RegScore: Scoring Systems for Regression Tasks

Michal K. Grzeszczyk[1,2], Tomasz Szczepański[1], Pawel Renc[2,3], Siyeop Yoon[2], Jerome Charton[2], Tomasz Trzciński[4,5], and Arkadiusz Sitek[2]

[1] Sano Centre for Computational Medicine, Cracow, Poland
m.grzeszczyk@sanoscience.org
[2] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
[3] AGH University of Krakow, Cracow, Poland
[4] Warsaw University of Technology, Warsaw, Poland
[5] Research Institute IDEAS, Warsaw, Poland

Abstract. Scoring systems are widely adopted in medical applications for their inherent simplicity and transparency, particularly for classification tasks involving tabular data. In this work, we introduce RegScore, a novel, sparse, and interpretable scoring system specifically designed for regression tasks. Unlike conventional scoring systems constrained to integer-valued coefficients, RegScore leverages beam search and k-sparse ridge regression to relax these restrictions, thus enhancing predictive performance. We extend RegScore to bimodal deep learning by integrating tabular data with medical images. We utilize the classification token from the TIP (Tabular Image Pretraining) transformer to generate Personalized Linear Regression parameters and a Personalized RegScore, enabling individualized scoring. We demonstrate the effectiveness of RegScore by estimating mean Pulmonary Artery Pressure using tabular data and further refine these estimates by incorporating cardiac MRI images. Experimental results show that RegScore and its personalized bimodal extensions achieve performance comparable to, or better than, state-of-the-art black-box models. Our method provides a transparent and interpretable approach for regression tasks in clinical settings, promoting more informed and trustworthy decision-making. We provide our code at https://github.com/SanoScience/RegScore.

Keywords: Pulmonary Hypertension · Scoring Systems · Regression.

1 Introduction

Scoring systems are sparse linear models that require the addition of points conditioned on binary features that sum to the final score. For example, in the CHADS$_2$ [7] system, if the age of the patient is higher or equal 75 ($age \geq 75$ binary feature), 1 point is added to the final score of stroke risk. Based on the sum of points, a probability can be derived from the pre-computed table or a non-linear function. There are a large number of scoring systems in healthcare, such as CHADS$_2$ or NEWS$_2$ [21] as clinicians tend to favor methods that are easier to use and interpret, even if they are less accurate than deep learning models. The most popular approaches for data-driven scoring systems creation involve training penalized logistic regression (LR). Then, the coefficients are rounded or have ±1 assigned depending on their sign as the Unit method in [1]. Ustun and Rudin introduced the Risk Supersparse Linear Integer Model (RiskSLIM) [23] whose discrete coefficients are found with Integer Programming. Based on this approach, Multiclass Interpretable Scoring Systems (MISS) expanded scoring system use beyond two classes [10]. Liu *et al.* [17] presented FasterRisk, which finds scoring systems in a three-step process of solving sparse logistic regression via beam search (BS), finding a pool of nearly optimal solutions with continuous coefficients and rounding them. Although interpretable, traditional scoring systems may fall short when diagnosing conditions defined by thresholding a continuous variable. For example, Pulmonary Hypertension (PH) is diagnosed when the invasively measured mean Pulmonary Artery Pressure (mPAP) exceeds 20 mmHg [13]. Instead of assigning arbitrary points, a more informative approach would be to develop scoring systems that reflect the relationship between features and mPAP.

Furthermore, diagnostics in modern healthcare involve collecting multimodal data in the form of images and tabular records. This aspect led to the development of methods that allow the injection of clinical data into vision deep learning models. Modules like Dynamic Affine Feature Map Transform (DAFT) [20], TabAttention [9] or TabMixer [8] enhance the interaction between imaging and tabular data via affine transformations, attention learning conditioned on tabular data or multilayer perceptron-based mixing of multimodal features. Such methods have surpassed naïve approaches for merging both modalities based on concatenation [22], maximum value selection [24] or multiplication [5]. Tabular data can also improve unimodal models when used during self-supervised learning (SSL) [11] or for guiding image feature learning [15]. The SSL on both modalities and bimodal inference in the Tabular Image Pretraining (TIP) achieves state-of-the-art results for imaging and tabular data [4]. Unfortunately, all these approaches are weakly explainable and do not take into account the interpretability of tabular features. Even though one can analyze feature attribution [11] or attention scores of the classification (CLS) token - serving as a learned representation for classification in transformers [4] - these methods offer only limited interpretability.

In this paper, we present RegScore, a sparse, interpretable, transparent scoring system for regression tasks. By relaxing integer-only constraints of points in scoring systems and changing the task from classification to regression, we show two ways to create RegScore. Firstly, we find the solution to a sparse ridge regression problem on binary features based on BS [17]. The second approach is to solve it with OKRidge (OKR) [16]. Further, we leverage the interpretability of tabular data and show how to produce interpretable predictions for bimodal deep learning models. Given the CLS token from the TIP transformer, instead of computing the final output, we generate weights for linear regression computed with tabular features. We dub this approach Personalized Linear Regression (PLR) since separate linear operations are performed for each of the samples. Similarly, we dynamically mask binary features to produce a Personalized RegScore (PRS) from the CLS token. Our contributions are as follows: (1) we introduce RegScore, a

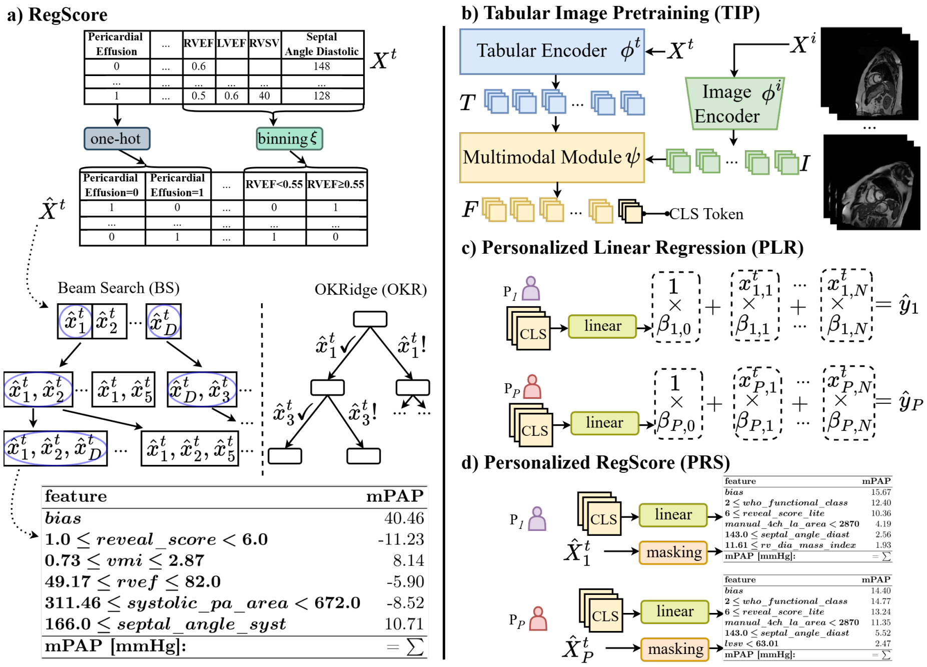
a) RegScore
Pericardial Effusion | ... | RVEF | LVEF | RVSV | Septal Angle Diastolic
0 | ... | 0.6 | | | 148
... | ... | ... | ... | ... | ...
1 | ... | 0.5 | 0.6 | 40 | 128
X^t
one-hot
binning ξ
Pericardial Effusion=0 | Pericardial Effusion=1 | ... | RVEF<0.55 | RVEF≥0.55
1 | 0 | ... | 0 | 1
... | ... | ... | ... | ...
0 | 1 | ... | 1 | 0
$\hat{X}^t$
Beam Search (BS)
$\hat{x}^t_1$ $\hat{x}^t_2$ ... $\hat{x}^t_D$
$\hat{x}^t_1, \hat{x}^t_2$ ... $\hat{x}^t_1, \hat{x}^t_5$ ... $\hat{x}^t_D, \hat{x}^t_3$ ...
$\hat{x}^t_1, \hat{x}^t_2, \hat{x}^t_D$ ... $\hat{x}^t_1, \hat{x}^t_2, \hat{x}^t_5$ ...
OKRidge (OKR)
$\hat{x}^t_1$✓ $\hat{x}^t_1$!
$\hat{x}^t_3$✓ $\hat{x}^t_3$!
feature | mPAP
bias | 40.46
$1.0 \leq reveal\_score < 6.0$ | -11.23
$0.73 \leq vmi \leq 2.87$ | 8.14
$49.17 \leq rvef \leq 82.0$ | -5.90
$311.46 \leq systolic\_pa\_area < 672.0$ | -8.52
$166.0 \leq septal\_angle\_syst$ | 10.71
mPAP [mmHg]: | $= \sum$
b) Tabular Image Pretraining (TIP)
Tabular Encoder ϕ^t ← X^t
X^i
Image Encoder ϕ^i
T
Multimodal Module ψ
I
F
CLS Token
c) Personalized Linear Regression (PLR)
P_1 CLS linear
$1 \times \beta_{1,0}$ + $x^t_{1,1} \times \beta_{1,1}$ + ... $x^t_{1,N} \times \beta_{1,N}$ = $\hat{y}_1$
P_P CLS linear
$1 \times \beta_{P,0}$ + $x^t_{P,1} \times \beta_{P,1}$ + ... $x^t_{P,N} \times \beta_{P,N}$ = $\hat{y}_P$
d) Personalized RegScore (PRS)
P_1 CLS linear
$\hat{X}^t_1$ masking
feature | mPAP
bias | 15.67
$2 \leq who\_functional\_class$ | 12.40
$6 \leq reveal\_score\_lite$ | 10.36
$manual\_4ch\_la\_area < 2870$ | 4.19
$143.0 \leq septal\_angle\_diast$ | 2.56
$11.61 \leq rv\_dia\_mass\_index$ | 1.93
mPAP [mmHg]: | $= \sum$
P_P CLS linear
$\hat{X}^t_P$ masking
feature | mPAP
bias | 14.40
$2 \leq who\_functional\_class$ | 14.77
$6 \leq reveal\_score\_lite$ | 13.24
$manual\_4ch\_la\_area < 2870$ | 11.35
$143.0 \leq septal\_angle\_diast$ | 5.52
$lvsv < 63.01$ | 2.47
mPAP [mmHg]: | $= \sum$

Fig. 1: Given the tabular (X^t) dataset, we produce the discretized set $(\hat{X}^t)$ to create RegScore with beam search or OKRidge (a). We further leverage the TIP method [4] with imaging data (X^i) (b) to compute CLS token for every sample and create interpretable regression predictions with PLR (c) and PRS (d).

scoring system for regression tasks, (2) we present PLR and PRS, two approaches for generating interpretable predictions from bimodal deep learning architecture, which while being more interpretable are competitive to other solutions, and (3) we apply the presented methods to the task of PH diagnosis and show that RegScore can outperform classification scoring systems by a significant margin.

2 Method

In this section, we present two methods for generating RegScore from tabular data, followed by a description of how the CLS token from the TIP transformer is used to derive PLR and PRS. An overview of our approach is shown in Fig. 1.

Let $(X^t = [x^t_1, \ldots, x^t_N] \in \mathbb{R}^{P \times N}, X^i \in \mathbb{R}^{P \times H \times W \times 3})$ be a dataset consisting of tabular-image pairs, where P is the number of samples and N is the number of features. We construct a binarized set $(\hat{X}^t = [\hat{x}^t_1, \ldots, \hat{x}^t_D] \in \{0, 1\}^{P \times D})$ by one-hot encoding N_{cat} categorical features and discretizing $N - N_{cat}$ continuous features using a discretization function ξ. Various implementations of ξ exist; in this work, we consider the Minimum Description Length Principle (MDLP) [6] and tertile binning. Given $\hat{X}^t$, we generate RegScore.

RegScore. The objective of RegScore is to minimize the mean squared loss under a sparsity constraint, resulting in a sparse ridge regression problem:

$$\min_{\beta} \|\mathbf{y} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq k, \tag{1}$$

Here β is the vector of weights (scores) assigned to each binary feature, λ_2 is the ℓ_2 regularization rate, and k is the model size (bias is omitted for clarity). The sparsity constraint renders the problem NP-hard. We find two methods to train RegScore. First, by relaxing integer constraints on weights in scoring systems, we adapt the BS step from FasterRisk [17] for regression. BS assumes that a model of size k inherently contains one of the best models of size $k-1$. It iteratively expands the solution by optimizing each non-zero weight and selecting the B best candidates. In the second step, BS fine-tunes non-zero weights and retains top-B solutions, ultimately returning the best model. The second approach leverages the OKRidge algorithm [16], which is based on the Branch-and-Bound (BnB) algorithm. First, a lower bound is computed for a node in the BnB tree. If this bound is lower than the current solution, a new solution is found and used to generate new nodes in the tree. Both methods yield highly effective RegScore.

Tabular Image Pretraining. To fully leverage imaging and tabular data, we employ TIP, which integrates a convolutional image encoder ϕ^i, a transformer-based tabular encoder ϕ^t, and a multimodal interaction module ψ. Given an image representation $I \in \mathbb{R}^{(H'W') \times E}$ from ϕ^i and tabular representation $T \in \mathbb{R}^{(N+1) \times E}$ from ϕ^t where E denotes the embedding dimension and $N+1$ includes the CLS token, ψ generates a multimodal representation $F \in \mathbb{R}^{(N+1) \times E}$. TIP employs three SSL losses: contrastive learning between I and T, image-tabular matching and tabular data reconstruction from F. For further implementation details refer to [4]. Following SSL, we utilize the CLS token during fine-tuning to derive PLR and PRS.

Personalized Linear Regression (PLR). In standard transformer architectures, predictions are typically generated by applying a linear layer to the CLS token, mapping it to the number of output classes (1 in the case of regression). Here, we instead transform the CLS token using a linear layer into a vector of size $N+1$, comprising N personalized weights ($\beta_{p,i}$ for each tabular feature $x_{p,i}^t$ per sample p) and a bias term ($\beta_{p,0}$). Given these personalized regression weights, the prediction $\hat{y}_p$ is computed using the linear regression equation:

$$\hat{y_p} = \beta_{p,0} + \sum_{i=1}^{N} x_{p,i}^t \times \beta_{p,i} \tag{2}$$

Personalized RegScore (PRS). PRS follows a similar approach but incorporates binarized tabular features as an additional input. We introduce a gating mechanism that retains only the top k binary features by setting the rest to zero. This is achieved by computing the mean embeddings of each feature and applying a linear transformation (W_g) to obtain scores S. k features with the

Table 1: Results of PH classification with scoring systems of size k and mPAP estimation using RegScore. We **bold** the best and underline second-best results. † indicates p-value <0.05 for statistically significant difference from RegScore.

Method	k	MAE ↓	R ↑	Accuracy ↑	F1 ↑
Unit [1]	-	-	-	66.05 ± 21.5	72.42 ± 23.0
MISS †[10]	5	-	-	84.52 ± 0.67	90.88 ± 0.40
RiskSLIM †[23]	5	-	-	85.21 ± 0.62	91.31 ± 0.47
FasterRisk †[17]	5	-	-	85.36 ± 1.03	91.40 ± 0.63
FasterRisk †[17]	50	-	-	86.51 ± 1.16	92.04 ± 0.71
RegScore$_{BS}$	5	8.53 ± 0.15	63.39 ± 1.44	86.59 ± 0.27	92.43 ± 0.15
RegScore$_{BS}$	50	7.75 ± 0.13	69.90 ± 1.02	88.05 ± 0.57	93.24 ± 0.31
RegScore$_{OKR}$	5	8.69 ± 0.30	61.73 ± 2.03	86.74 ± 0.79	92.54 ± 0.46
RegScore$_{OKR}$	50	$\mathbf{7.73 \pm 0.13}$	$\mathbf{70.06 \pm 0.96}$	$\mathbf{88.12 \pm 0.72}$	$\mathbf{93.28 \pm 0.38}$

highest scores pass through the gating mechanism. During training, we use a soft gating function K_s (a sigmoid function with steepness controlled by $\frac{1}{\tau}$), while during inference, we apply a hard gating function K_h:

$$S = W_g \left(\frac{1}{E} \sum_{i=1}^{E} F_i \right), \tau_k = \mathrm{topk}(S)_{\min}, K_{\mathrm{h}} = \mathbb{1}(S \geq \tau_k), K_{\mathrm{s}} = \sigma \left(\frac{S - \tau_k}{\tau} \right) \quad (3)$$

3 Experiments and results

In what follows, we describe the dataset used for mPAP estimation and PH classification. We compare the performance of RegScore against other methods for constructing classification scoring systems. Additionally, we benchmark PLR and PRS against various tabular and/or image-based approaches.

Dataset. This study was approved by the Ethics Committee. The dataset originates from the ASPIRE Registry (Assessing the Severity of Pulmonary Hypertension In a Pulmonary Hypertension REferral Centre) [14] and comprises 2051 invasively measured mPAP values matched with Cardiac MRI (CMR) videos of one cardiac cycle (short-axis plane). It includes data from 1918 patients (1171 females, 747 males, aged 64 ± 14 years) with some undergoing repeated procedures over time. We select demographic features and MRI-derived measurements with fewer than 500 missing values. We impute missing values using the mean for continuous and the mode for categorical features. The CMRs were acquired using devices from multiple vendors including Siemens, Philips and GE. Instead of using full videos, we extract systolic, diastolic and in-between frames as 3-channel images [11,4].

Implementation details. We split the dataset into the training set (1790 samples) used for a 5-fold cross-validation and test set (261 samples), ensuring that each patient's data appears in only one split. The splits are stratified based on mPAP (divided into four bins) to maintain similar distributions. For the scoring systems' classification task, cases with mPAP exceeding 25 mmHg are considered

Table 2: Results of mPAP regression and PH classification using imaging (I) and/or tabular (T) methods with SSL and supervised learning (FT). † indicates p-value <0.05 between the performance of PLR, PRS and other methods.

Method	SSL	FT	MAE ↓	R ↑	Accuracy ↑	F1 ↑
			Machine Learning methods			
DT †[19]	-	T	10.97 ± 0.29	51.43 ± 2.80	81.23 ± 1.46	88.74 ± 0.96
XGB †[2]	-	T	8.05 ± 0.07	69.27 ± 1.23	87.36 ± 0.61	92.77 ± 0.35
LR †[19]	-	T	7.76 ± 0.06	69.66 ± 0.79	87.89 ± 0.34	93.19 ± 0.18
GBR †[19]	-	T	7.67 ± 0.06	70.80 ± 0.63	87.89 ± 0.21	93.13 ± 0.11
RF †[19]	-	T	7.61 ± 0.04	70.42 ± 0.68	87.13 ± 0.88	92.69 ± 0.53
			Deep Learning methods			
ResNet-50 †[12]	-	I	8.24 ± 0.28	68.25 ± 2.74	84.83 ± 1.62	91.55 ± 0.67
SimCLR †[3]	I	I	7.91 ± 0.30	72.25 ± 2.12	85.59 ± 0.75	91.76 ± 0.33
DAFT †[20]	-	IT	7.74 ± 0.30	71.24 ± 1.02	88.51 ± 1.27	93.29 ± 0.80
TabMixer$_{2D}$ †[8]	-	IT	7.69 ± 0.14	72.08 ± 0.44	88.35 ± 0.79	93.23 ± 0.41
VIME †[25]	T	T	7.53 ± 0.09	73.22 ± 0.88	$\underline{88.97 \pm 1.25}$	93.56 ± 0.73
TabAttention$_{2D}$ †[9]	-	IT	7.47 ± 0.11	72.83 ± 1.14	88.89 ± 0.72	$\underline{93.59 \pm 0.31}$
MMCL [11]	IT	I	7.45 ± 0.37	75.04 ± 1.60	87.89 ± 0.44	92.93 ± 0.25
TIP [4]	IT	IT	**6.88 ± 0.25**	**77.30 ± 1.46**	88.58 ± 0.63	93.39 ± 0.35
PLR	IT	IT	$\underline{7.14 \pm 0.14}$	$\underline{75.07 \pm 0.87}$	88.66 ± 1.29	93.46 ± 0.73
PRS$_5$	IT	IT	7.19 ± 0.16	74.85 ± 1.26	**89.43 ± 1.00**	**93.84 ± 0.60**

positive (1678 positive vs. 373 negative cases). For other methods trained on regression task, mPAP serves as ground truth and classification is achieved by thresholding the predicted value. We use Mean Absolute Error (MAE) and Pearson's correlation coefficient (R) as regression metrics while accuracy and F1 as classification metrics. Mean and standard deviation are reported across the test set over five folds. We standardize numerical features, retaining only those with statistical significance based on f-regression [19]. We present all features as part of Fig. 2. CMRs are resampled to a pixel spacing of 0.9375mm×0.9735mm and resized to 128×128 pixels. Deep learning models are implemented in PyTorch and trained on an NVIDIA A100 80GB GPU for up to 500 SSL and fine-tuning epochs with the Adam optimizer. The best model is selected by validation performance, with SSL and fine-tuning learning rates chosen from $\{3 \times 10^{-3}, 3 \times 10^{-4}, 3 \times 10^{-5}\}$ and $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$ respectively. We set $B = 10$, $\tau = 0.1$, $k = 5$, $\lambda_2 = 10^{-8}$, and use tertiles for ξ in PRS and MDLP in RegScore.

Comparison with state-of-the-art methods. We compare the performance of RegScore against other scoring system methods, including Unit [1], MISS [10], RiskSLIM [23] and FasterRisk [17]. The results of these experiments are presented in Table 1. Both RegScore training approaches outperform competing methods on classification metrics (with statistically significant differences, paired t–test p–value <0.05) for $k = 5$ and $k = 50$, while also providing interpretable mPAP estimation. RegScore trained with OKRidge achieves slightly better performance than the BS version, however, the difference is not statistically significant. We present examples of scoring systems in Table 4. We also compare

Table 3: Ablation study of the key components in proposed methods.

Method	MAE ↓	R ↑	Acc. ↑	F1 ↑
RegScore$_{BS5}$	**8.53 ± 0.15**	**63.39 ± 1.44**	**86.59 ± 0.27**	**92.43 ± 0.15**
w/ tertile bins	9.38 ± 0.36	53.47 ± 3.54	86.28 ± 1.81	92.28 ± 0.89
RegScore$_{OKR5}$	**8.69 ± 0.30**	**61.73 ± 2.03**	**86.74 ± 0.79**	**92.54 ± 0.46**
w/ tertile bins	9.41 ± 0.27	52.50 ± 3.36	86.21 ± 1.51	92.24 ± 0.81
PLR	**7.14 ± 0.14**	**75.07 ± 0.87**	**88.66 ± 1.29**	**93.46 ± 0.73**
w/o Image	8.78 ± 0.91	62.28 ± 9.13	86.59 ± 2.12	92.41 ± 1.13
w/o SSL	7.79 ± 0.29	70.00 ± 1.56	88.28 ± 0.58	93.32 ± 0.30
PRS$_5$	**7.19 ± 0.16**	**74.85 ± 1.26**	**89.43 ± 1.0**	**93.84 ± 0.6**
w/ MDLP bins	7.19 ± 0.33	74.35 ± 2.29	88.74 ± 0.34	93.50 ± 0.21
w/o SSL	7.93 ± 0.22	70.10 ± 3.00	88.97 ± 0.32	93.58 ± 0.17
w/o Image	7.60 ± 0.13	71.53 ± 1.00	88.97 ± 0.50	93.60 ± 0.24

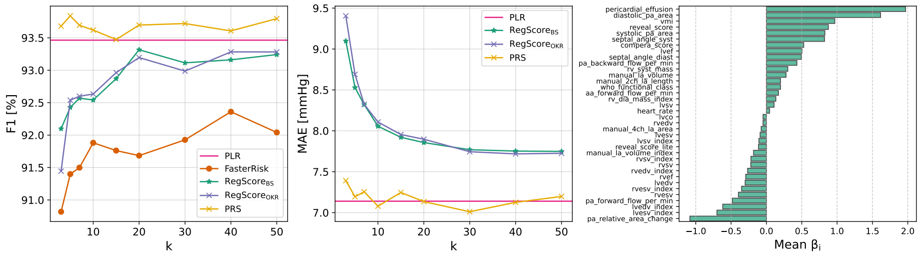

Fig. 2: Performance comparison of classification (left) and regression (center) across model sizes (k). The right panel shows the mean feature weights in PLR.

PLR and PRS against machine learning models trained on tabular data, including LR, XGBoost [2], Gradient Boosting Decision Trees (GBDT), and Random Forest (RF). Additionally, we benchmark them against deep learning methods for imaging and/or tabular data, including ResNet-50 [12], SimCLR [3], DAFT [20], TabMixer [8], VIME [25], TabAttention [9], MMCL [11], and TIP [4] (Table 2). Although PLR and PRS yield slightly higher MAE values, 7.14 and 7.19, respectively, compared to TIP (6.88), they outperform all other methods in regression performance, with all but one difference being statistically significant. Notably, PRS achieves the highest classification metrics (F1 = 93.84) among the evaluated approaches, while also offering interpretability.

Ablation study. We conduct an ablation study (Table 3) to assess key aspects of our methods. The performance of both PLR and PRS worsens when trained without image data or SSL, highlighting the importance of the training procedure and bimodality. In all methods, modifying the discretization function leads to a decline in performance, underscoring the need to carefully select an appropriate binning strategy for the algorithm.

Table 4: Examples of RegScore for mPAP estimation and other scoring systems for PH classification.

a. **RegScore**

feature	mPAP
$bias$	40.46
$1.0 \leq reveal\_score < 6.0$	-11.23
$0.73 \leq vmi \leq 2.87$	8.14
$49.17 \leq rvef \leq 82.0$	-5.90
$311.46 \leq systolic\_pa\_area < 672.0$	-8.52
$166.0 \leq septal\_angle\_syst$	10.71
mPAP [mmHg]:	$= \sum$

b. FasterRisk

feature	points
$bias$	4
$1.0 \leq reveal\_score < 6.0$	-3
$72.9 \leq rv\_syst\_mass \leq 259.04$	3
$270 \leq diastolic\_pa\_area < 545$	-3
$166 \leq septal\_angle\_syst$	4
$pericardial\_effusion = No$	-2
risk PH: $1/(1 + exp(-score))$	

c. RiskSLIM

feature	points
$bias$	3
$1.0 \leq reveal\_score < 6.0$	-2
$rv\_dia\_mass\_index < 12.56$	-1
$270 \leq diastolic\_pa\_area < 545$	-2
$166 \leq septal\_angle\_syst$	2
$pericardial\_effusion = No$	-1
risk PH: $1/(1 + exp(-score))$	

d. MISS

feature	No PH	PH
$bias$	-5	-4
$1.0 \leq reveal\_score < 6.0$	1	-1
$12.56 \leq rv\_dia\_mass\_index$	-4	-3
$270 \leq diastolic\_pa\_area < 545$	0	-2
$166 \leq septal\_angle\_syst$	-3	-1
$pericardial\_effusion = UNK$	-1	0
score:	$= \sum$	$= \sum$

4 Discussion and Conclusions

In this paper, we introduced scoring systems for regression tasks. For diseases like PH which are diagnosed by thresholding specific measurements, RegScore offers higher clinical interpretability. Unlike traditional scoring systems, RegScore not only provides an interpretable prediction but also directly relates it to the measure of interest. RegScore outperformed other scoring systems on the classification task (p–value <0.05). RegScore is efficient to calculate - it can be generated in minutes (≈55 seconds for BS variant), whereas deep learning models require hours for training. This speed enables the exploration of multiple near-optimal models to select the best one by domain experts. Furthermore, PLR and PRS introduce almost no computational overhead compared to TIP, as they only require a linear layer with additional parameters (N or D outputs instead of 1).

We examined the impact of model size k on the performance of RegScore and PRS, with results presented in Fig. 2. Across all model sizes, RegScore achieves better classification results than FasterRisk, with larger models yielding improved classification and regression performance. PRS results remain stable due to the model's personalized nature, which adapts to the number of selected features. Because PLR directly couples tabular features to predictions, we can analyze feature importance. In Fig. 2, we present mean weight values for each feature in PLR. High coefficients are assigned to features also selected by RegScore (e.g. systolic septal angle, reveal score), aligning with findings from other studies [18].

Our methods have limitations. Similar to other scoring systems, our approach may underperform on datasets characterized by highly non-linear feature interactions. What is more, there is a trade-off between interpretability and performance. Shifting from the most effective black-box TIP toward RegScore increases interpretability but reduces regression performance. This trade-off arises because PLR, PRS, and RegScore constrain their predictions by coupling them with tabular data. However, this flexibility allows clinicians to choose between more interpretable or higher-performing models based on their needs. Future work could address this trade-off by incorporating binning into the optimization process to enhance performance while maintaining interpretability.

In summary, we introduced RegScore, a novel approach for interpretable scoring in regression tasks, along with PLR and PRS, which enhance interpretability in bimodal models. Our results show that RegScore outperforms existing scoring systems in PH classification and holds promise for broader clinical applications.

Acknowledgments. This research has been made possible by the Kosciuszko Foundation. The American Centre of Polish Culture. This research was funded in whole or in part by National Science Centre, Poland 2023/49/N/ST6/01841. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. This work is supported by the EU's Horizon 2020 programme (grant no. 857533, Sano) and the Foundation for Polish Science's International Research Agendas programme (MAB PLUS/2019/13), co-financed by the EU under the European Regional Development Fund and the Polish Ministry of Science and Higher Education (contract no. MEiN/2023/DIR/3796).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Burgess, E.W.: Factors determining success or failure on parole. The workings of the indeterminate sentence law and the parole system in Illinois pp. 221–234 (1928)
2. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Du, S., Zheng, S., Wang, Y., Bai, W., O'Regan, D.P., Qin, C.: TIP: Tabular-image pre-training for multimodal classification with incomplete data. In: 18th European Conference on Computer Vision (ECCV 2024) (2024)
5. Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 242–252. Springer (2020)
6. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Ijcai. vol. 93, pp. 1022–1029. Citeseer (1993)
7. Gage, B.F., Waterman, A.D., Shannon, W., Boechler, M., Rich, M.W., Radford, M.J.: Validation of clinical classification schemes for predicting stroke: results from the national registry of atrial fibrillation. Jama **285**(22), 2864–2870 (2001)

8. Grzeszczyk, M.K., Korzeniowski, P., Alabed, S., Swift, A.J., Trzciński, T., Sitek, A.: Tabmixer: Noninvasive estimation of the mean pulmonary artery pressure via imaging and tabular data mixing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 670–680. Springer (2024)
9. Grzeszczyk, M.K., Płotka, S., Rebizant, B., Kosińska-Kaczyńska, K., Lipa, M., Brawura-Biskupski-Samaha, R., Korzeniowski, P., Trzciński, T., Sitek, A.: Tabattention: Learning attention conditionally on tabular data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 347–357. Springer (2023)
10. Grzeszczyk, M.K., Trzciński, T., Sitek, A.: Miss: Multiclass interpretable scoring systems. In: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM). pp. 55–63. SIAM (2024)
11. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hoeper, M.M., et al.: Pulmonary hypertension. Dtsch Arztebl Int **114**, 73–84 (2017). https://doi.org/10.3238/arztebl.2017.0073
14. Hurdman, J., Condliffe, R., Elliot, C., Davies, C., Hill, C., et al.: Aspire registry: Assessing the spectrum of pulmonary hypertension identified at a referral centre. European Respiratory Journal **39**, 945–955 (4 2012). https://doi.org/10.1183/09031936.00078411
15. Jiang, J.P., Ye, H.J., Wang, L., Yang, Y., Jiang, Y., Zhan, D.C.: Tabular insights, visual impacts: Transferring expertise from tables to images. In: Forty-first International Conference on Machine Learning (2024)
16. Liu, J., Rosen, S., Zhong, C., Rudin, C.: Okridge: Scalable optimal k-sparse ridge regression. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 41076–41258. Curran Associates, Inc. (2023)
17. Liu, J., Zhong, C., Li, B., Seltzer, M., Rudin, C.: Fasterrisk: Fast and accurate interpretable risk scores. In: Proceedings of Neural Information Processing Systems (2022)
18. Lungu, A., Swift, A.J., Capener, D., Kiely, D., Hose, R., Wild, J.M.: Diagnosis of pulmonary hypertension from magnetic resonance imaging–based computational models and decision tree analysis. Pulmonary circulation **6**(2), 181–190 (2016)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
20. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3d image and tabular data via the dynamic affine feature map transform. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 688–698. Springer (2021)
21. Smith, G.B., Redfern, O.C., Pimentel, M.A., Gerry, S., Collins, G.S., Malycha, J., Prytherch, D., Schmidt, P.E., Watkinson, P.J.: The national early warning score 2 (news2). Clinical Medicine **19**(3), 260 (2019). https://doi.org/10.7861/clinmedicine.19-3-260, https://www.sciencedirect.com/science/article/pii/S1470211824011862

22. Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., Initiative, A.D.N., et al.: A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. Neuroimage **189**, 276–287 (2019)
23. Ustun, B., Rudin, C.: Learning optimized risk scores. Journal of Machine Learning Research **20**(150), 1–75 (2019)
24. Vale-Silva, L.A., Rohr, K.: Long-term cancer survival prediction using multimodal deep learning. Scientific Reports **11**(1), 13505 (2021)
25. Yoon, J., Zhang, Y., Jordon, J., van der Schaar, M.: Vime: Extending the success of self- and semi-supervised learning to tabular domain. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 11033–11043. Curran Associates, Inc. (2020)