Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 381–386
DOI: 10.15439/2025F5371
ISSN 2300-5963 ACSIS, Vol. 43

Symbolic vs Black-Box Explanations: A Model-Driven Approach Using Grammatical Evolution

Dominik Sepioło, Antoni Ligęza
0000-0001-7746-3781, 0000-0002-6573-4246
AGH University of Krakow,
Department of Applied Computer Science,
al. Adama Mickiewicza 30, 30-059 Kraków, Poland
Email: {sepiolo, ligeza}@agh.edu.pl

***Abstract*—Black-box explainability tools like LIME and SHAP are widely used to interpret machine learning models. However, their post-hoc, local nature often results in inconsistent and semantically opaque explanations. This paper presents a model-driven explainability approach using grammatical evolution (GE), enabling the discovery of symbolic, human-readable models. We compare black-box explanations to symbolic GE-generated models on two benchmark tasks: a quadratic equation classification problem and the Iris dataset. GE produces interpretable, consistent, and semantically meaningful expressions consistent with domain knowledge, offering a more trustworthy foundation for explainable AI. The use of Meaningful Intermediate Variables (MIVs) further improves the clarity and expressiveness of the symbolic models.**

***Keywords*—Explainable Artificial Intelligence (XAI); Grammatical Evolution; Symbolic Regression; Model-Driven XAI; Black-Box Explanations; Post-Hoc Explainability; Transparent AI Models**

I. Introduction

THE INCREASING deployment of machine learning (ML) systems in critical domains has renewed attention to the need for transparency, interpretability, and trust in artificial intelligence (AI) models. In practical applications, such as biomedical diagnostics and regulatory environments, it is not sufficient for a model to produce a correct prediction; it must also be able to explain *how and why* it provided a decision in a way understandable to human users [1], [2].

Explainable Artificial Intelligence (XAI) aims to make AI systems more interpretable and trustworthy [1], [3]. XAI methods include inherently transparent models (e.g. decision trees) and post hoc techniques such as LIME [4] and SHAP [5], which approximate black-box models locally. However, post-hoc explanations are often inconsistent, sensitive to sampling, and lack semantic grounding in the domain [6], [7].

These challenges have led to a growing interest in *Model-Driven* approaches to XAI, which construct interpretable models directly from data using symbolic techniques [8], [9]. Among them, Grammatical Evolution (GE) is a promising evolutionary algorithm that evolves human-readable expressions guided by context-free grammars [10]. GE has shown effectiveness in symbolic regression, functional modeling, and equation discovery - especially when domain knowledge is embedded through grammar design or Meaningful Intermediate Variables[1] (MIV) [11], [13].

This paper presents a comparative exploration of symbolic model-driven explanations using GE and black-box post-hoc tools. We argue that symbolic models offer greater semantic clarity, stability, and alignment with human reasoning. Using two benchmarks, a synthetic quadratic classification task and the Iris dataset, we show how GE produces simple, meaningful models that either recover known functional relationships (e.g. $d = b^2 - 4ac$) or leverage interpretable MIVs (e.g. petal area) to achieve high accuracy with full transparency.

Our main contributions are as follows:

- We contrast symbolic and black-box explanations on two classification tasks, highlighting the strengths and weaknesses of each approach.
- We show that GE provides clear, interpretable models aligned with domain knowledge, unlike the often unstable outputs of LIME and SHAP.
- We demonstrate the use of MIVs to enhance interpretability and simplify the model structure.
- We propose a methodology for integrating GE into a transparent MD-XAI pipeline.

The remainder of the paper is organized as follows. Section 2 outlines the background and methods. Section 3 presents experiments. Section 4 discusses the results and implications, and Section 5 concludes with insights and future work.

II. Related Work and Methods

This section outlines the two primary paradigms of explainability considered in this study: post-hoc explanation methods for black-box models, and model-driven symbolic approaches based on Grammatical Evolution. We discuss their principles, strengths, and known limitations, focusing on their impact on semantic clarity, model transparency, and alignment with domain knowledge.

[1] A Meaningful Intermediate Variable (MIV) has clear semantics, is derived from simple functions, and depends on data or other MIVs.

Topical area: Advanced Artificial Intelligence in Applications

A. Black-Box Explanation Techniques

Post-hoc explainability interprets black-box models (e.g. ensembles, neural networks) by analyzing the output without accessing internal structure. Two widely used methods are LIME [4] and SHAP [5]:

- **LIME** fits a local linear surrogate model by perturbing input data to explain predictions.
- **SHAP** uses game theory to calculate feature contributions by calculating Shapley values.

Although popular for their visual intuitiveness, LIME and SHAP face intrinsic key limitations [7], [3], [6]:

- *Instability:* Results vary with random seeds or sampling.
- *Locality:* Explanations are instance-specific and are not generalizable.
- *Poor semantics:* Attributions lack domain or causal grounding.
- *Explainer opacity:* The logic behind them is often difficult to validate.

These weaknesses have motivated the development of model-driven alternatives that prioritize transparency and interpretability by discovering and designing structures [15], [7], [8].

B. Symbolic Regression with Grammatical Evolution

Symbolic regression seeks interpretable mathematical expressions that fit the data accurately. Unlike classical regression with fixed structures (e.g. linear, polynomial), it explores a broad space of formulas composed of variables, constants, and operators, producing models that are directly interpretable and verifiable by humans.

Model-driven explainability seeks to construct models that are inherently interpretable. Among symbolic machine learning techniques, GE provides a flexible and powerful approach to discovering human-readable models [10].

GE is a genetic programming technique that uses context-free grammars to define valid symbolic expressions. Solutions are encoded as genomes that are decoded into mathematical formulas or rules based on the grammar.

- **Expressive Power:** GE supports arithmetic, logic, and domain-specific operations via customizable grammars.
- **Interpretability:** The output models are human-readable symbolic expressions.
- **Incorporation of Domain Knowledge:** Grammar rules or MIVs embed prior knowledge into the search space.

This approach has been successfully applied in various domains, including bioinformatics [9], symbolic regression [10], [12], and model-driven XAI [11], [8], [13]. GE discovers interpretable functions that reflect the data structure while allowing integration of constraints and semantic guidance.

Using user-defined grammars, GE can enforce domain-specific constraints or promote particular structural patterns. For instance, intermediate variables or known transformations can be explicitly included in the grammar, guiding the search process toward semantically valid and explainable models. This makes GE especially suitable for applications that require compliance with expert knowledge or regulatory compliance.

Our prior work [13] showed that GE can rediscover symbolic relationships (e.g. the quadratic discriminant) and generate compact and transparent classifiers. For example, using the hand-made variable $PLmPW = \text{petal.length} \cdot \text{petal.width}$, GE accurately classified the Iris species with high accuracy and full transparency.

Recent research explores hybrid methods that combine GE with neural networks or extract symbolic rules from black-box models [14].

C. Comparison Criteria

In this work, we evaluate post-hoc and symbolic explanation methods based on the following key dimensions:

- **Fidelity:** Accuracy of the explanation relative to the model or the true function.
- **Simplicity:** Complexity of the resulting explanation (e.g. expression length, tree depth).
- **Semantic Alignment:** Degree to which the explanation uses meaningful or domain-relevant variables.
- **Stability:** Consistency of the explanation across multiple runs or perturbed inputs.

D. Incorporating GE into XAI Pipelines

To formalize the use of GE as an alternative to post-hoc explainers, we propose a transparent pipeline for integrating it into explainability workflows. This replaces black-box interpretation with directly evolved, interpretable models:

1) **Feature Preprocessing:** Prepare input data; optionally derive MIVs.
2) **Grammar Specification:** Define a context-free grammar with domain semantics.
3) **GE Model Training:** Apply GE to evolve models mapping features/MIVs to outputs.
4) **Symbolic Rule Extraction:** Select models according to fitness, simplicity, and semantic alignment.
5) **Validation and Interpretation:** Assess accuracy and interpret model structure.

Figure 1 illustrates the structure of this methodology.

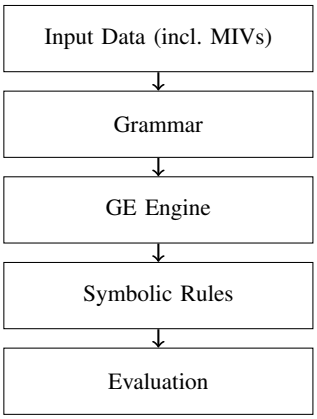

Fig. 1. Pipeline for symbolic model construction using GE

III. Experimental Comparison

To compare black-box explanation methods with symbolic GE models, we performed experiments on two tasks: a synthetic quadratic discriminant classification and the Iris dataset, a standard benchmark for interpretability studies.

A. Quadratic Classification Task

This task involves a simple symbolic function: the discriminant of a quadratic equation $d = b^2 - 4ac$ to classify input into two ($d > 0$), one ($d = 0$) or no real solutions ($d < 0$).

The dataset contains synthetic (a, b, c) tuples labeled with the corresponding value of d.

In the initial phase of the experiment, we trained models to classify the number of real roots using only (a, b, c), without providing $d = b^2 - 4ac$. This tested whether standard methods, linear regression, decision trees, and random forest (RF), could learn the decision boundary from data alone. The performance of the model was assessed using two metrics: RMSE (Root Mean Squared Error), defined as $\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, and its normalized form, NRMSE, $NRMSE = \frac{RMSE}{\bar{y}}$.

The prediction accuracy is presented in Table I. Although random forest had the lowest test error, none of the models produced interpretable rules or captured the underlying structure. Their generalization, especially near $d = 0$, was unstable.

TABLE I
RMSE and NRMSE of selected machine learning models trained on (a, b, c) without access to the discriminant

Model	RMSE (train)	NRMSE (train)	RMSE (test)	NRMSE (test)
Linear Regression	0.8267	0.8026	0.8239	0.8368
Decision Tree	0.5431	0.5273	0.8142	0.8224
Random Forest	0.2284	0.2218	0.5883	0.5942

This motivated the introduction of a domain-informed variable representing the discriminant to evaluate the benefits of symbolic modeling and explainability.

a) LIME and SHAP Results: We applied both explanation tools to the RF classifier and observed notable inconsistencies in predictions:

- LIME frequently attributed great importance to the variable b, with fewer or inconsistent roles for a and c.
- SHAP produced attributions in which c often appeared dominant, sometimes contradicting LIME.

Neither method revealed the discriminant formula or suggested that a quadratic combination of a, b, and c was relevant.

As shown in Fig. 2, LIME and SHAP sometimes disagreed on both the importance and direction of the feature. Similar instability was observed in [13], where shallow explanations lacked semantic consistency.

b) GE Results: In contrast, the GE model directly rediscovered the symbolic form of the discriminant: $d = b^2 - 4ac$ GE then generated an interpretable rule-based classifier:

$$\texttt{ifelse}(d > 0, \text{"two"}, \texttt{ifelse}(d == 0, \text{"one"}, \text{"zero"}))$$

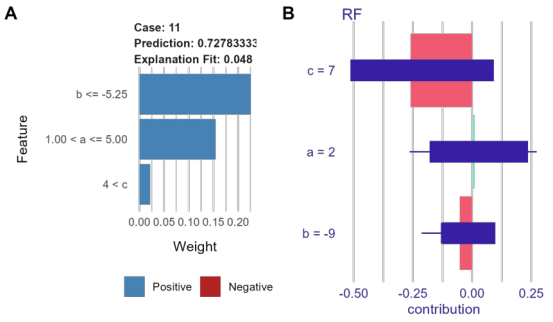

Fig. 2. (**A**) LIME and (**B**) SHAP explanations for a given prediction

This compact and interpretable structure is exactly consistent with the known semantics of the problem. Figure 3 illustrates the classification process: calculate the discriminant d, then assign the class based on its value.

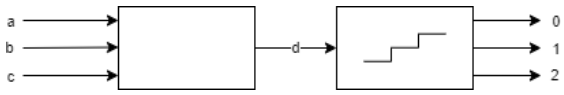

Fig. 3. GE-evolved symbolic rule with intermediate variable d representing the discriminant. The rule exactly matches the theoretical solution structure.

The symbolic rule for d was discovered by evolutionary search on a grammar encoding arithmetic operations and constants relevant to the discriminant:

```
<expr> ::= <var> | <op>(<n>, <var>)
    | <op>(<var>, <n>) | <op>(<expr>, <expr>)
<op>   ::= "-" | "*" | "^"
<var>  ::= a | b | c
<n>    ::= 1 | 2 | 3 | 4
```

The grammar for the final classification, which assigns the number of solutions of a quadratic equation, is listed below:

```
<result>     ::= ifelse(<expr>, "one", "other")
<expr>       ::= (<expr> & <sub_expr>)
        | (<expr> | <sub_expr>) | <sub_expr>
<sub_expr>   ::= <comparison>(<var>, <func_var>)
<comparison> ::= > | < | ==
<func_var>   ::= <num> | <var> | <func>(<var>)
<func>       ::= mean | max | min | sd
<var>        ::= a | b | c | d
<num>        ::= 0 | 1
```

The GE model achieved 100% accuracy on the training and test sets due to perfect alignment with the true decision boundary. In contrast, the random forest misclassified 4–5% near $d = 0$ and failed to generalize. This case demonstrates that GE provides strong performance and formal transparency, providing stable, reproducible, and verifiable expressions, unlike post-hoc black-box explainers. This experiment highlights the value of MIVs, like d, which capture complex feature relationships. MIVs reflect domain knowledge and decompose learning into interpretable parts. As shown in [13], they help GE efficiently discover accurate and human-readable rules, which post-hoc tools such as LIME or SHAP cannot achieve due to their lack of structural abstraction.

B. Iris Dataset

The Iris dataset includes 150 samples labeled: Setosa, Versicolor, or Virginica, each with four features: sepal length/width and petal length/width. It is commonly used to assess the interpretability of the model.

a) LIME and SHAP Results: RF trained on the Iris dataset was explained with LIME and SHAP. Compared to the quadratic task, the results were more consistent. Both tools often highlighted *Petal.Length* and *Petal.Width*, though discrepancies remained. As shown in Fig. 4, LIME negatively emphasized all features, while SHAP prioritized petal features over sepal ones.

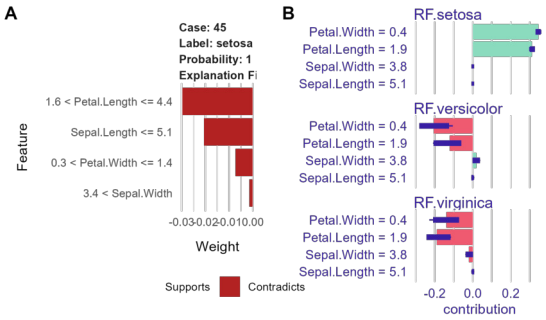

Fig. 4. (**A**) LIME and (**B**) SHAP explanations for the same Iris sample

Despite improved consistency compared to the delta experiment, neither method offered a compact or semantically meaningful rule that could generalize throughout the dataset.

b) GE Results: GE was applied to the full feature set to minimize misclassifications, running for 1000 iterations with default settings and the grammar shown below:

```
<expr>       ::= (<expr> & <sub_expr>)
       | <expr> | <sub_expr>) | <sub_expr>
<sub_expr>  ::= <comparison>(<var>, <func_var>)
<comparison> ::= > | < | == | >= | <=
<func_var>   ::= <num> | <var> | <func>(<var>)
<func>       ::= mean | max | min | sd
<var>        ::= Sepal.Length | Sepal.Width
               | Petal.Length | Petal.Width
<num>        ::= 1 | 1.5 | 2 | 2.5 | 3 | 4 | 5
```

Representative symbolic models and their corresponding plots are shown below, highlighting the effective separation of the *setosa* and *versicolor* classes.

```
ifelse(Petal.Length <= Sepal.Width, "setosa",
"other")

ifelse((Petal.Width <= 1.5)&(Petal.Length >= 2.5),
"versicolor", "other")
```

Although the model identified all *setosa* samples, its classification of *versicolor* at 94.66% accuracy indicates potential for further improvement.

To improve the classification, we extended the feature set by introducing a Meaningful Intermediate Variable that approximates the petal area:

$$PLmPW = petal.length * petal.width \quad (1)$$

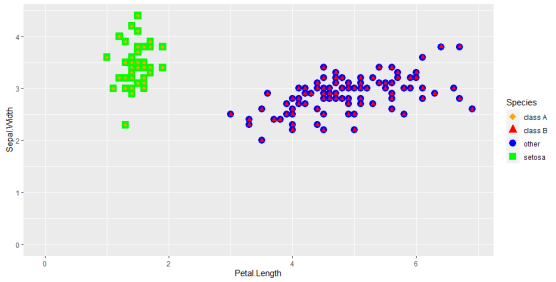

Fig. 5. *Setosa* classification with sepal width and petal length

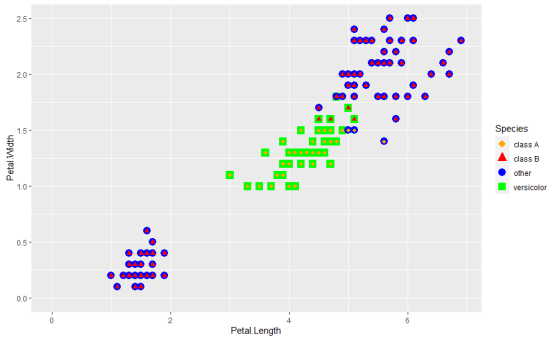

Fig. 6. *Versicolor* classification with petal width and petal length

Using this MIV, GE developed a simple and accurate rule-based model:

- If $PLmPW \leq 2$: Setosa.
- If $2 < PLmPW \leq 8$: Versicolor.
- If $PLmPW > 8$: Virginica.

This rule achieved approximately 95% precision, with a decision structure that is semantically meaningful and easy to understand.

Figure 7 presents the functional model obtained by combining the individual rules.

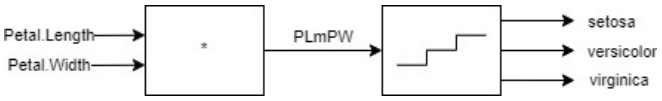

Fig. 7. Functional model for Iris species classification with MIV

The results of the classification process performed with the functional model provided are depicted in Figure 8.

We also introduced additional MIVs including the sepal area:

$$SLmSW = sepal.length * sepal.width, \quad (2)$$

petal and sepal perimeters, and various length-to-width ratios to improve model accuracy.

MIVs used in the model may be derived not only from the original input features but also from other previously defined MIVs. One such example is the petal-to-sepal area ratio, proposed using the MIVs defined in Equations 1 and 2:

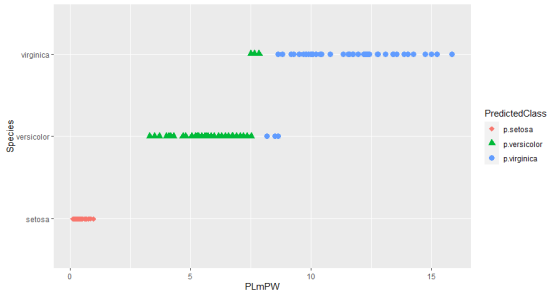

Fig. 8. Iris species classification with $MIV = petal.length * petal.width$

$$PdS = \frac{PLmPW}{SLmSW} \tag{3}$$

Using the GE approach, the following classification rules were generated on the basis of the derived PdS value:

```
ifelse(PdS <= 0.2, "setosa", "other")
ifelse((PdS <= 0.43) & (PdS >= 0.22),
"versicolor", "other")
ifelse(PdS >= 0.43, "virginica", "other")
```

These rules define a functional model, presented in Figure 9.

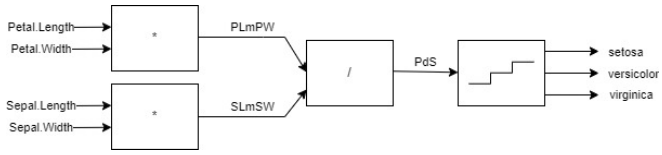

Fig. 9. Functional model for Iris classification with multiple MIVs

The classification results obtained using the functional model based on multiple MIVs are presented in Figure 10.

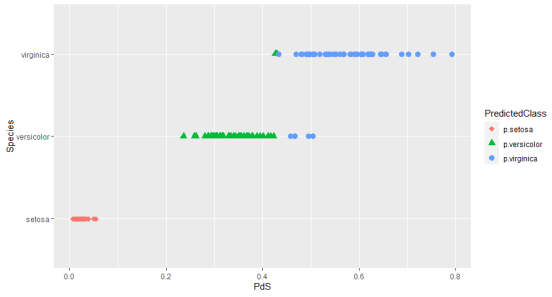

Fig. 10. Iris species classification with multiple MIVs

The model demonstrates an improvement over the previous one, achieving an accuracy of 96%.

IV. Analysis and Discussion

The experimental results highlight differences between post-hoc explainability techniques and model-driven symbolic methods. Our comparative analysis focuses on four critical dimensions: stability, interpretability, semantic alignment, and predictive performance.

A. Stability and Reproducibility

An intrinsic limitation of post-hoc methods is the *instability*. Attribution values vary across runs or small input changes, especially in the quadratic task. For example, LIME overemphasized b, while SHAP inconsistently prioritized c, reducing reliability in high-stake settings.

In contrast, GE consistently produced models with stable structure and clear semantics across runs, offering strong support for reproducibility, system auditing, and model understanding.

B. Interpretability and Semantic Coherence

GE produces symbolic models based on mathematical or logical rules that are directly interpretable. In the quadratic task, it rediscovered $d = b^2 - 4ac$. In the Iris dataset, MIV-based rules (e.g. petal area, petal-sepal ratios) reflect domain structure and improved transparency.

LIME and SHAP offer intuitive local insights but lack structural integration. Their explanations are not reusable and generalizable, limiting their value for model critique and design.

C. Role and Value of Intermediate Variables

MIVs significantly enhanced interpretability and performance. In the Iris task, handcrafted variables like `PLmPW` and `PdS` helped GE derive simple domain-aligned rules by embedding biological knowledge into the search process. We also showed that MIVs can be composed hierarchically—for example, `PdS` combines two others, supporting scalable symbolic modeling in complex domains with important but implicit constructs.

D. Prediction Accuracy

Although symbolic modeling prioritizes interpretability, GE models also achieved high precision. In the quadratic task, GE accurately captured the decision boundary. For the Iris dataset, MIV-based models reached 96% accuracy—matching black-box classifiers, but with much greater transparency. In particular, some gains came from structured variables, emphasizing the value of hybrid approaches that combine domain knowledge with symbolic learning.

E. Limitations and Practical Considerations

Despite its advantages, GE has practical constraints. Grammar design requires domain expertise, and poor design either limits the search space excessively or fails to guide the model toward meaningful solutions. As a stochastic method, GE may require multiple runs or tuning. However, in domains with expert input or bounded problems, GE offers a controllable and verifiable alternative to opaque black-box explanations.

The comparative evaluation described in Section II-C established four key criteria: reproducibility, interpretability, semantic consistency, and structural clarity. These dimensions provided a principled framework for analyzing the strengths and limitations of each method. The experimental findings validate the relevance of this framework: post-hoc methods

like LIME and SHAP showed weaknesses in reproducibility and semantic alignment, often generating inconsistent or context-insensitive attributions. In contrast, the symbolic models developed through GE satisfied all four criteria. They were reproducible across runs, structurally coherent, interpretable by domain experts, and semantically aligned with known functional relationships. The use of this criteria-driven perspective was essential in highlighting that precision alone is insufficient when evaluating explainable models, and structural and semantic properties must also be considered.

To consolidate the results and observations discussed in the experimental sections, Table II provides a qualitative comparison of the three approaches evaluated. LIME, SHAP, and Grammatical Evolution. The comparison is organized around the four criteria defined in Section II-C: fidelity, simplicity, semantic alignment, and stability.

TABLE II
QUALITATIVE COMPARISON OF EXPLANATION METHODS

Criterion	LIME	SHAP	GE
Fidelity	Low–Medium	Low-Medium	High (in known domains)
Simplicity	Medium	Medium	High (compact rules)
Semantic Alignment	Low	Low	High (uses MIVs)
Stability	Medium	Medium	High

As shown in the table, LIME and SHAP provide partial fidelity and limited semantic coherence, particularly in tasks where the decision boundary depends on specific functional structures. Their attributions are also sensitive to sampling variability and lack cross-instance consistency. GE, by contrast, produces symbolic models that offer high fidelity when the target rule is recoverable, and do so using semantically meaningful constructs such as intermediate variables. These symbolic rules are compact and structurally transparent, providing an interpretable basis for classification decisions. This distinction is particularly evident in the quadratic task, where GE recovered the exact mathematical discriminant, while LIME and SHAP failed to detect the underlying structure.

V. CONCLUSION

This study presents a comparative evaluation of black-box post-hoc explanation methods (LIME and SHAP) and symbolic, model-driven approaches using Grammatical Evolution. In two classification tasks, GE consistently demonstrated superior semantic clarity, interpretability, and reproducibility.

The core findings are as follows:

- **Interpretability and Semantic Alignment:** GE-generated symbolic models are inherently interpretable, expressed as explicit formulas aligned with domain semantics (e.g., discriminants or feature ratios). In contrast, LIME and SHAP offer fragmented, instance-specific attributions that are often semantically opaque and inconsistent across similar inputs.
- **Structured Abstraction through MIVs:** GE incorporates structured domain knowledge via Meaningful Intermediate Variables, such as petal area or petal-to-sepal ratios. These enhance model clarity, simplify decision boundaries, and support composability by enabling hierarchical MIVs built from previously defined variables.
- **Accuracy and Trustworthiness:** While GE emphasizes interpretability, it also achieves competitive accuracy—up to 96% on the Iris dataset, generating compact and transparent rules. This combination improves trust and auditability in high-stakes or regulated applications.

These results suggest that symbolic regression techniques, and particularly GE, provide a viable alternative to post-hoc explainers in contexts where explanation fidelity and transparency are critical.

Future work will focus on intelligent and automated grammar generation, discovering intermediate variables, and improving convergence strategies for grammar-guided evolution.

REFERENCES

[1] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, Elsevier (2019), https://doi.org/10.1016/j.inffus.2019.12.012.

[2] Guidotti, R., Monreale, A., Ruggieri, S., et al.: A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* **51**(5), 1–42 (2019), https://doi.org/10.1145/3236009.

[3] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**(1), 18 (2021), https://doi.org/10.3390/e23010018.

[4] Ribeiro, M. T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *KDD 2016*, pp. 1135–1144. ACM (2016), https://doi.org/10.1145/2939672.2939778.

[5] Lundberg, S. M., Lee, S. I.: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* **30**, pp. 4765–4774 (2017), https://dl.acm.org/doi/10.5555/3295222.3295230.

[6] Sepioło, D., Ligęza, A.: Towards Explainability of Tree-Based Ensemble Models: A Critical Overview. In: *New Advances in Dependability of Networks and Systems*, pp. 287–296. Springer (2022), https://doi.org/10.1007/978-3-031-06746-4_28.

[7] Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019), https://doi.org/10.1038/s42256-019-0048-x.

[8] Ligęza, A., et al.: Explainable Artificial Intelligence. Model Discovery with Constraint Programming. In: *ISMIS 2020*, Springer, pp. 171–191 (2020), https://doi.org/10.1007/978-3-030-67148-8_13.

[9] Hu, T.: Can Genetic Programming Perform Explainable Machine Learning for Bioinformatics? In: *Genetic and Evolutionary Computation*, Springer, pp. 63–77 (2020), https://doi.org/10.1007/978-3-030-39958-0_4.

[10] Ryan, C., O'Neill, M., Collins, J. J. (eds.): *Handbook of Grammatical Evolution*. Springer (2018), https://doi.org/10.1007/978-3-319-78717-6.

[11] Sepioło, D., Ligęza, A.: Towards Model-Driven Explainable Artificial Intelligence. An Experiment with Shallow Methods Versus Grammatical Evolution. In: *ECAI 2023 Workshops*, Springer, pp. 360–365 (2024), https://doi.org/10.1007/978-3-031-50485-3_36.

[12] Orzechowski, P., La Cava, W., Moore, J.H.: Where Are We Now? A Large Benchmark Study of Recent Symbolic Regression Methods. IN: *GECCO '18: Proceedings of the Genetic and Evolutionary Computation Conference*, Association for Computing Machinery, pp. 1183–1190 (2018), https://doi.org/10.1145/3205455.3205539.

[13] Sepioło, D., Ligęza, A.: Towards Model-Driven Explainable Artificial Intelligence: Function Identification with Grammatical Evolution. IN: *Applied Sciences* 14, 5950, (2024), https://doi.org/10.3390/app14135950.

[14] Tsoulos, I. G., Tzallas, A., Karvounis, E.: Using Optimization Techniques in Grammatical Evolution. In: *Future Internet*, 16, 172, (2024), https://doi.org/10.3390/fi16050172.

[15] Ligęza, A.: An experiment in causal structure discovery. A constraint programming approach. In: In: *ISMIS 2017*, Springer, pp. 261-268 (2017), https://doi.org/10.1007/978-3-319-60438-1_26.