DESY-22-120

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

Davide Ferrari MSc[1, 2, 3], Veronica Guidetti PhD[4], Yanzhong Wang PhD[1,3], Vasa Curcin PhD[1, 3]

[1] School of Population Health and Environmental Sciences, King's College London, London, UK; [2] Centre for Clinical Infection & Diagnostics Research, St. Thomas' Hospital, London, UK; [3] NIHR Biomedical Research Centre, Guy's and St Thomas' NHS Foundation Trust and King's College London, London, UK; [4] Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

Abstract

Symbolic Regression (SR) is a data-driven methodology based on Genetic Programming, and it is widely used to produce arithmetic expressions for modelling learning tasks. Compared to other popular statistical techniques, SR outcomes are given by an arbitrary set of mathematical operations, representing arbitrarily complex linear and non-linear functions without a predefined fixed structure. Another advantage is that, unlike other machine learning algorithms, SR produces interpretable results. In this paper, we explore the qualities and limitations of this technique in a novel implementation as a binary classifier for in-hospital or short-term mortality prediction in patients with Covid-19. Our results highlight that SR provides a competitive alternative to popular statistical and machine learning methodologies to model relevant clinical phenomena thanks to good classification performance, stability in unbalanced dataset management, and intrinsic interpretability.

Introduction

Mortality is one of the most popular clinical outcomes used in medical statistics, both for clinical and methodological reasons. On the one hand, it is the most extreme and severe clinical outcome, and there is a huge interest in understanding how it can be predicted and prevented. On the other hand, it is one of the most accessible outcomes to model from a statistical point of view, as the phenotyping rules determining a patient's death are the most unequivocal. In standard applications, mortality is interpreted as a binary classification problem, the simplest classification form. In medical applications, the most common approaches used for mortality prediction include Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF).

In clinical applications, especially those that influence healthcare professionals' decision-making process, interpretability is a key element that provides an overview of how the predictive model generates results. This transparency became a necessary aspect of all healthcare applications due to the policies introduced by data protection regulations, the European[1] and the UK GDPR[2]. Approaches like LR are intrinsically interpretable as they produce a fixed-structure explicit mathematical formula, whereas most of the others are not and need dedicated techniques like SHAP[3] and LIME[4] to be interpreted.

Apart from the clinical interest in mortality risk prediction, this task is de-facto a widespread benchmark for statistical learning approaches as it is easily modellable in several clinical databases and can provide a reliable comparison of methods' performance. In this manuscript, we propose a novel approach based on Symbolic Regression[5] (SR) to implement this binary classification and compare it to other popular alternatives using standard predictive performance measures, and ease and clarity of model interpretation. Our implementation of SR is open source and available at https://github.com/davideferrari92/multiobjective_symbolic_regression.

1) Symbolic Regression and Genetic Programming

Genetic Programming (GP) is a technique inspired by biology where a set of programs are subject to mutations and alterations like those that the genome undergoes in passing from one generation to the next. The most performing individuals are kept in the genetic pool to propagate their qualities, helping new generations to produce better performance. SR is a GP-based approach aiming to find a set of mathematical expressions that best fit a prediction task and apply GP alterations to the pool of expressions keeping only the best ones. Unlike other fixed-model approaches, such as LR, the shape of the formula in SR is dynamical and can include any desired mathematical operator, including non-linear ones.

The SR algorithm starts generating a population of n independent individuals where operators and operands in the starting formula are randomly chosen. Internal nodes are given by operations (Operation Nodes, or OpNode), while the leaves are terminal nodes with variables or constants (Feature Nodes or FeatNode). Operations are chosen from a predefined set based on what possibilities the experiments aim to explore; in our case, the allowed operations are addition, subtraction, multiplication, division, natural logarithm, exponential, power, square root, maximum and minimum. The terminal nodes can be variables from the training dataset or constants. The constants are initialized randomly in each expression, but their value is adjusted using a method we will describe further in the manuscript. At each training generation, a pool of k individuals are randomly selected from the population and, through a tournament selection process, one of the expressions is duplicated and subject to a genetic mutation; we call this new individual *offspring*; this process is repeated n times doubling the dimension of the population. Only the n best individuals are then kept in the pool.

Genetic operations are always applied to a random node of the offspring. Our list of genetic operations include *point mutation* (an OpNode, and its sub-tree are replaced by a newly generated sub-tree), *crossover* (an OpNode and its sub-tree is replaced by an OpNode and its sub-tree from another individual of the pool), *node insertion* (an new OpNode is inserted at a randomly selected point in the tree making it deeper of one level), *node deletion* (a random OpNode is deleted and its operators are shifted one level above), *leaf mutation* (a random FeatNode is replaced by another from the allowed ones), *operator mutation* (a random OpNode operation is replaced by another with the same number of operators, i.e., arity) and *simplification* (the individual can be a non-minimal representation of an expression and therefore, although being numerically equivalent, a reorganized and simplified tree is considered in place of the selected one). In our setup, all operations happen with the same probability.

1.1) Related Works

In recent years, evolutionary approaches, and SR in particular, have proven to be competitive methodologies in the data science scenario. Clinical applications can be found in metabolic medicine[6–11], oncology[12,13], cardiology [14], and others[15–19]. Mortality prediction is of wide interest, and multiple approaches have been developed to predict this outcome: in intensive care, there are commonly used scoring systems like APACHE[20], SAPS[21,] and SOFA[22]. Besides the most common machine learning techniques, the search for effective predictive tools using GP is conducted using approaches like those presented in [13,23–25].

The interpretability of SR expressions may be easier than other ML techniques as the mathematical formula is explicitly defined. Nevertheless, understanding its clinical meaning is a complicated and open challenge. This is a relatively unexplored topic, and a further joint effort by both the computing and the clinical experts is required. To the best of our knowledge, this is the first time an SR is explicitly implemented and trained as a binary classifier for mortality prediction on structured electronic health records.

1.2) Binary Classification in Symbolic Regression

For the optimization of a binary classification task, we used the Binary Cross-Entropy (BCE) evaluated on all m datapoints of the training dataset as fitness function to be minimized. In Equation 1 we report the BCE definition ($\hat{y}$ is the ground truth and y is the predicted probability for the i-th data point in a dataset of m samples), in Equation 2 the logistic (sigmoid) function used to scale the expression result between 0 and 1 for the binary classification and in Figure 1 is depicted one example of expression. Given the remarkable imbalance of the dataset, described in detail in the next section, all the training procedures need to be weighted on the prevalence of the two classes to have a more appropriate behaviour of the predictive model. We therefore adopted a Weighted Binary Cross-Entropy (WBCE) in both our experiments. For the sake of simplicity, for the rest of the manuscript we will refer to WBCE just as BCE.

$$BCE = -\frac{1}{m}\sum_{i=1}^{m}[y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)]$$

Equation 1 Binary Cross-Entropy

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Equation 2 Logistic (sigmoid) function

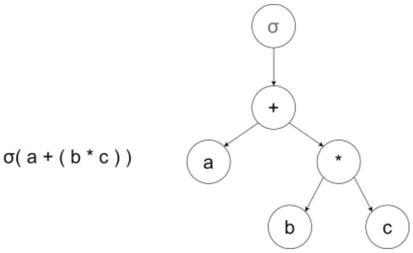

Figure 1 Expression with sigmoid function as root node

1.3) Multi-objective training with Pareto optimality

Symbolic Regression is an algorithm that rewards the best individuals and propagates them over the generations. In single-objective optimization, individuals are sorted from lowest to highest fitness function values, e.g., BCE, at the end of each generation. Instead, in multi-objective optimization, population ranking is implemented in such a way as to support more than one fitness function at a time. Therefore, program ranking cannot be based on a single *fitness function* (or *objectives*) and must rely on more complex criteria. We implemented the *Non-dominated Sorting Genetic Algorithm II (NSGA-II)* [26,27], an elitist multi-objective genetic algorithm technique where program ranking is based on non-dominance criterium and crowding distance between individuals belonging to the same optimality ranking. Given n fitness functions, each individual is represented by an n-dimensional vector whose coordinates are given by its fitness function values. In NSGAII, one individual is said to dominate another one if all fitness values of the former are smaller than or equal to those of the latter and at least one is strictly smaller. Following the Pareto optimality principle [28], non-dominated individuals are attributed to rank 1, which is indeed called "*first Pareto front*". To identify rank 2 (second Pareto front) individuals, the first Pareto front is subtracted from the population and the non-dominated individuals are sought; every rank gathers individuals not dominating each other. The process is repeated until all individuals are attributed to a Pareto front. In a Pareto optimality ranking algorithm, individuals belonging to the same rank $\boldsymbol{R_i}$ are considered equally optimal although their different behaviour needs to be critically evaluated to choose the desired expression or set of expressions. Therefore, to further discriminate between individuals of the same rank, NSGA-II leverages the concept of crowding distance according to which the more "isolated" individuals, that is, the one belonging to a less dense portion of the Pareto front, should be privileged among other non-dominating individuals of the same rank. Figure 2 depicts one generation of the training process.

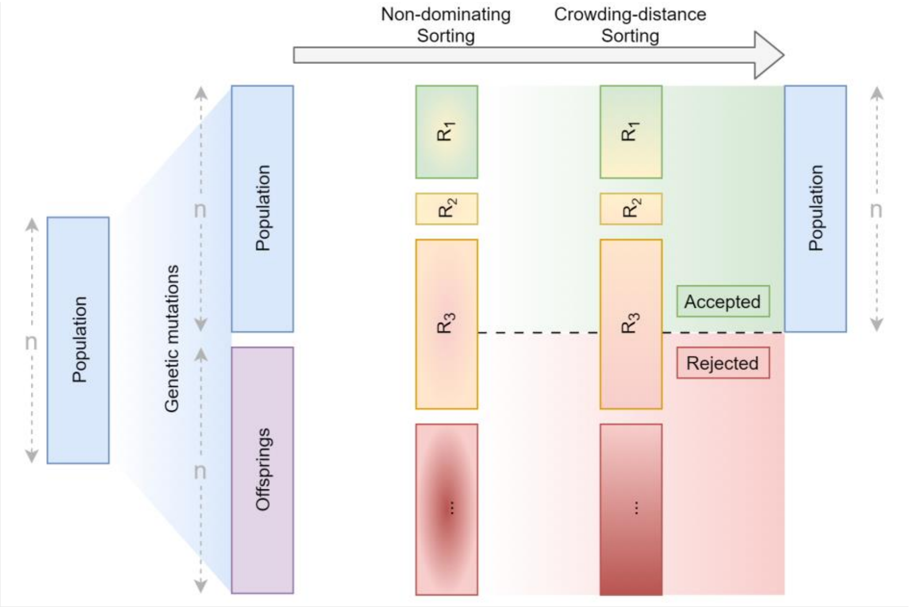

Figure 2 NSGA-II and Crowding distance algorithm schema

1.4) Neuron-based constants optimization

All terminal nodes in an expression can be either features from the training dataset or numerical constants. As for the constants, even assuming that the structural form of the expression is already the best one to fit the learning task, it is rather unlikely that the random values provided by the GP algorithm are the best ones. Therefore, we use an optimization process to update the numerical values inside each expression. To do so, we create a one-neuron neural network in which the activation function is the expression itself (hence the need for a differentiable method), and the trainable parameters are the constants to be optimized. The parameters are initialized to the same value as the constants inside the expression. Before evaluating the fitness of the expression on the training set, we train the neuron for 10 epochs at every generation using weighted BCE loss function and ADAM optimizer[29], i.e., a stochastic gradient descent method based on adaptive estimation of first order and second order moments. At the end of training, the neuron parameters represent the optimized constant values that are then replaced inside each expression. Iterating this process among consecutive generations will eventually converge to the best possible values for all the constants in each expression.

2) The learning task, inclusion criteria, and the dataset selection

The experimental dataset we used is constituted of a population of 2400 individual patients admitted to St. Thomas' Hospital in London with a laboratory-confirmed diagnosis of Covid-19. Each data point reports only the first collected clinical data of each patient as the goal of this task is to predict the short term in-hospital mortality risk of patients at hospital admission. The variables available in the dataset include demographics, signs and symptoms, laboratory results, and past medical history.

One of the main drawbacks of SR is that it tends to perform poorly on very high-dimensional data, for this reason we pre-process our dataset using non-linear feature selection methods to improve the model efficiency and increase its generalization [30]. We focus on two well-known methods that use both input-output and input-input relevance, i.e., *Hilbert-Schmidt Independence Criterion Lasso (HSIC Lasso)*[31,32]. It aims to find the minimal and optimal subset of features to explain a given phenomenon. The HSIC Lasso method finds the global optimal features subset solving a feature-wise kernelized Lasso optimization problem with a non-negative constraint. With this process we selected 10 features among the 25 available; they are described in Table 1.

Table 1 Description of the 10 selected features

Feature		***Total***	***Class 0 (survived)***	***Class 1 (dead)***	***p-value***
Population (%)		2400 (100%)	2105 (87.71 %)	295 (12.29 %)	
Demographic (continuous)	**Age, mean (± SD) [No.]**	59.77 (20.58) [2400]	57.75 (20.42) [2105]	74.14 (15.36) [295]	< 0.001
Laboratory Data (continuous)	**C-Reactive Protein, mean (± SD) [No.]**	82.08 (81.53) [2400]	76.89 (78.1) [2105]	119.09 (94.99) [295]	< 0.001
	Creatinine, mean (± SD) [No.]	121.76 (145.7) [2400]	116.37 (144.42) [2105]	160.26 (149.19) [295]	< 0.001
	D-Dimer, mean (± SD) [No.]	3.66 (9.0) [2400]	3.39 (8.39) [2105]	5.66 (12.39) [295]	< 0.001
	Neutrophils, mean (± SD) [No.]	6.0 (3.8) [2400]	5.8 (3.62) [2105]	7.41 (4.69) [295]	< 0.001
	Platelets, mean (± SD) [No.]	232.39 (103.85) [2400]	235.03 (103.45) [2105]	213.54 (104.91) [295]	< 0.001
	Urea, mean (± SD) [No.]	7.94 (6.44) [2400]	7.2 (5.62) [2105]	13.28 (8.92) [295]	< 0.001
Past medical history (binary)	**Cardiovascular Diseases, False (%)**	1651 [68.79 %]	1528 [72.59 %]	123 [41.69 %]	< 0.001
	Cardiovascular Disease, True (%)	749 [31.21 %]	577 [27.41 %]	172 [58.31 %]	
	COPD*, False (%)	2223 [92.62 %]	1970 [93.59 %]	253 [85.76 %]	< 0.001
	COPD*, True (%)	177 [7.38 %]	135 [6.41 %]	42 [14.24 %]	
Signs (continuous)	**Respiratory Rate, mean (± SD) [No.]**	20.64 (5.0) [2400]	20.41 (4.8) [2105]	22.24 (6.04) [295]	< 0.001

*COPD - Chronic obstructive pulmonary disease

The available dataset includes one record for each of the 2400 patients and has been split with an 80/20 ratio into a training (1920) and a test (480) dataset; we stratified the split on the target variable to preserve class proportion. To limit the negative effects of the imbalance on the predictive models, we assigned a weight to all the samples proportional to the prevalence of the other class whenever the algorithm allowed.

3) Generation of the binary classifiers with Symbolic Regression

The binary classification task is trained to minimize the BCE on the training set, although we also leveraged the possibility of a multi-objective training process for SR and conducted a second experiment with an additional objective: the maximization of the average precision score on the training set. In the following section, we will introduce the optimal expressions coming from the two approaches and present their performance. We train both models on a population of 200 randomly generated expressions that are evolved through 1000 generations without early stopping conditions.

3.1) Two new mortality predictors

The two resulting expressions are reported in Equation 3 and Equation 4. Both must be post-processed with the logistic function to generate the classification probability.

Equation 3 Expression generated by single-objective SR

$$\begin{aligned} &0.719 + CardiovascularDiseases \\ &\qquad + min\left(2 \cdot log(Urea) - log\left(Creatinine \cdot Urea + \sqrt{2 \cdot Platelets} \cdot (CRP + Neutrophils)\right), 1.623\right) \end{aligned}$$

Equation 4 Expression generated by multi-objective SR

$$-3.094 + CardiovascularDiseases + 1.630 \cdot min\big(log(\,0.7 \cdot Urea), log(CRP)\big)$$

The multi-objective approach based on Pareto optimality produces a set of equally optimal individuals; this implies that choosing the final model is a nontrivial task, as the expressions show different behaviour. We decided to represent here only the expression with the lowest BCE but, depending on the goal of the learning task, a different expression could be preferred. Many factors can influence this decision, and the number of variables used can be one of them: in real-world scenarios where collecting variables can be expensive and technically complicated, choosing a model with few variables, or with cheaper ones, or both, but with equal or comparable optimality towards the learning task can favour its daily clinical use. Moreover, the complexity of the expression can make the model clinically hard to interpret, regardless of its performance. Therefore, choosing a simpler formula, like Equation 4 over Equation 3, leads to a more straightforward clinical interpretation, increasing the final user's trust. Finally, the analysis of equally optimal expressions can also be used to identify recurrent (and thus potentially meaningful) sub-expression; the same structure (e.g., $constant + CardiovascularDiseases$) found in multiple expressions could lead to the discovery of "*first principle*" mechanisms underlying a physiological behaviour that will help to unlock new clinical knowledge. The interpretation of the presented expression is still an open challenge and will not be discussed further in this manuscript.

The performance measures are expressed by means of the Area Under the Receiver Operating Characteristics (AUROC), Accuracy, Precision, Sensitivity, and F-score defined in Table 2. All the performances presented in this manuscript are calculated for each model on the same test dataset.

Table 2 Classification performance definition

Accuracy	**Precision** **(Positive Predicted Values)**	**Sensitivity** **(True Positives Rate)**	**F-score**
$\frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}}$	$\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$	$\frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$	$\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\frac{\boldsymbol{FP+FN}}{\boldsymbol{2}}}$
$\boldsymbol{TP}$ - True Positives	$\boldsymbol{TN}$ - True Negative	$\boldsymbol{FP}$ - False Positive	$\boldsymbol{FN}$ - False Negative

3.2) Comparison with machine learning algorithms

To contextualize our implementation of a mortality predictor, we implemented other known machine learning algorithms with the ambition of comparing their predictive performance and assessing the validity of our proposal. All the models have been trained on the same training dataset and validated on the same test dataset we used for the SR. Each algorithm has been tuned by means of a standard grid search on its hyperparameters and has been evaluated with a 10-fold cross-validation method. We report the average predictive performance measures on the test dataset in Table 3. The ROC curves for all the models are shown in Figure 3.

Table 3 Classification performance comparison with machine learning algorithms

Model	***AUROC***	***Accuracy***	***Precision (PPV)***	***Sensitivity (TPR)***	***F-score***
Symbolic Regression	**0.84**	0.77	0.32	0.77	**0.45**
Multi-objective Symbolic Regression *	**0.82**	0.72	0.27	0.76	**0.40**
Decision Tree	0.50	0.16	0.12	**0.93**	0.21
Random Forest	0.75	0.71	0.26	0.72	0.38
Gradient Boosting	0.45	0.79	0.06	0.05	0.05
Light Gradient Boosting	0.70	0.83	0.30	0.25	0.27
Extreme Gradient Boosting	0.69	0.84	0.29	0.20	0.24
Ada Boost	0.76	**0.87**	-	-	-
Linear Discriminant Analysis	0.56	**0.87**	-	-	-

* Trained minimizing BCE and maximizing the average precision score in a multi-objective approach

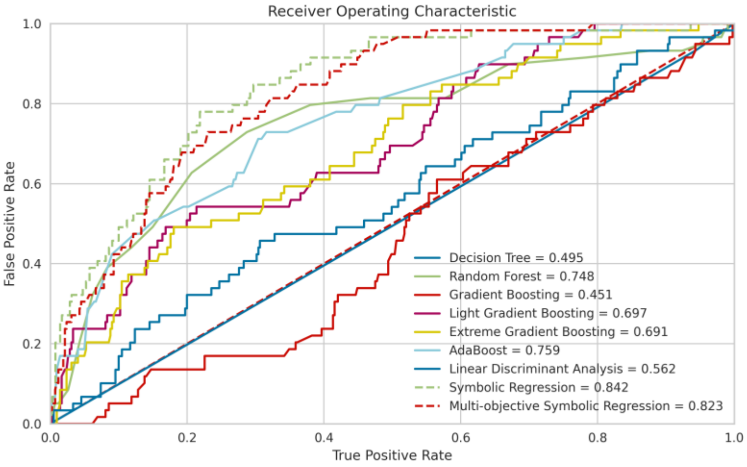

Figure 3 ROC Curve for SR and ML models

In healthcare, not all misclassifications are equally wrong; in this case, an FN (a patient classified as safe, but who will die) is a much more dangerous prediction than an FP (a patient classified as high risk, but who will survive). In a preventive medicine mindset, we believe predictive models should be more conservative on FNs and avoid this clinically dangerous error. Therefore, maximization of sensitivity should be preferred over accuracy or precision. The most evident behaviour is the poor handling of data imbalance: standard tree-based algorithms tend to produce better accuracy, but the few samples belonging to class 1 were systematically misclassified (low precision and, especially, low sensitivity) regardless of sample weighting. AdaBoost and LDA did not handle data imbalance at all and only learned to predict class 0 (the percentage of samples in that class is indeed about 0.87, as their accuracy). On the other hand, both SR models provide the best AUROC and sensitivity measures demonstrating better handling of this phenomenon.

An alternative approach for the comparison could have been to use data augmentation (such as SMOTE[33]) to balance the classes; however, we have chosen not to do so to avoid the risk of adding spurious information to the data. In fact, this exploratory work aims to compare the benchmarks of various algorithms on real clinical data as it is collected in hospital wards. We have not discussed data imputation techniques in SR as the dataset under study was complete; nonetheless, we want to emphasize that the topic has been addressed in the literature, comparing the effect of various existing techniques, and developing new ones[34].

We can conclude that our implementation of SR, single-objective and multi-objective, is a valid and competitive alternative to standard ML algorithms. Most importantly, SR can handle data imbalances much better, provides an improved sensitivity measure, and grants better care for the few samples belonging to the most critical clinical group. From these results, it seems that the training purely based on BCE minimization could be a better choice than a multi-objective one, which also aims to maximize the average precision score. However, suppose a marginal decrease in performance is bearable for the final application. In that case, multi-objective SR can be preferred as it produces a set of optimal solutions from which to choose. Comparing the formulas presented in this work, the one arising from multi-objective optimization is simpler. This may be preferable as it favours clinical interpretation at the cost of slightly lower performance. This experiment highlights the existence of more than one suitable solution with comparable performance, introducing the possibility (and responsibility) of choosing the most suitable approach for each application

3.3) Comparison with other known predictive scores

This manuscript aims to support the competitiveness of SR as a methodology for creating interpretable and performing mortality classifiers. Much more than the clinical interpretation of the compared indices, this section focuses on the performance of our approach against the more common LR odds ratios. Also, we used a systematic literature review[35] and related exvalidation study[36] of predictive models for Covid-19 to identify known risk scores for predicting short-term mortality in hospitalized patients. We included models developed on patient data that used variables available to us, so to validate these approaches on our test dataset. The performance measures can be found in Table 4. In Figure 4 we show the ROC curves for all the scores.

Table 4 Classification performance comparison with other mortality risk scores

Model	*AUROC*	*Accuracy*	*Precision (PPV)*	*Sensitivity (TPR)*	*F-score*
Symbolic Regression §	**0.84**	0.77	0.32	**0.77**	**0.45**
Multi-objective Symbolic Regression §	0.82	0.72	0.27	**0.76**	0.40
Logistic Regression Odds Ratios §	**0.85**	**0.89**	**0.75**	0.20	0.32
Bello-Chavolla Model [37] Φ	0.65	0.63	0.22	0.48	0.30
Hu Model [38]	0.65	0.64	0.33	0.65	0.44
Zhang DCS Model [39]	0.72	0.82	0.33	0.05	0.09
Zhang DCSL Model [39]	0.66	0.80	0.10	0.02	0.04
Wang Model [40]	0.76	0.74	0.33	0.56	0.42
Knights 4C Model [41]	0.81	0.86	0.70	0.32	0.44

§ Trained on our training dataset and validated on our test dataset
Φ Post-processing: scaled between 0 and 1

Using our dataset alone undoubtedly biases our evaluation of existing risk scores. Indeed, our training and test datasets are more similar to each other than to the dataset on which other scores have been created. For this reason, it is not surprising to find better performances. Nevertheless, as happened with ML models, our SR model produces better overall performance in AUROC, sensitivity, and F-score and comparable accuracy and precision. The LR Odds Ratio approach still performs better in classifying the most prevalent class, but it also shows poor sensitivity. Given the importance of its maximisation, we can safely say that SR is a preferable alternative to other more traditional approaches for high-stake domains like medicine.

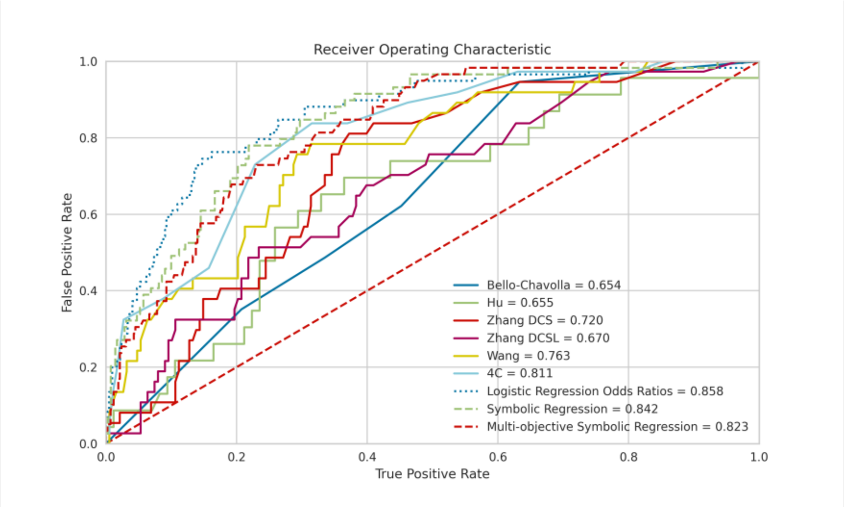

Figure 4 ROC Curve for SR and the other mortality risk scores

4) Conclusions and future developments

This paper presented a way to use SR as a binary classifier for mortality prediction in patients with Covid-19. We proposed two alternative implementations, one trained to minimize the BCE and another using multi-objective optimization to minimize the BCE and maximize the average precision score. This manuscript successfully shows the potential of SR as a competitive learning methodology to be used in clinical scenarios thanks to its good performance, intrinsic interpretability, and its ability to model arbitrarily complex expressions. Further research should be devoted to the model interpretation process. Indeed, the generated expressions do not have a fixed structure and can contain non-linearities; this requires an additional joint effort from clinicians and data scientists to contextualize each expression in the clinical environment properly. Furthermore, some studies in the literature confirm that the use of a multi-objective approach largely prevents premature convergence of the algorithm, ensuring that the search for optimal solutions does not stop too soon[42]. A major limitation of this study is its single-hospital design. This implies that the population examined is relatively homogeneous, and the overfitting risk increases. A relevant improvement would be given by a multi-centre application, especially if equipped with a distributed training process for improved privacy preservation. In addition, further analysis should be done to better understand the behaviour of SR, especially when trained on unbalanced datasets. Finally, the choice of the best expression, or set of expressions, is an important challenge that needs to be studied case by case, depending on the desired behaviour of the model at any given time.

Ethics, funding, and acknowledgements

Ethical approval for data informatics was granted by The London Bromley Research Ethics Committee (reference (20/HRA/1871)) to the King's Health Partners Data Analytics and Modelling COVID-19 Group to collect clinically relevant data points from patient's electronic health records. This research was part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre programme of Infection and Immunity (RJ112/N027) based at Guy's and St Thomas' National Health Service NHS Foundation Trust, King's College London, and DRIVE-Health, KCL funded Centre for Doctoral Training (CDT) in Data-Driven Health. Veronica Guidetti is supported by the ERC Consolidator Grant STRINGFLATION under the HORIZON 2020 grant agreement no. 647995. The authors acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the National Institute for Health Research (NIHR) Biomedical Research Centres at South London & Maudsley and Guy's & St. Thomas' NHS Foundation Trusts and NIHR Applied Research Collaboration (ARC) South London at King's College Hospital (KCH) NHS Foundation Trust and King's College London, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.
We express our gratitude to Dr. Wenjuan Wang (KCL) for the support provided during the data pre-processing at the early stages of the study.

References

1. GPDR.eu. General Data Protection Regulation (GDPR) Compliance Guidelines. *Gpdr.Eu* https://gdpr.eu/ (2020).
2. UK GDPR Updated for Brexit | UK GDPR. https://uk-gdpr.org/.
3. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017-Decem**, 4766–4775 (2017).
4. Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should i trust you?” Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-Augu**, 1135–1144 (2016).
5. Augusto, D. A. & Barbosa, H. J. C. Symbolic regression via genetic programming. *Proceedings - Brazilian Symposium on Neural Networks, SBRN* **2000-January**, 173–178 (2000).
6. de Falco, I. *et al.* An evolutionary methodology for estimating blood glucose levels from interstitial glucose measurements and their derivatives. *Proceedings - IEEE Symposium on Computers and Communications* **2018-June**, 1158–1163 (2018).
7. Contador, S., Ignacio Hidalgo, J., Garnica, O., Manuel Velasco, J. & Lanchares, J. Can clustering improve glucose forecasting with genetic programming models? *GECCO 2019 Companion - Proceedings of the 2019 Genetic and Evolutionary Computation Conference Companion* 1829–1836 (2019) doi:10.1145/3319619.3326809.
8. Contactor, S., Velasco, J. M., Garnica, O. & Hidalgo, J. I. Profiled glucose forecasting using genetic programming and clustering. *Proceedings of the ACM Symposium on Applied Computing* 529–536 (2020) doi:10.1145/3341105.3374003.
9. Hidalgo, J. I. *et al.* Data Based Prediction of Blood Glucose Concentrations Using Evolutionary Methods. *Journal of Medical Systems* **41**, 1–20 (2017).
10. Colmenar, J. M. *et al.* Predicting glycemia in diabetic patients by evolutionary computation and continuous glucose monitoring. *GECCO 2016 Companion - Proceedings of the 2016 Genetic and Evolutionary Computation Conference* 1393–1400 (2016) doi:10.1145/2908961.2931734.
11. Golap, M. A. uddowla, Raju, S. M. T. U., Haque, M. R. & Hashem, M. M. A. Hemoglobin and glucose level estimation from PPG characteristics features of fingertip video using MGGP-based model. *Biomedical Signal Processing and Control* **67**, 102478 (2021).
12. van der Meer, M. C. *et al.* Better and faster catheter position optimization in HDR brachytherapy for prostate cancer using multi-objective real-valued GOMEA. *GECCO 2018 - Proceedings of the 2018 Genetic and Evolutionary Computation Conference* 1387–1394 (2018) doi:10.1145/3205455.3205505.
13. Vanneschi, L. *et al.* Identification of individualized feature combinations for survival prediction in breast cancer: A comparison of machine learning techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6023 LNCS**, 110–121 (2010).
14. Wilstup, C., Cave, C. & Abzu, †. Combining symbolic regression with the Cox proportional hazards model improves prediction of heart failure deaths. *medRxiv* 2021.01.15.21249874 (2021) doi:10.1101/2021.01.15.21249874.
15. Tao, Y., Zhang, Y. & Jiang, B. Evolutionary learning-based modeling for warfarin dose prediction in Chinese. *GECCO 2017 - Proceedings of the Genetic and Evolutionary Computation Conference Companion* 1380–1386 (2017) doi:10.1145/3067695.3082492.
16. Hughes, J. A., Houghten, S. & Brown, J. A. Models of Parkinson’s Disease Patient Gait. *IEEE Journal of Biomedical and Health Informatics* **24**, 3103–3110 (2020).
17. Hughes, J. A., Houghten, S. & Brown, J. A. Descriptive Symbolic Models of Gaits from Parkinson’s Disease Patients. *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2019* (2019) doi:10.1109/CIBCB.2019.8791459.
18. Moore, J. H., Parker, J. S. & Hahn, L. W. Symbolic discriminant analysis for mining gene expression patterns. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2167**, 372–381 (2001).
19. D’Angelo, G., Pilla, R., Tascini, C. & Rampone, S. A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees. *Soft Computing* **23**, 11775–11791 (2019).

20. Zimmerman, J. E., Kramer, A. A., McNair, D. S. & Malila, F. M. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Critical Care Medicine* **34**, 1297–1310 (2006).

21. Gall, J. R., Lemeshow, S. & Saulnier, F. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. *JAMA: The Journal of the American Medical Association* **270**, 2957–2963 (1993).

22. Toma, T., Abu-Hanna, A. & Bosman, R. J. Discovery and inclusion of SOFA score episodes in mortality prediction. *Journal of Biomedical Informatics* **40**, 649–660 (2007).

23. Chan, C. L. & Ting, H. W. Constructing a novel mortality prediction model with Bayes theorem and genetic algorithm. *Expert Systems with Applications* **38**, 7924–7928 (2011).

24. Jiménez, F., Sánchez, G. & Juárez, J. M. Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. *Artificial Intelligence in Medicine* **60**, 197–219 (2014).

25. Bannister, C. A., Halcox, J. P., Currie, C. J., Preece, A. & Spasić, I. A genetic programming approach to development of clinical prediction models: A case study in symptomatic cardiovascular disease. *PLOS ONE* **13**, e0202685 (2018).

26. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6**, 182–197 (2002).

27. Yusoff, Y., Ngadiman, M. S. & Zain, A. M. Overview of NSGA-II for optimizing machining process parameters. *Procedia Engineering* **15**, 3978–3983 (2011).

28. Smits, G. F. & Kotanchek, M. Pareto-Front Exploitation in Symbolic Regression. *Genetic Programming Theory and Practice II* 283–299 (2006) doi:10.1007/0-387-23254-0_17.

29. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

30. Chen, Q., Zhang, M. & Xue, B. Feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. *IEEE Transactions on Evolutionary Computation* **21**, 792–806 (2017).

31. Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. & Sugiyama, M. High-Dimensional feature selection by feature-Wise kernelized lasso. *Neural Computation* **26**, 185–207 (2014).

32. Climente-González, H., Azencott, C. A., Kaski, S. & Yamada, M. Block HSIC Lasso: Model-free biomarker detection for ultra-high dimensional data. *Bioinformatics* **35**, i427–i435 (2019).

33. Chawla, N. v., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002).

34. Al-Helali, B., Chen, Q., Xue, B. & Zhang, M. Data Imputation for Symbolic Regression with Missing Values: A Comparative Study. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020* 2093–2100 (2020) doi:10.1109/SSCI47803.2020.9308216.

35. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, 26 (2020).

36. Jong, V. M. T. de *et al.* Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis. *BMJ* **378**, e069881 (2022).

37. Bello-Chavolla, O. Y. *et al.* Predicting Mortality Due to SARS-CoV-2: A Mechanistic Score Relating Obesity and Diabetes to COVID-19 Outcomes in Mexico. *Journal of Clinical Endocrinology and Metabolism* **105**, 2752–2761 (2020).

38. Hu, C. *et al.* Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *International Journal of Epidemiology* **49**, 1918–1929 (2020).

39. Zhang, H. *et al.* Risk prediction for poor outcome and death in hospital in-patients with COVID-19: Derivation in Wuhan, China and external validation in London, UK. *medRxiv* 2020.04.28.20082222 (2020) doi:10.1101/2020.04.28.20082222.

40. Wang, K. *et al.* Clinical and Laboratory Predictors of In-hospital Mortality in Patients with Coronavirus Disease-2019: A Cohort Study in Wuhan, China. *Clinical Infectious Diseases* **71**, 2079–2088 (2020).

41. Knight, S. R. *et al.* Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *The BMJ* **370**, 22 (2020).

42. Mahrach, M., Miranda, G., León, C. & Segredo, E. Comparison between Single and Multi-Objective Evolutionary Algorithms to Solve the Knapsack Problem and the Travelling Salesman Problem. *Mathematics 2020, Vol. 8, Page 2018* **8**, 2018 (2020).