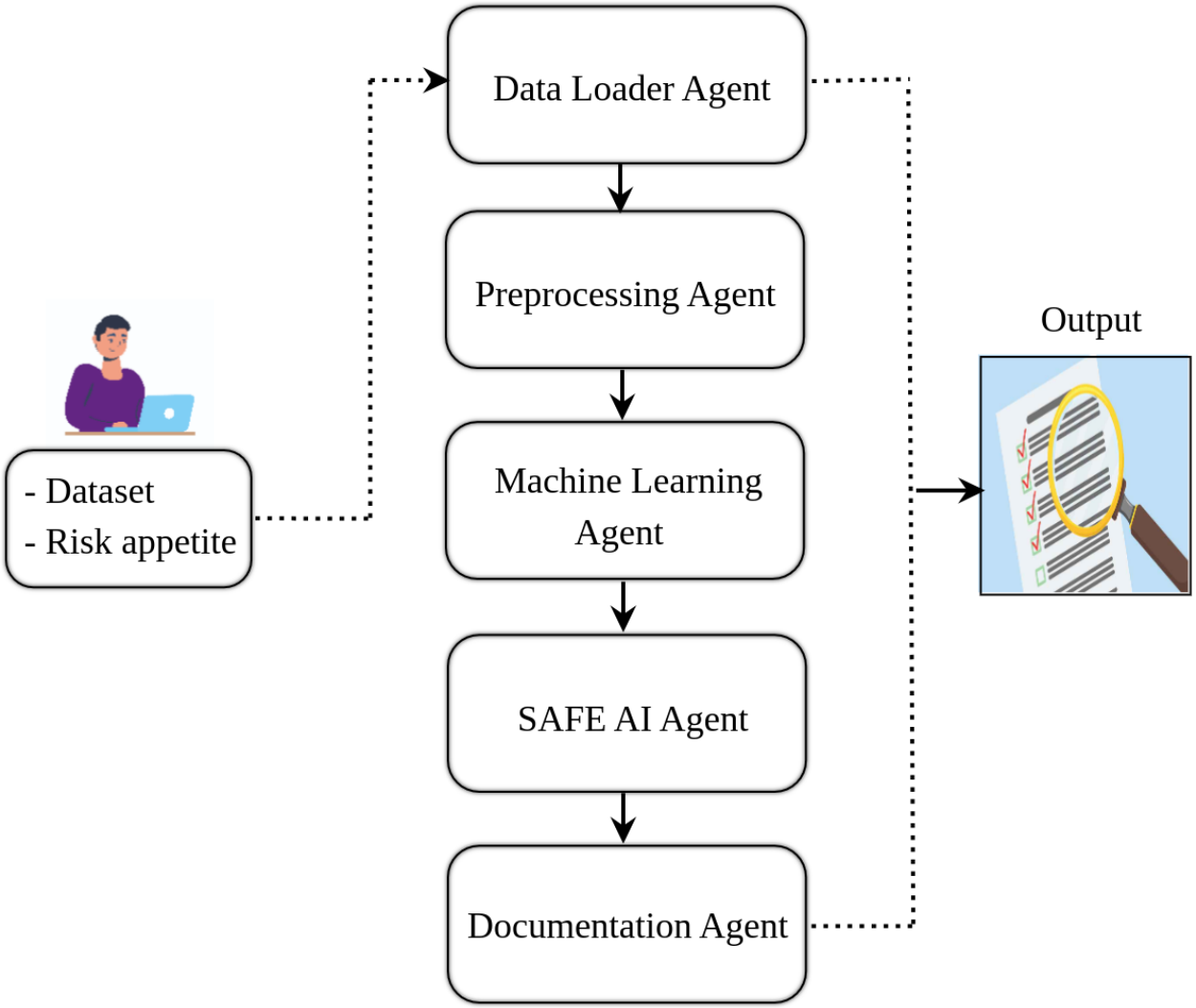


Figure: SAFE Agentic Systems. (Babaei, Giudici, Piergallini, Zieni, 2026)

Agents' report

- *The computed RGA score is* 0.6875*. This significant score is above the acceptable threshold of* 0.6*, indicating that the model exhibits a strong alignment with true labels...*
- *...The computed RGR score is* 0.91*. This indicates the model's strong performance in making predictions, regardless of variations in the data...*
- *...The Highest Explainability is for Gender (0.1673) and Loan Purpose (0.1106). The Lowest Explainability is for Race (0.0001), indicating a lack of understanding of the influence of this feature on model predictions....*
- *...In summary, the model displays a high level of accuracy and robustness across all features, making it a reliable choice for classification tasks. However, there is an evident disparity in explainability across the features, which suggests a need for further investigation into how different features contribute to model predictions to enhance transparency.*

SAFE quantum machine learning: problem

We have applied a variational Quantum Circuit based on a 9-qubit amplitude encoding representation of the output of a ResNet neural network on 6056 512*512 pixel MRI images aimed at predicting cancer type.

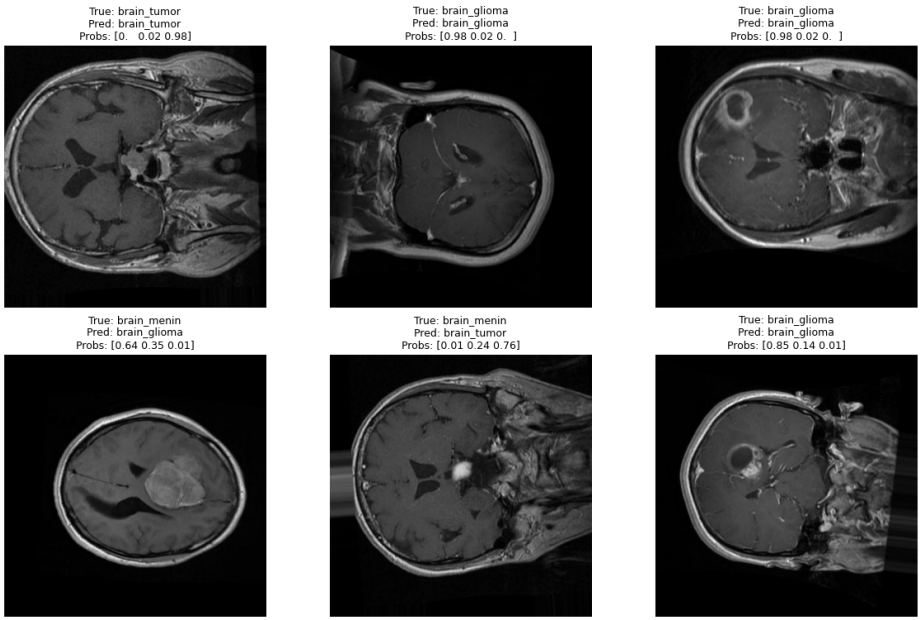


Figure: **Quantum-based tumor classification on unseen MRI scans.** Each image shows the true label, predicted class, and the model softmax probabilities.

SAFE quantum machine learning: encoding

The standardized feature vector $z \in \mathbb{R}^{512}$ is normalized

$$z \longrightarrow x = \frac{z}{\|z\|},$$

and then embedded into a quantum state using a 9-qubit amplitude encoding:

$$|\psi(x)\rangle = \sum_{i=0}^{511} x_i |i\rangle.$$

.

SAFE quantum machine learning: architecture

- The quantum state is processed by a variational quantum circuit with a Strongly Entangling Layer acting on 9 qubits.
- The layer contains single-qubit rotation gates, each associated with three variational parameters (the angles α_i, β_i, and γ_i defining the rotation R applied to the ith qubit), followed by a fully connected pattern of entangling CNOT gates.
- The circuit measures the expectation values of the Pauli-Z operator on each qubit:

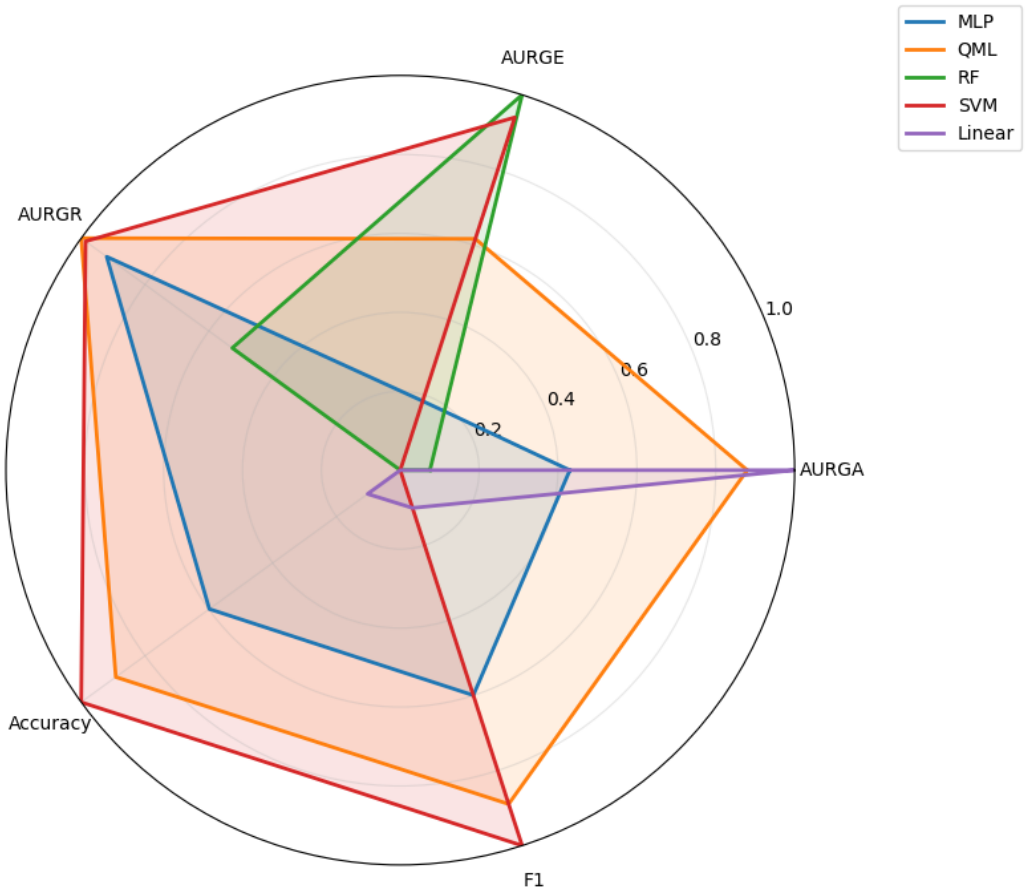
$$q_i = \langle \psi_{\text{out}} | Z_i | \psi_{\text{out}} \rangle, \quad q_i \in [-1, 1].$$

- The feature vector q is finally mapped to class logits:

$$y = W_c q + b_c.$$

262k trainable parameters, learned minimising a cross-entropy loss.

SAFE quantum machine learning: results



- **VQC**: most robust and most balanced SAFE profile
- RF: strongest AURGE, weak elsewhere
- SVM: best Accuracy/F1, low AURGA (resilience to data removal)
- Linear: strong AURGA.

Essential References


Giudici P, Raffinetti E. (2025).RGA: a unified measure of predictive accuracy. *Advances in data analysis and classification, 19, 67â93*


Babaei, G., Giudici P, Raffinetti E. (2025). A Rank Graduation Box for SAFE AI *Expert Systems with Applications, 259, 125239*


Babaei, G. and Giudici, P. (2024).GPT classification, with application to credit lending *Machine learning with applications, 16, 100534*


Giudici, P., Piergallini, A., Raffinetti, E., Recchioni M.C. (2024). Explainable AI methods for financial time series. *Physica A: statistical mechanics with applications, 2024, 655, 130176.*


Babaei, G., Giudice, O., Giudici, P., Maggi, A. (2025). SAFE Natural Language Processing *IEEE Proceedimgs of the IJCNN conference*


Auricchio, G., Giudici, P. , Toscani, G. (2024). Extending the Gini index to higher dimensions via whitening processes. *Rendiconti Lincei, 35 (3).*


Giudici P, Kolesnikov, V. (2026). SAFE AI metrics: an integrated approach. *Machine Learning With Applications.*


Babaei, G., Giudici, P., Piergallini, A., Zieni, R. (2026). SAFE Agentic AI systems. *IEEE proceedings of the first Agentic AI conference.*