CW-B: Class Weighted Boosting Framework for Imbalance Resilient Multi Class Cardiac Phenotyping

Sijia Li[1*], Xiaoyu Tan[2*], Chen Zhan[1], Yuanji Ma[3], Haoyu Wang[4], and Xihe Qiu[1✉]

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China
qiuxihe1993@gmail.com
[2] Tencent YouTu Lab, Shanghai 200232, China
[3] Department of Cardiology, Zhongshan Hospital, Fudan University, Shanghai Institute of Cardiovascular Diseases, Shanghai 200032, China
[4] Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai 200092, China

Abstract. Cardiac discharge phenotyping informs post-discharge treatment and follow-up, but real-world records are often incomplete and class-imbalanced, increasing the risk of missed high-risk phenotypes. We propose **CW-B**, a clinical risk-aligned class-weighted XGBoost pipeline for five-class cardiac discharge phenotyping under real-world class imbalance and missingness. CW-B combines fold-specific class-balanced instance weighting, missingness-indicator augmentation, and classwise error auditing to improve recognition of clinically prioritized phenotypes while preserving interpretable and auditable decision logic. In five-fold stratified cross-validation, CW-B achieves the best Accuracy, Macro-F1, Balanced Accuracy, and Prioritized F1 among tree-based, ensemble, and neural baselines. Overall, CW-B provides a practical and deployment-oriented approach for more reliable cardiac discharge phenotyping in real-world clinical settings.

Keywords: Class-weighted learning · XGBoost · Multiclass classification.

1 Introduction

Automated phenotyping of hospital discharge diagnoses serves as a critical bridge between in-hospital care and post-discharge management, directly informing risk stratification, therapeutic planning, and longitudinal follow-up pathways [17]. Compared with single-outcome prediction, discharge phenotyping more closely matches real clinical decision entry points and therefore requires models to produce stable and consistent assignments under substantial physiological heterogeneity, incomplete documentation, and noisy measurements [8]. Moreover, the

[*] Equal contribution. ✉ Corresponding author.

clinical consequences of errors are inherently asymmetric, as missed identification of acute high-risk phenotypes can divert subsequent management and incur disproportionately severe harm. Consequently, algorithms for discharge phenotyping should achieve strong overall discrimination while explicitly auditing and reducing high-risk missed detections, supported by interpretable and auditable evidence to facilitate clinical adoption.

Prior work in clinical phenotyping and computational phenomics spans multiple methodological paradigms [2, 1]. Rule-based and expert systems offer transparency but often depend on institution-specific coding practices and heuristic thresholds, limiting generalizability across populations and sites [12]. Conventional machine learning and ensemble models are effective on structured data [18], yet under skewed class distributions they frequently optimize global metrics in ways that underemphasize clinically salient phenotypes [20, 16]. More recently, end-to-end deep models and large language models have been explored for EHR representation learning and clinical reasoning; however, their failure modes remain difficult to characterize, and limitations in interpretability and auditability continue to impede deployment [31, 7]. Overall, there remains a need for a practical framework that explicitly models class bias and asymmetric risk while providing a controllable trade-off and reproducible evidence suitable for clinical auditing [9, 10].

To address this gap, we propose **CW-B**, a risk-aligned class-weighted boosting pipeline for cardiac discharge phenotyping. Rather than treating imbalance handling as a generic preprocessing step, CW-B integrates fold-specific class-balanced instance weighting, explicit missingness modeling, and clinically prioritized error auditing within a unified five-class prediction task. In addition, CW-B employs robust preprocessing and explicit modeling of missingness to improve resilience to incomplete records and measurement variability commonly observed in real-world data. While retaining the computational efficiency and deployment practicality of tree-based models, CW-B preserves interpretability through feature-based split structures, supporting traceable decision rationales and systematic auditing [4].

Our contributions are threefold:

1. We formulate cardiac discharge phenotyping as a risk-aligned five-class task with explicit emphasis on clinically actionable missed detections.
2. We implement CW-B with fold-specific class-balanced instance weighting and leakage-free missingness-aware feature construction.
3. We benchmark CW-B against tree-based, ensemble, and neural baselines using global metrics and prioritized classwise error auditing.

2 Methods

2.1 Task Definition and Notation

We study discharge phenotyping as a multi class classification problem over structured clinical records. An overview of the overall pipeline and the proposed

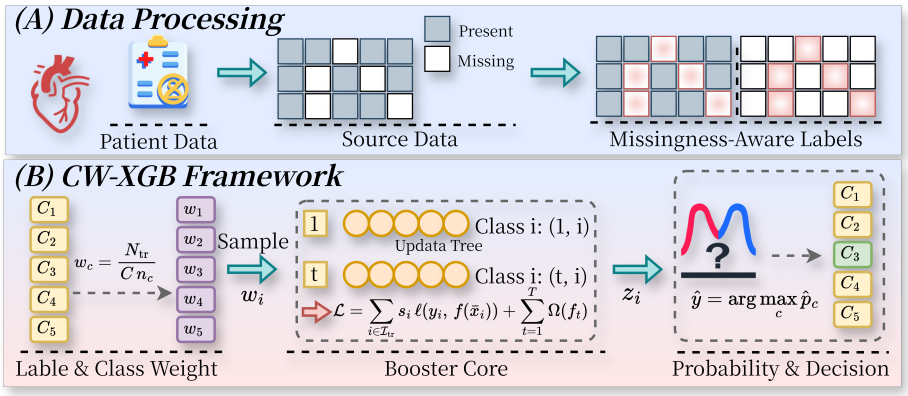

Fig. 1. Overview of the proposed CW-B pipeline. (A) Data processing and construction of missingness-aware labels, where binary missingness indicators encode whether each clinical variable is originally observed or missing. (B) The CW-XGB classifier core, which performs class-weighted training and probability-based multiclass prediction.

CW-B framework is shown in Fig. 1. Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^d, \qquad y_i \in \{0, \ldots, C-1\}, \tag{1}$$

where x_i is the feature vector and y_i is the discharge phenotype label. In our main setting, $C = 5$ with the label mapping

$$0 = \text{stableCAD}, \quad 1 = \text{ACS}, \quad 2 = \text{oldMI}, \quad 3 = \text{CAS}, \quad 4 = \text{nonCAD}. \tag{2}$$

Here, **stableCAD** (label 0) refers to stable coronary artery disease with chronic or stable ischemic presentation; **ACS** (label 1) denotes acute coronary syndrome characterized by time-sensitive acute ischemic events; **oldMI** (label 2) indicates a prior myocardial infarction documented as historical disease; **CAS** (label 3) denotes non-obstructive coronary artery disease and, in our cohort, corresponds to clinically suspected CAD patients from the younger, initially non-CAD subgroup who underwent coronary angiography demonstrating $< 50\%$ luminal narrowing. Disease burden was further quantified using the Gensini score (GS), and patients were followed longitudinally to monitor incident CAD-related events [28, 30]; and **nonCAD** (label 4) represents patients without a CAD-related discharge phenotype. In clinical practice, **0/1/3** (stableCAD, ACS, and CAS) are associated with established diagnostic and therapeutic pathways, making them clinically actionable categories for downstream management, which motivates our emphasis on these classes in deployment-oriented interpretation while retaining a unified five-class learning objective.

The model outputs a probability vector $p_i = f(x_i) \in [0,1]^C$ and a prediction $\hat{y}_i = \arg\max_c p_{ic}$. Before model evaluation, we define the clinically prioritized set

$$\mathcal{C}_p = \{0, 1, 3\}, \tag{3}$$

Algorithm 1 Class weighted gradient boosted trees for one stratified fold

Require: Indices $\mathcal{I}_{\text{tr}}, \mathcal{I}_{\text{ev}}$, features $X \in \mathbb{R}^{N \times d}$, labels $y \in \{0, \ldots, C-1\}^N$, classes C, hyperparameters Θ, missing sentinel s_{miss}
Ensure: Fold model f_k, preprocessing parameters Π_k, predictions $(\hat{y}, \hat{p})$
$r \leftarrow \mathbb{I}(X = s_{\text{miss}})$
Set $X_{ij} \leftarrow \text{NaN}$ wherever $r_{ij} = 1$
Compute (μ, σ, m) on $X[\mathcal{I}_{\text{tr}}]$ and set $\Pi_k = (\mu, \sigma, m)$
Standardize using (μ, σ) and impute missing values using m
Form $\bar{X}_{\text{tr}} = [\tilde{X}_{\text{tr}}; r[\mathcal{I}_{\text{tr}}]]$ and $\bar{X}_{\text{ev}} = [\tilde{X}_{\text{ev}}; r[\mathcal{I}_{\text{ev}}]]$
$n_c \leftarrow \sum_{i \in \mathcal{I}_{\text{tr}}} \mathbb{I}(y_i = c),\ \forall c$
$w_c \leftarrow \frac{|\mathcal{I}_{\text{tr}}|}{C\, n_c},\ \forall c$
$\tilde{w}_c \leftarrow \frac{w_c}{\frac{1}{C}\sum_j w_j},\ \forall c$
Assign $s \leftarrow \tilde{w}_{y[\mathcal{I}_{\text{tr}}]}$
Train f_k on $(\bar{X}_{\text{tr}}, y[\mathcal{I}_{\text{tr}}])$ with instance weights s and hyperparameters Θ
Predict probabilities $\hat{p}$ on $\bar{X}_{\text{ev}}$ and set $\hat{y} = \arg\max_c \hat{p}_c$
return $f_k, \Pi_k, (\hat{y}, \hat{p})$

corresponding to stableCAD, ACS, and CAS. This set is pre-defined based on clinical actionability and missed-detection risk rather than class frequency or experimental results. Specifically, ACS is a time-sensitive acute ischemic phenotype, stableCAD informs secondary prevention, medication planning, and follow-up, and CAS in this cohort refers to clinically suspected CAD patients with less than 50% luminal narrowing who still require risk assessment and symptom management. Importantly, $\mathcal{C}_p$ is used only for deployment-oriented evaluation and classwise error auditing; it does not change the training labels, model objective, or prediction rule.

2.2 Preprocessing and Explicit Missingness Modeling

We treat preprocessing as an integral component of the learning pipeline and enforce a leakage free protocol. Missing entries in the raw feature table are encoded by a sentinel value s_{miss} and are converted to missing values prior to transformation. All preprocessing statistics are estimated using the training split only and then applied to the corresponding evaluation split [25, 29].

Given a training index set $\mathcal{I}_{\text{tr}}$, each feature dimension j is standardized using the training mean μ_j and standard deviation σ_j,

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}. \tag{4}$$

To ensure numerical stability, invalid or very small σ_j values are stabilized to a positive constant during fitting. Missing entries are imputed using the training median m_j,

$$\tilde{x}_{ij} = \begin{cases} m_j, & \text{if } x'_{ij} \text{ is missing,} \\ x'_{ij}, & \text{otherwise.} \end{cases} \tag{5}$$

Because missingness itself can be informative, we additionally construct a binary indicator $r_{ij} \in \{0, 1\}$ that represents whether feature j is missing for sample i [5, 13]. In our cohort, 55 of 57 features contain missing values, with 38,571 missing entries among 248,178 total feature entries, corresponding to a 15.54% missingness rate. This motivates preserving documentation patterns through explicit missingness indicators, allowing the model to distinguish an imputed numerical value from the clinical fact that the value was originally unrecorded. The final augmented representation concatenates imputed values and missingness indicators,

$$\bar{x}_i = \left[\tilde{x}_i;\, r_i\right] \in \mathbb{R}^{2d}, \tag{6}$$

allowing the model to jointly leverage observed values and recording patterns under incomplete documentation.

2.3 Class Balanced, Instance Weighted Gradient Boosted Trees

Our primary model is a class balanced, instance weighted gradient boosted decision tree classifier [6]. The objective is to mitigate label imbalance by controlling the contribution of each class during learning while preserving transparent, auditable tree structured decision rules [15].

Let n_c denote the number of training samples in class c and N_{tr} denote the number of training samples. We define class balanced weights

$$w_c = \frac{N_{\mathrm{tr}}}{C\, n_c}, \qquad \tilde{w}_c = \frac{w_c}{\frac{1}{C}\sum_{j=0}^{C-1} w_j}, \qquad s_i = \tilde{w}_{y_i}, \tag{7}$$

where s_i is the instance weight assigned to sample i. This construction reduces the dominance of frequent phenotypes and improves class balanced learning.

Let $\ell(\cdot)$ denote the multiclass negative log likelihood and let $\{f_t\}_{t=1}^{T}$ be the additive sequence of trees with regularizer $\Omega(\cdot)$. The weighted learning objective is

$$\mathcal{L} = \sum_{i \in \mathcal{I}_{\mathrm{tr}}} s_i\, \ell(y_i,\, f(\bar{x}_i)) + \sum_{t=1}^{T} \Omega(f_t). \tag{8}$$

Optimization follows gradient boosting with a second order approximation, where instance weights modulate both split selection and leaf value estimation through weighted gradients and curvatures [32].

3 Experiments

3.1 Evaluation Protocol

We evaluate the proposed method on a five-class discharge phenotyping task with $N = 4354$ samples and $d = 57$ structured features. The class counts for stableCAD, ACS, oldMI, CAS, and nonCAD are 1675, 483, 266, 461, and 1469, respectively. Stratified five-fold cross-validation is used with a fixed random seed

Table 1. Hyperparameter configuration for the proposed method and baselines.

Method	Key settings
CW-B (Ours)	600 boosting iterations, depth 5, learning rate 0.07, subsampling 0.95, ℓ_2 1.0
XGB	1200 boosting iterations, depth 7, learning rate 0.03, subsampling 0.95, ℓ_2 1.0
CB	1200 boosting iterations, depth 8, learning rate 0.03
STK	CatBoost and XGBoost base learners, logistic regression meta learner, out of fold training
Neural baselines	MLP(256) for DQN and BC, MLP(256,128) with early stopping; Adam learning rate 10^{-3}

Table 2. Mean CV performance. Prioritized metrics average stableCAD, ACS, and CAS; lower is better for miss rate.

Meth	Acc	M-F1	Bal Acc	Pri Rec	Pri F1	Pri Miss	M-F1 Std
CW-B(ours)	**0.81**	**0.72**	**0.73**	0.67	**0.69**	0.33	0.0103
XGB	0.80	0.69	0.66	0.65	0.67	0.35	0.0108
CB	0.78	0.70	0.70	0.66	0.66	0.34	0.0136
STK	0.75	0.69	0.71	**0.69**	0.66	**0.31**	0.0069
DQN	0.76	0.68	0.68	0.66	0.64	0.34	0.0070
BC	0.77	0.68	0.67	0.62	0.64	0.38	0.0109
MLP	0.76	0.65	0.63	0.58	0.60	0.42	0.0180

[19]. Preprocessing statistics are estimated on the training split only and then applied to the corresponding evaluation split, ensuring that evaluation data do not influence feature construction. All component ablations are conducted under the same five-fold stratified cross-validation protocol, random seed, and main model configuration unless otherwise specified.

We report overall classification accuracy (Accuracy), macro averaged F1 score (Macro-F1), and balanced accuracy (Balanced Accuracy), where Balanced Accuracy is defined as the average of per class recall [3, 24]. To emphasize deployment relevant failure modes, we additionally report averaged recall, averaged F1 score, and averaged miss rate over the clinically prioritized set $\mathcal{C}_p = \{0, 1, 3\}$, corresponding to stableCAD, ACS, and CAS.

3.2 Baselines and Hyperparameter Configuration

We benchmark CW-B against tree based, ensemble, and neural baselines under the same preprocessing constraints and the same evaluation protocol. XGB is an XGBoost baseline with a higher capacity configuration, using more boosting iterations and deeper trees to assess whether gains are attributable to capacity rather than the proposed weighting strategy. CB is a CatBoost baseline that

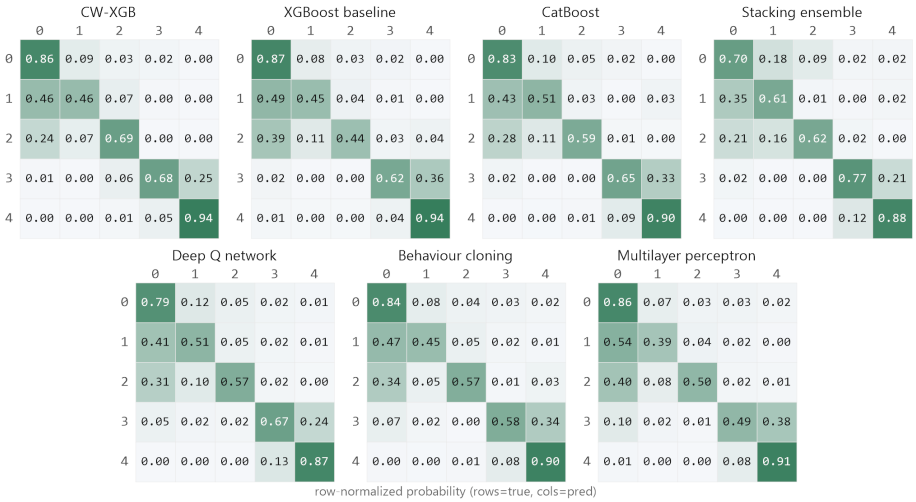

Fig. 2. Concatenated confusion matrices. Each subplot corresponds to one method. Class order: stableCAD, ACS, oldMI, CAS, nonCAD.

evaluates an alternative gradient boosting implementation under class balancing [27]. STK is a stacking ensemble that combines CatBoost and XGBoost base learners and fits a logistic regression meta learner on out of fold predictions, providing a controlled test of whether model fusion can improve clinically prioritized errors beyond single learners [23]. DQN and BC are multilayer perceptron baselines trained with cost sensitive objectives based on focal loss and class balancing, serving as neural references that emphasize minority sensitivity [26, 11]. MLP is a standard multilayer perceptron trained with early stopping, serving as a non cost sensitive neural comparator [21, 22].

These baselines are selected to isolate and validate the key components of the proposed approach. Tree based baselines (XGB, CB) test whether class balanced instance weighting and the explicit missingness representation yield consistent gains under strong tabular inductive bias [14]. The higher capacity XGBoost configuration tests whether the proposed benefits persist beyond capacity scaling. The stacking ensemble tests whether complementary decision boundaries can further reduce missed detections in the prioritized phenotypes. Neural baselines test whether cost sensitive training alone is sufficient to match boosted trees on this structured, moderately sized cohort, thereby supporting the choice of a transparent tree based model for deployment. Table 1 summarizes the core hyperparameter configuration and imbalance handling strategy.

3.3 Metrics and Main Results

Let $M \in \mathbb{N}^{C\times C}$ denote the confusion matrix, where rows represent true labels and columns represent predicted labels. We report Accuracy, Macro-F1, and

Balanced Accuracy using their standard definitions, with Balanced Accuracy computed as the mean class-wise recall. For class c, let R_c denote recall and $\mathrm{F1}_c$ denote the class-wise F1 score. Specifically,

$$R_c = \frac{M_{cc}}{\sum_{b=0}^{C-1} M_{cb}}, \qquad \mathrm{MissRate}_c = 1 - R_c. \tag{9}$$

For the clinically prioritized set $\mathcal{C}_p = \{0, 1, 3\}$, corresponding to stableCAD, ACS, and CAS, we further report

$$\begin{aligned} \text{Prioritized Recall} &= \frac{1}{|\mathcal{C}_p|} \sum_{c \in \mathcal{C}_p} R_c, \\ \text{Prioritized F1} &= \frac{1}{|\mathcal{C}_p|} \sum_{c \in \mathcal{C}_p} \mathrm{F1}_c, \\ \text{Prioritized MissRate} &= 1 - \text{Prioritized Recall}. \end{aligned} \tag{10}$$

Figure 2 visualizes the concatenated confusion matrices, enabling inspection of systematic confusions involving stableCAD, ACS, and CAS. Table 2 reports mean performance for all methods, sorted by Macro-F1, then by the prioritized F1, and then by Accuracy.

3.4 Discussion

CW-B achieves the best Accuracy, Macro-F1, Balanced Accuracy, and Prioritized F1 among the compared methods, while remaining competitive on prioritized recall and miss rate. As shown in Table 2, CW-B improves over the higher-capacity XGBoost baseline, suggesting that the gain is not simply attributable to model capacity. Additional component analyses under the same five-fold stratified cross-validation protocol further support the contribution of the main design choices. Removing class-balanced instance weighting reduces Macro-F1, Balanced Accuracy, and Prioritized F1 to 0.6968, 0.6621, and 0.6631, respectively. Removing the missingness indicator yields 0.6804, 0.6666, and 0.6444, while relying only on XGBoost native missing-value handling yields 0.7082, 0.6909, and 0.6780. Alternative imputation with indicators also underperforms the full pipeline; for example, KNN imputation yields 0.7008, 0.6791, and 0.6703. These results indicate that both class-balanced learning and explicit missingness modeling contribute to macro-level and prioritized-class performance. Fig. 2 provides complementary qualitative evidence by visualizing classwise error patterns. The stacking ensemble achieves the lowest prioritized miss rate, but this sensitivity-oriented operating point comes with a clear trade-off in overall accuracy and macro-level balance. In contrast, CW-B preserves strong global performance while delivering the best prioritized F1 for the physician-emphasized phenotypes.

4 Conclusion

We propose CW-B, a class weighted gradient boosted tree framework with explicit missingness modeling for imbalance resilient multi class discharge pheno-

typing. In stratified cross validation, CW-B achieves the strongest macro level performance with competitive overall accuracy and balanced accuracy, and it yields more clinically coherent confusion patterns with fewer clinically undesirable errors. Clinically, these improvements are meaningful because discharge phenotypes directly guide secondary prevention, medication selection, follow up intensity, and referral pathways, and more reliable phenotype assignments can reduce downstream management variability and support earlier, more consistent risk aligned care after hospitalization. In addition, the transparent tree based structure facilitates post hoc auditing and error review, which is essential for building clinical trust and enabling safe integration into real world discharge workflows.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alzoubi, H., Alzubi, R., Ramzan, N., West, D., Al-Hadhrami, T., Alazab, M.: A review of automatic phenotyping approaches using electronic health records. Electronics **8**(11), 1235 (2019)
2. Banda, J.M., Seneviratne, M., Hernandez-Boussard, T., Shah, N.H.: Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annual review of biomedical data science **1**(1), 53–68 (2018)
3. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. pp. 3121–3124. IEEE (2010)
4. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1721–1730 (2015)
5. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific reports **8**(1), 6085 (2018)
6. 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 (2016)
7. De Freitas, J.K., Johnson, K.W., Golden, E., Nadkarni, G.N., Dudley, J.T., Bottinger, E.P., Glicksberg, B.S., Miotto, R.: Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns **2**(9) (2021)
8. Denaxas, S.C., Gonzalez-Izquierdo, A., Fitzpatrick, N.K., Direk, K., Hemingway, H.: Phenotyping uk electronic health records from 15 million individuals for precision medicine: the caliber resource. In: ICIMTH. pp. 220–223 (2019)
9. Ding, S., Zhang, S., Hu, X., Zou, N.: Identify and mitigate bias in electronic phenotyping: A comprehensive study from computational perspective. Journal of Biomedical Informatics **156**, 104671 (2024)

10. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
11. Foster, D.J., Block, A., Misra, D.: Is behavior cloning all you need? understanding horizon in imitation learning. Advances in Neural Information Processing Systems **37**, 120602–120666 (2024)
12. Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote Jr, J., Moseley, E., Grant, D.W., Tyler, P.D., et al.: A comparison of rule-based and deep learning models for patient phenotyping. preprint, arxiv. org (2017)
13. Gillies, C.E., Taylor, D.F., Cummings, B.C., Ansari, S., Islim, F., Kronick, S.L., Medlin Jr, R.P., Ward, K.R.: Demonstrating the consequences of learning missingness patterns in early warning systems for preventative health care: a novel simulation and solution. Journal of Biomedical Informatics **110**, 103528 (2020)
14. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems **35**, 507–520 (2022)
15. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications **73**, 220–239 (2017)
16. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)
17. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association **20**(1), 117–121 (2013)
18. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent data analysis **6**(5), 429–449 (2002)
19. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14, pp. 1137–1145. Montreal, Canada (1995)
20. Lewis, A.E., Weiskopf, N., Abrams, Z.B., Foraker, R., Lai, A.M., Payne, P.R., Gupta, A.: Electronic health record data quality assessment and tools: a systematic review. Journal of the American Medical Informatics Association **30**(10), 1730–1740 (2023)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
22. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)
23. Naimi, A.I., Balzer, L.B.: Stacked generalization: an introduction to super learning. European journal of epidemiology **33**(5), 459–464 (2018)
24. Opitz, J.: From bias and prevalence to macro f1, kappa, and mcc: A structured overview of metrics for multi-class evaluation. Heidelberg University (2022)
25. Perez-Lebel, A., Varoquaux, G., Le Morvan, M., Josse, J., Poline, J.B.: Benchmarking missing-values approaches for predictive models on health databases. GigaScience **11**, giac013 (2022)
26. Pomerleau, D.A.: Alvinn: An autonomous land vehicle in a neural network. Advances in neural information processing systems **1** (1988)
27. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31** (2018)

28. Reynolds, H.R., Diaz, A., Cyr, D.D., Shaw, L.J., Mancini, G.J., Leipsic, J., Budoff, M.J., Min, J.K., Hague, C.J., Berman, D.S., et al.: Ischemia with nonobstructive coronary arteries: insights from the ischemia trial. Cardiovascular Imaging **16**(1), 63–74 (2023)
29. Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj **338** (2009)
30. Widmer, R.J., Samuels, B., Samady, H., Price, M.J., Jeremias, A., Anderson, R.D., Jaffer, F.A., Escaned, J., Davies, J., Prasad, M., et al.: The functional assessment of patients with non-obstructive coronary artery disease: expert review from an international microcirculation working group. EuroIntervention **14**(16), 1694–1702 (2019)
31. Yang, S., Varghese, P., Stephenson, E., Tu, K., Gronsbell, J.: Machine learning approaches for electronic health records phenotyping: a methodical review. Journal of the American Medical Informatics Association **30**(2), 367–381 (2023)
32. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: Third IEEE international conference on data mining. pp. 435–442. IEEE (2003)