Proceedings of Machine Learning Research vol 284:1–3, 2025 19th Conference on Neurosymbolic Learning and Reasoning

Evolutionary Surrogate-Assisted Prescription: Neuro-Symbolic Framework for Trustworthy Decisioning

Hormoz Shahrzad HORMOZ@COGNIZANT.COM; HORMOZ@CS.UTEXAS.EDU
Risto Miikkulainen RISTO@COGNIZANT.COM; RISTO@CS.UTEXAS.EDU
Cognizant AI Lab, San Francisco, CA & The University of Texas at Austin

Editors: Leilani H. Gilpin, Eleonora Giunchiglia, Pascal Hitzler, and Emile van Krieken

Abstract

Abundant records now link organizational contexts, actions, and outcomes. Evolutionary Surrogate-Assisted Prescription (ESP) converts such data into trustworthy policies through a two-stage neuro-symbolic pipeline: a neural network **Predictor** surrogate is trained first using supervised learning, after which an interpretable **Prescriptor** is evolved against it using the EVOTER rule grammar. Decoupling prediction from prescription yields high sample-efficiency, low on-line risk, and explicit regularization, while the resulting rule sets remain compact and auditable. Across diverse, safety-critical domains, ESP attains accuracy on par with—or exceeding—neural network models, yet retains full transparency, establishing a robust platform for large-scale, trustworthy decision optimization.

1. Introduction

Deep networks excel at accuracy but obscure their reasoning— which is unacceptable in regulated areas such as health, finance, and law. Post-hoc methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) offer only local or approximate insight and can mislead (Linardatos et al., 2020; Lage et al., 2019). ESP (Francon et al., 2020) takes an alternative route: first learn an accurate neural network surrogate of the environment, then evolve a fully symbolic, transparent policy, resolving the accuracy–interpretability tension highlighted by Doshi-Velez and Kim (2017).

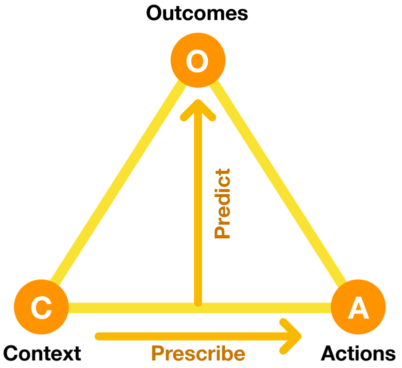


Figure 1: *Elements of ESP.* A neural network **Predictor** maps context–action pairs to outcomes, allowing sample-efficient evolution of an interpretable **Prescriptor**.

2. ESP

Figure 1 summarises the two-stage workflow.

Stage 1: Learning the Predictor. Given data $\mathcal{D} = \{(\mathbf{x}_i, a_i, y_i)\}$ a function $\hat{f}(\mathbf{x}, a) \approx y$ is learned e.g. with gradient-boosted trees or neural networks.

Stage 2: Evolving the Prescriptor. An evolutionary algorithm searches a symbolic space Π of rule sets, maximising expected surrogate outcome $\mathbb{E}_{\mathbf{x}}[\hat{f}(\mathbf{x}, \pi(\mathbf{x}))]$ while respecting safety or cost constraints. The EVOTER grammar (Section 3) ensures each individual is a compact, auditable policy.

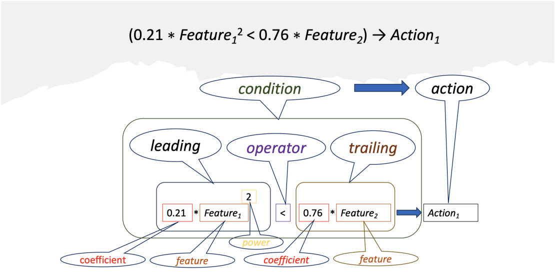
Such an offline optimization approach lets ESP enforce fairness and budget constraints during search, achieving *trust by construction* rather than by post-hoc explanation.

3. EVOTER: The Symbolic Backbone of ESP

EVOTER (Figure 2; Shahrzad et al., 2025) evolves intrinsically interpretable rule sets through a *flat*, list-based grammar enriched with time-lag operators, feature–feature comparisons, and nonlinear transformations. This design simplifies variation operators, mitigates bloat, and captures relational and temporal patterns essential in applications, such as early sepsis detection and reinforcement learning.

$$
\begin{array}{rcl}
< rules > & ::= & < rule > | < rule > \textbf{OR} < rules > \\
< rule > & ::= & < conditions > \longrightarrow [< certainty >] < action > \\
< conditions > & ::= & < condition > | < condition > \textbf{AND} < conditions > \\
< condition > & ::= & < leading > < operator > < trailing > \\
< leading > & ::= & < coefficient > * < feature >^{[<power>]} [(lag)] \\
< coefficient > & ::= & 0.[0-9]+ \\
< feature > & ::= & \text{(input feature)} \\
< power > & ::= & [1-3] \\
< lag > & ::= & [0-9] \\
< operator > & ::= & = | \neq | < | \leq | > | \geq \\
< trailing > & ::= & < leading > | < value > \\
< value > & ::= & \text{(scalar)} \\
< certainty > & ::= & < coefficient > \\
< action > & ::= & \text{(class or action)}
\end{array}
$$

(a) BNF grammar of EVOTER rules



(b) Example EVOTER rule

Figure 2: *Rule-set representation in EVOTER.* (a) BNF with time lags, feature comparisons, and exponents. (b) Colour-coded example rule. A rule set contains multiple such rules, a default fallback, and usage counters.

4. Experimental Evaluation

Time-Series Prescription. On a blood-pressure time-series benchmark for early sepsis detection (Hemberg et al., 2014), ESP evolved a two-clause rule `IF (BP_mean_t-2 < 60) & (HR/BP_mean > 1.2) THEN Alert` that lowered the false-negative rate by 8.2% relative to a clinically tuned baseline while remaining immediately interpretable to medical staff.

Diabetes Treatment Insights. On the *Diabetes10Y* dataset (Strack et al., 2014), ESP evolved a Pareto set of treatment rules; the top policy achieved 99 % "no-readmission" and "sent-home" outcomes (vs. 60 % / 78 % historically) and exposed a race-conditioned clause that clinicians could audit or remove.

Transparent vs. Opaque Prescriptors. On heart-failure (Chicco and Jurman, 2020) and two-objective diabetes (Shahrzad et al., 2025) benchmarks, rule sets scored within 0.1 % of neural networks (non-significant for heart failure, marginal for diabetes), indicating that ESP's transparency comes at virtually no accuracy cost.

5. Conclusion

ESP couples a neural network **Predictor** with an rule-set **Prescriptor**, delivering performance similar to neural networks while staying transparent and editable. The prescriptor can first be evolved as a neural network for speed, then *distilled* into concise rules for deployment. Ongoing work on GPU-scale evolution and continual retraining will extend ESP to larger datasets and non-stationary environments, further broadening its industrial reach.

References

Davide Chicco and Giuseppe Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1):1–16, 2020.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702:08608*, 2017.

Olivier Francon, Santiago Gonzalez, Babak Hodjat, Elliot Meyerson, Risto Miikkulainen, Xin Qiu, and Hormoz Shahrzad. Effective reinforcement learning through evolutionary surrogate-assisted prescription. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 814–822, 2020.

Erik Hemberg, Kalyan Veeramachaneni, Babak Hodjat, Prashan Wanigasekara, Hormoz Shahrzad, and Una-May O'Reilly. Learning decision lists with lags for physiological time series. In *Third Workshop on Data Mining for Medicine and Healthcare, at the 14th SIAM International Conference on Data Mining*, pages 82–87, 2014.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *arXiv:1902.00006*, 2019.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Hormoz Shahrzad, Babak Hodjat, and Risto Miikkulainen. EVOTER: Evolution of transparent explainable rule-sets. *ACM Trans. Evol. Learn. Optim.*, 5(2), May 2025. doi: 10.1145/3702651.

Beata Strack, Jonathan Deshazo, Chris Gennings, Juan Luis Olmo Ortiz, Sebastian Ventura, Krzysztof Cios, and John Clore. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.