Symbolic Regression for Transparent Clinical Decision Support: A Data-Centric Framework for Scoring System Development

Veronica Guidetti[1,*], Federica Mandreoli[1]

[1]*Department of Physics, Informatics and Mathematics, University of Modena e Reggio Emilia, Modena, Italy*

Abstract

Machine learning (ML) has transformed healthcare, improving diagnostics, treatment, research, and patient care. However, clinical decision support (CDS) still relies heavily on classical statistical models and manual rules, lacking transparency and accuracy. Starting from the mid-20th century, scoring systems offer a transparent approach to CDS development. Nevertheless, classical methods for scoring systems like logistic regression may lack predictive accuracy and suffer in handling complex high-dimensional electronic health record data, while black-box ML models pose risks due to their lack of interpretability. To address these challenges, our group focuses on developing interpretable symbolic ML approaches, leveraging multi-objective symbolic regression (MOSR) to accelerate index development, mitigate human bias, and allow for the exploration of new aggregation functions and weighting systems. MOSR optimizes multiple objectives simultaneously, distilling complex phenomena into non-linear yet understandable constructs, a crucial aspect for gaining trust from healthcare professionals. Moreover, MOSR is highly flexible and extendable to classical statistical models. This paper presents our experience in developing data-driven scoring systems, building on real-world applications such as COVID-19 mortality prediction and risk estimation post-liver transplantation. Our methodology involves designing the entire data pipeline, from feature selection to scoring formula generation, highlighting the importance of developing data-centric and interpretable ML techniques for high-risk domains.

Keywords

Scoring systems, Symbolic Machine Learning, Data-Centric AI, Multi-Objective Optimization, Clinical Decision Making

1. Introduction

Artificial Intelligence (AI) systems have revolutionized healthcare by offering opportunities to enhance diagnostics, treatment, research, and patient care. Among these tasks, we want to focus on the development of data-driven solutions to clinical decision support (CDS). Currently, CDS predominantly relies on classical statistical models and manually curated rules or heuristics that are designed to identify patient cohorts with specific characteristics of interest.

Medical scoring systems are considered as the most common CDS tool because they provide healthcare professionals with objective, intelligible, and quantifiable measures to aid in clinical decision-making. Indeed, by interpreting scores in conjunction with clinical expertise and patient-specific factors, healthcare professionals can make more informed decisions regarding diagnosis, treatment planning, and patient management. The history of scoring systems in medicine dates back to the mid-20th century. One of the earliest and most well-known scoring

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

veronica.guidetti@unimore.it (V. Guidetti); federica.mandreoli@unimore.it (F. Mandreoli)

systems is the Apgar score, developed by Virginia Apgar in 1952, which assesses the health of newborns and is still widely used today for quick assessment. Creating and validating clinical scoring systems involves several key steps. Clinicians begin by identifying a patient population with a specific disease or condition and gathering relevant data, including demographics, medical history, and laboratory results. Statistical methods are then employed to identify predictive factors for the outcome of interest, such as mortality or disease progression. Using these factors, a model is developed to generate a score predicting the likelihood of the outcome occurring. Finally, the scoring system is validated by testing its performance on a separate group of patients with the same condition to ensure its accuracy and reliability.

Over the years, scoring systems continued to be developed, incorporating data from new technologies, such as imaging and laboratory tests, to improve their accuracy and reliability. However, traditional index creation methods, successful in the past, faced challenges in the context of complex clinical phenotypes and the vast landscape of Electronic Health Records (EHRs). In fact, without automated pipelines to reduce variables or identify their importance, EHR-derived datasets are often high-dimensional, scarce, sparse, and unbalanced [1]. For these reasons, the emergence of AI and sophisticated ML models able to analyze extensive datasets encompassing genetics, lifestyle factors, Patient Reported Outcomes, and EHRs, has sparked a renewed emphasis on developing medical scoring systems.

This work aims to describe the state of the art in the development of data-driven scoring systems and illustrate the progress and applications studied by our research group in recent years for the development of interpretable CDS tools. Specifically, in Section 2, we introduce the concept of scoring systems and their traditional development methods. We then delve into the challenges posed by the rise of AI in automating index creation and explain why interpretable ML techniques are the primary tools for this task. Moving forward, we present state-of-the-art methods for automatically generating scoring systems, with a focus on symbolic regression. This technique serves as the foundation for a new approach developed by our team for constructing flexible and parsimonious indices in real-world scenarios. Section 3 outlines our comprehensive data pipeline for scoring system development, highlighting key aspects and featuring real case studies from our recent research endeavors. In Section 4, we draw some concluding remarks and summarize the outstanding challenges in this field that we aim to address in the near future.

2. Preliminaries: Scoring Systems Development and Symbolic Regression

A score is a mathematical combination of a set of elementary indicators (EIs) representing the different components of a multidimensional concept to be measured (e.g. development, quality of life, wealth, risk, etc.). Hence, synthetic scores are used to measure concepts that cannot be captured by a single indicator. In general, a composite index should be based on a theoretical framework, allowing the selection, combination, and weight of the EIs to reflect the size or structure of the phenomenon being measured. Building scores often requires a series of decisions/choices to be made. The process involves multiple steps: i) *theoretical framework definition* where the concept to be measured and the EIs to be considered are identified; ii) *EIs selection* based on their relevance, validity, availability, cost, etc., and *standardization* so as to work with adimensional quantities; iv) *aggregation function definition* where the final shape of the index combining the selected EI is determined; v) *index validation* in terms of robustness, generalization and discriminating ability.

There is no general method for the construction of synthetic indices so each score is linked to the particular application. Classically, finding the right aggregation function and the weighting of EIs is a highly non-trivial task [2] that involves human decision by default and in which data sets are mostly used to validate the score after it is built or, when using parametric statistical models, to learn its numerical coefficients (see, e.g., [3]). Despite its widespread use in high-stake domains, such an approach may lead to scores that suffer from the lack of formal guarantees in terms of performance and constraint compliance [4].

2.1. Clinical Score Generation in the Era of Big Data and AI

Despite the definition of the theoretical context and the identification of the relevant EIs will always require some human knowledge to be reliable, the other stages of index creation may be automated. Indeed, the promise of systematic, targeted, and data-driven scores for CDS becomes apparent with recent advances in ML and the ever-increasing availability of EHR data. This would not only speed up index creation but also eliminate any human bias in its construction, as well as explore further aggregation functions and weighting systems than what was done in the past. However, the potential benefits are met with equally substantial challenges. Achieving systematic index creation necessitates the development and deployment of highly accurate clinical prediction models for a wide range of clinical problems. Moreover, the CDS framework should be robust enough to cope with and adapt to significant variations in clinical practices and documentation standards across different healthcare providers and systems. For these reasons, comprehensive assessments should be performed to weigh the potential advantages and risks associated with each automated CDS solution [5].

Meeting the previous requirements is made easier by interpretable ML techniques that can be inspected and questioned by domain experts, ultimately leading to a higher level of trust and, consequently, facilitating their integration into existing decision-making frameworks. In fact, it is a view increasingly shared by the scientific community that black box models cannot be the final answer to increase accuracy or fix cases where the assumptions of classical models are not met [6, 7]. Opaque models cannot be fully understood, were shown to be dangerous in many real-life situations, and so should not be used in high-stakes domains [8]. Moreover, it was recently shown that simpler intelligible models achieve comparable, if not better, results in several real-world cases [9], especially when dealing with tabular data such as EHRs.

2.2. State-of-the-art Approaches

State-of-the-art approaches for interpretable data-driven scoring systems formalize finding the index aggregation formula as an optimization problem. Most of them translate index creation into a classification problem, primarily focusing on binary outcomes [10, 4, 11, 12]. Most works use integer programming to build scoring systems with fixed aggregation function shapes and use training data to learn numerical coefficients. While these works have the merit of supporting high-stake decisions through interpretable data-driven models, they marginally explore the space of scoring functions. In fact, most methods rely on dummy indicators, based on thresholds, and use decision tree models to construct the aggregation functions [10, 11, 12, 13]. Constraining the space of aggregation functions to be searched to some linear structure may prevent the model from finding the right solution because of structural constraints introduced for the sake of simplifying the phenomenon. As a matter of fact, non-linear scoring systems are already widely used (see [2] for a technical introduction). For instance, the Body Mass Index (BMI) [14] represents a simple and widely known example in clinical practice.

The problems previously listed can be solved by using symbolic regression (SR), the data-driven ML method for regression analysis. In fact, as we will elaborate in the next section, SR is able to search a much wider space of mathematical formulas to find the model that best fits a given dataset. Moreover, the use of SR is not limited to non-linear model fitting but can be adapted to any parametric statistical model. For the above reasons, the use of SR in scoring system development for high-stakes domains may lead to extremely important results in the coming years [5, 15].

2.3. Symbolic Regression

SR refers to data-driven ML methods for regression analysis that search the space of mathematical formulas to find the model that best fits a given dataset. The most classical approach to SR relies on genetic programming (GP) algorithms [16, 17, 15]. Standard GP approaches to SR require the selection of a predefined set of mathematical operators and variables that are used to construct candidate models (syntax trees in SR, see Figure 1 for an example). The algorithms need to perform optimization starting from a large set of randomly initialized models (population) that is evolved through several generations where individuals undergo transformations inspired by genetic mutations. Our interest in SR comes from its remarkable results on non-linear data-driven modeling, even on small training data sets [18]. Indeed, SR was shown to generalize better than standard ML methods on hundreds of tasks, providing interpretable results in most cases.

When the optimization process relies on meta-heuristic processes such as GP, SR can be easily extended to multi-objective tasks. Multi-objective SR (MOSR) has been recently developed to address constrained minimization problems and prevent overfitting the training data. A classical approach to solve the overfitting problem is to consider a bi-objective setup optimizing both complexity and accuracy [19, 20]. In fact, without a goal associated with complexity, models tend to become excessively verbose, losing generalization performance and transparency. Only recently, some work started combining accuracy and parsimony into a single objective through model selection principles such as the Akaike or Bayes Information Criterion (AIC and BIC), or the minimum description length [21]. Other applications of MOSR aim to incorporate knowledge-driven (in)equalities to be satisfied [22, 23, 24]. While these types of constraints can be easily incorporated in scoring system development, score construction itself can be a natively multi-dimensional problem that must satisfy or track more than one property.

3. Symbolic Data-centric AI for Scoring Systems

Our interest in developing data-driven scoring systems has stemmed from our close collaboration with the Department of Infectious Diseases of the Modena University Hospital, particularly with members of the HIV Metabolic Clinic (MHMC), a tertiary-level referral center for the

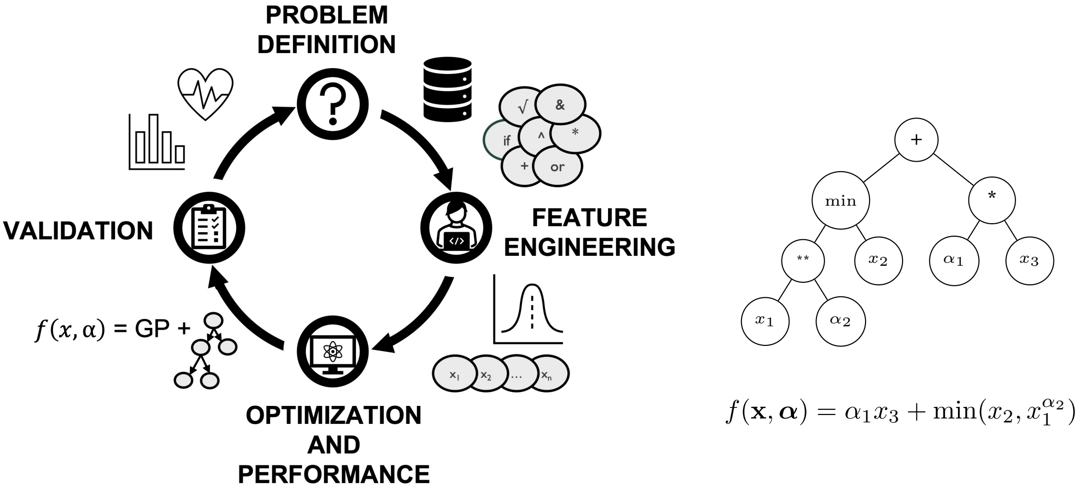

Figure 1: Left: Pictorial view of the data pipeline construction. Right: Example of model syntax tree.

diagnosis and treatment of non-infectious co-morbidities in people living with HIV (PWH). In 2022 [25], our research group pioneered the introduction of the first MOSR approach for managing data-driven, continuous, and non-linear scoring systems. The necessity for a multi-objective approach in continuous medical index creation arises from the intricate nature of medical outcomes, often too complex to predict precisely. This challenge is compounded by the characteristics of EHRs, which are typically not only small but also unbalanced. To address these challenges, unlike standard strategies that validate results post-hoc using metrics like stratification power and index balancing, our approach incorporates these desirable properties directly into the optimization phase. This integration provides greater control over the behavior and complexity of the models.

Moreover, despite the scientific community's main focus being on identifying index aggregation functions, our fruitful collaboration with the MHMC highlighted that the *automation of data-driven scoring system development should rely on a complete data pipeline that encompasses the entire data lifecycle*, from data preprocessing and feature engineering, to model creation, optimization, and validation. A pictorial view of the data pipeline needed to create scoring systems with MOSR is depicted in Figure 1. While certain phases, such as the problem definition step, necessitate a synergistic approach between data scientists and clinicians, other steps can be partially optimized. In our experience, these are:

- **EI engineering and selection.** The construction of normalized EI can be automatized once the data type, its meaning, and the kind of non-linearities are specified in index formulation. A major challenge in SR is the vast function space to be explored, which grows exponentially with the number of mathematical operators and variables [26]. To improve convergence rate we usually rely on minimum-redundancy-maximum-relevance non-linear feature selections [27, 28, 29] whenever needed.
- **Optimization objective.** The ability to abstract the problem enables the identification of the appropriate class of models wherein risk assessment should be integrated, whether

it's classification, regression, survival analysis, etc. By comprehending the properties essential for the index's effectiveness, one can discern additional constraints or desiderata, such as constraints/objectives on correlation, calibration, or distribution. Lastly, identifying the final application domain aids in understanding the tolerable nonlinearities and the maximum acceptable complexity level of the final formulas. These considerations collectively facilitate the identification of optimal optimization goals.

- **Performance metrics and comparison.** Performance metrics beyond the optimization criteria should be identified to check the predictive power of the generated scores. These metrics need to be tailored to the specific problem under study. For instance, in the context of a risk score based on binary classification, it is essential to calculate metrics such as sensitivity, specificity, and the area under the ROC curve. Furthermore, it is paramount to compare the results with those obtained from classical statistical methods and other benchmark ML approaches.
- **Model selection and validation.** Model selection should consider performance metrics, safety, and domain expertise. In addition to standard performance evaluation, validating the score typically entails analyzing its distribution within the target population and its correlation with other clinically relevant measures. This ensures the score's consistency and compatibility with existing medical knowledge.

Below, we will demonstrate that since the publication of the methodological work, our group has persistently advanced in developing, systematizing, and automating various aspects necessary for score development automation, also expanding the methodology to encompass different case studies. We start by briefly summarizing two situations where scores are modeled as regression and binary classification problems, afterward, we show in detail how we extended the method to survival analysis problems.

Improving Frailty index assessment for PWH [30]. This study developed a data-driven tool for geriatric assessment of PWH, focusing on enhancing the Frailty Index (FI), a clinical score based on 35 variables. Leveraging data from the MHMC cohort, we designed a pipeline for constructing reduced FIs. Starting from a set of knowledge driven 54 binary EIs, we employed a non-linear feature selection method and identified the 27 most relevant EIs in predicting the FI. We modeled index simplification as a regression problem and used MOSR to replicate the values, distribution, and risk stratification ability of the original FI, minimizing weighted mean square error, Wasserstein distance, and maximizing pairwise Kendall correlation. We evaluated optimal model predictiveness through calibration, correlation with the original index, and associations with established geriatric outcomes such as age, the EQ-5D-5L score, and the SPPB index. The simplest optimal model used only 16 readily available variables, meeting all requirements and was incorporated into the MHMC's automatic data collection pipeline.

Short-term mortality prediction in patients with Covid-19 [31]. This study aimed to predict short-term in-hospital mortality risk using data collected upon hospital admission. EHRs from 2400 patients with Covid-19 diagnoses were gathered. A non-linear feature selection method reduced covariates from 25 to 10, validated by the medical team and standardized. Index creation involved a tailored non-linear logistic regression using formulas generated by MOSR. Experiments minimized weighted Binary Cross Entropy and maximized the F1 score to optimize the sensitivity-specificity tradeoff. MOSR outperformed popular machine-learning

Table 1
Comparison between classical Cox model and a MOSR solution showing examples of risk factors mined by MOSR. Only statistically significant variables were reported for the Cox model.

	Model	**Cov.**	α_j	$[5\%, 95\%]_{\alpha_j}$	$-\log_2(\mathrm{p})$	**PHA**
Cox model	$CLH = \sum_i \alpha_i x_i$	ICU	0.68	[0.29,1.08]	10.62	Failed
		cMELD	0.67	[0.26,1.08]	9.54	Satisfied
		PTI	1.50	[0.74,2.25]	13.35	Satisfied
MOSR	$CLH = \alpha_{\mathsf{F}_1}\, \mathsf{F}_1 + \alpha_{\mathsf{F}_2} \mathsf{F}_2$	F1	1.82	[1.07,2.57]	18.89	Satisfied
	$F_1 = \mathsf{PTI}^{\mathsf{HD}}$	F2	1.15	[0.78,1.52]	30.11	Satisfied
	$F_2 = \max(\frac{2}{3}HD, \min(\mathsf{ICU}, \max(\mathsf{MELD}, \mathsf{PTI}^{\mathsf{HD}}\ \mathsf{Tx2015}^{\mathsf{PTI}}))))$					

algorithms and classical human-generated indices, prioritizing false negative reduction for timely interventions.

3.1. MOSR for survival: Mining post Liver Transplantation (LT) risk factors

The study [32] aimed to quantify the interaction between risk factors in predicting death within the first four months post-LT. The data consisted of 485 EHRs of patients who underwent LT at the University Hospital of Modena between 2010 and 2020. Available exposure variables included patient admission details, preoperative conditions such as colonization by multidrug-resistant bacteria, and postoperative risk factors like bloodstream infections. Due to data scarcity and imbalance, EI selection and construction required collaboration between data scientists and clinicians. The chosen set of covariates was: Hospitalization days (**HD**); Intensive Care Unit days (**ICU**); Model for End-Stage Liver Disease (**MELD**); Duration of surgery (**DS**); LT year (**LTY**); Post-LT infection (**PTI**); MDR-Gram negative pre-operative colonization (**GnC**); on top of death observation and censoring times. Variable selection was followed by feature engineering to create comparable and informative EIs. Specifically, we discretized continuous variables based on clinically meaningful intervals.

Cox's regression, a classic semi-parametric survival model, is the most classic and easy-to-interpret model that can estimate the effects of exposure variables and adjust for confounding effects. Classically, Cox's hazard model is written as:

$$H(\mathbf{x}, t) = H_0(t) \times \exp\left\{\boldsymbol{\alpha}^T \mathbf{x}\right\} \tag{1}$$

where $\mathbf{x} = \{x_i\}, i = 1 \ldots k$, are the covariates and $\boldsymbol{\alpha} = \{\alpha_i\}$ are the parameters to be estimated by the model. Therefore, the hazard function H is given by the product of the baseline hazard $H_0(t)$ and the covariate-dependent relative risk. As the only time dependence lies in $H_0(t)$, a fundamental assumption underlying the application of the Cox model is the Proportional Hazard Assumption (PHA). The application of the classical Cox model turned out not to be suitable for the indicators created, since, as shown in Table 1, some covariates did not meet PHA. Classical methods for extending the Cox model in this situation are difficult to apply to sparse and unbalanced data, as well as being less interpretable. Therefore, we embedded SR into Cox's regression by making the covariate-dependent log relative hazard function (CLH) trainable and potentially non-linear: $\exp\left\{\boldsymbol{\alpha}^T\mathbf{x}\right\} \to \exp\left\{f(\boldsymbol{\alpha},\mathbf{x})\right\}$. By optimizing partial AIC ($\mathcal{P}$AIC) and model complexity, parametrized as the number of nodes in the syntax tree, we maximized model likelihood while minimizing the number of numerical parameters and the degree of non-linearity for enhanced interpretability. MOSR survival models not only outperformed classical Cox regression in predictive performance but also successfully mined composite data-driven risk factors, overcoming classical model limitations such as the PHA. See the bottom panel in Table 1 for a representative optimal solution. Finally, model selection was guided by differences in $\mathcal{P}$AIC values for theoretical support, calibration and predictive performance metrics for efficiency and usefulness, and out-of-distribution prediction for safety assessment. More details about feature selection and construction and model performance and validation can be found in the original text.

4. Concluding Remarks, Open Challenges, and Future Research Opportunities

Scoring system development stands out as a prominent application of data-centric AI, where the fusion of data mining, big data analytics, and ML plays a pivotal role. In this work, we have presented the state of the art in automatic scoring system generation, covering data frameworks, algorithms, and select case studies successfully tackled by our research group. By distilling knowledge from retrospective clinical data, our data processing pipeline and analyses offer seamless adaptability to diverse case studies.

Despite the advancements obtained during the last years in the development of scoring systems, state of the art data pipeline for data-driven scoring system development still lack some crucial milestones to be reached. In addition to being interpretable, CDSs must be able to handle uncertainty. The problem of quantifying and managing epistemic uncertainty becomes paramount when the system under study evolves with time or when dealing with multicenter studies. The ML frameworks devoted to these issues are continual learning (CL) and federated learning (FL) [33].

CL focuses on enabling AI models to learn and adapt continuously over time, incorporating new information while retaining previously acquired knowledge. On the other hand, FL aims to facilitate the development of multicenter studies by overcoming the problems related to centralized learning settings that need to transfer sensitive medical data from multiple centers to a central location. FL and CL approaches to SR have received little attention, with the literature focusing on DL models. Only a few works exist about federated SR algorithms [34, 35], and none of them deal with concept drift.

To foster trust among domain experts, model interpretability may not suffice and their full engagement in model creation and selection should be allowed. In a true human-in-the-loop scenario, physicians should be allowed to contribute their clinical expertise directly to ML model development and selection. This can be facilitated through flexible graphical interfaces enabling iterative feedback processes.

Finally, proposing SR as a method capable of automating the creation of CDS tools will require consolidating and extending it to various parametric models commonly used in clinical score development, such as sub-distribution hazard models and time series analysis.

References

[1] F. Mandreoli, D. Ferrari, V. Guidetti, F. Motta, P. Missier, Real-world data mining meets clinical practice: Research challenges and perspective, Frontiers in big Data 5 (2022) 1021621.

[2] M. Mazziotta, A. Pareto, Methods for constructing composite indices: One for all or all for one, Rivista Italiana di Economia Demografia e Statistica 67 (2013) 67–80.

[3] M. Than, et al., Development and validation of the emergency department assessment of chest pain score and 2h accelerated diagnostic protocol, EMA 26 (2014) 34–44.

[4] B. Ustun, C. Rudin, Learning optimized risk scores, Journal of Machine Learning Research 20 (2019) 1–75.

[5] W. G. La Cava, P. C. Lee, I. Ajmal, X. Ding, P. Solanki, J. B. Cohen, J. H. Moore, D. S. Herman, A flexible symbolic regression method for constructing interpretable clinical prediction models, NPJ Digital Medicine 6 (2023) 107.

[6] G. Kantidakis, H. Putter, C. Lancia, J. d. Boer, A. E. Braat, M. Fiocco, Survival prediction models since liver transplantation-comparisons between cox models and machine learning techniques, BMC medical research methodology 20 (2020) 1–14.

[7] G. Kantidakis, E. Biganzoli, H. Putter, M. Fiocco, et al., A simulation study to compare the predictive performance of survival neural networks with cox models for clinical trial data, Computational and Mathematical Methods in Medicine 2021 (2021).

[8] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[9] L. Semenova, C. Rudin, R. Parr, On the existence of simpler machine learning models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1827–1858.

[10] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Machine Learning 102 (2016) 349–391.

[11] N. Sokolovska, Y. Chevaleyre, J.-D. Zucker, A provable algorithm for learning interpretable scoring systems, in: Proc. of the 21 Int'l Conf. on Artificial Intelligence and Statistics, 2018, pp. 566–574.

[12] C. Q. Zhu, M. Tian, L. Semenova, J. Liu, J. Xu, J. Scarpa, C. Rudin, Fast and interpretable mortality risk scores for critical care patients, 2023. arXiv:2311.13015.

[13] R. Zhang, R. Xin, M. Seltzer, C. Rudin, Optimal sparse survival trees, arXiv preprint arXiv:2401.15330 (2024).

[14] A. Romero-Corral, et al., Accuracy of body mass index in diagnosing obesity in the adult general population, International journal of obesity 32 (2008) 959–966.

[15] D. Angelis, F. Sofos, T. E. Karakasidis, Artificial intelligence in physical sciences: Symbolic regression trends and perspectives, Archives of Computational Methods in Engineering 30 (2023) 3845–3865.

[16] J. R. Koza, Genetic programming as a means for programming computers by natural selection, Statistics and computing 4 (1994) 87–112.

[17] W. La Cava, P. Orzechowski, B. Burlacu, F. O. de França, M. Virgolin, Y. Jin, M. Kommenda, J. H. Moore, Contemporary symbolic regression methods and their relative performance, arXiv preprint arXiv:2107.14351 (2021).

[18] C. Wilstrup, J. Kasak, Symbolic regression outperforms other models for small data sets, arXiv preprint arXiv:2103.15147 (2021).

[19] B. Burlacu, G. Kronberger, M. Kommenda, M. Affenzeller, Parsimony measures in multi-objective genetic programming for symbolic regression, in: Proceedings of the genetic and evolutionary computation conference companion, 2019, pp. 338–339.

[20] Q. Chen, B. Xue, M. Zhang, Rademacher complexity for enhancing the generalization of genetic programming for symbolic regression, IEEE transactions on cybernetics 52 (2020) 2382–2395.

[21] D. J. Bartlett, H. Desmond, P. G. Ferreira, Exhaustive symbolic regression, IEEE Transactions on Evolutionary Computation (2023).

[22] J. Kubalík, E. Derner, R. Babuška, Symbolic regression driven by training data and prior knowledge, in: Proceedings of the 2020 Genetic and Evolutionary Computation Conference, 2020, pp. 958–966.

[23] C. Haider, F. O. de França, B. Burlacu, G. Kronberger, Using shape constraints for improving symbolic regression models, arXiv preprint arXiv:2107.09458 (2021).

[24] J. Kubalík, E. Derner, R. Babuška, Multi-objective symbolic regression for physics-aware dynamic modeling, Expert Systems with Applications 182 (2021) 115210.

[25] D. Ferrari, V. Guidetti, F. Mandreoli, Multi-objective symbolic regression for data-driven scoring system management, in: 2022 IEEE International Conference on Data Mining (ICDM), IEEE, 2022, pp. 945–950.

[26] M. Virgolin, S. P. Pissis, Symbolic regression is np-hard, Transactions on Machine Learning Research (2022).

[27] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on pattern analysis and machine intelligence 27 (2005) 1226–1238.

[28] M. Yamada, J. Tang, J. Lugo-Martinez, E. Hodzic, R. Shrestha, A. Saha, H. Ouyang, D. Yin, H. Mamitsuka, C. Sahinalp, et al., Ultra high-dimensional nonlinear feature selection for big biological data, IEEE Transactions on Knowledge and Data Engineering 30 (2018) 1352–1365.

[29] Z. Zhao, R. Anand, M. Wang, Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform, in: 2019 IEEE international conference on data science and advanced analytics (DSAA), IEEE, 2019, pp. 442–452.

[30] V. Guidetti, F. Motta, J. Milic, D. Ferrari, F. Mandreoli, G. Guaraldi, Unlocking frailty index: Knowledge distillation with symbolic machine learning to simplify frailty assessment in standard hiv clinics, *Under review* (2024).

[31] D. Ferrari, V. Guidetti, Y. Wang, V. Curcin, Multi-objective symbolic regression to generate data-driven, non-fixed structure and intelligible mortality predictors using ehr: Binary classification methodology and comparison with state-of-the-art, in: AMIA Annual Symposium Proceedings, volume 2022, American Medical Informatics Association, 2022, p. 442.

[32] V. Guidetti, G. Dolci, E. Franceschini, E. Bacca, G. J. Burastero, D. Ferrari, V. Serra, F. Di Benedetto, C. Mussini, F. Mandreoli, Death after liver transplantation: Mining interpretable risk factors for survival prediction, in: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2023, pp. 1–10.

[33] L. Cao, H. Chen, X. Fan, J. Gama, Y.-S. Ong, V. Kumar, Bayesian federated learning: A survey, arXiv preprint arXiv:2304.13267 (2023).

[34] J. Dong, J. Zhong, W.-N. Chen, J. Zhang, An efficient federated genetic programming framework for symbolic regression, IEEE Transactions on Emerging Topics in Computational Intelligence (2022).

[35] D. Nguyen Duy, M. Affenzeller, R. Nikzad-Langerodi, Towards vertical privacy-preserving symbolic regression via secure multiparty computation, in: Proceedings of the Companion Conference on Genetic and Evolutionary Computation, 2023, pp. 2420–2428.