Clinically interpretable risk stratification in Chagas disease using novel derived variables and calibrated decision trees

Henrique C. Parro[1], Alexandra Hurtado Ortíz[2], Alan Demétrius Baria Valejo[3], Liliana Torcoroma García[4], Lina Garcés[1]

[1] Universidade de São Paulo – USP
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brazil

[2]Hospital Internacional de Colombia – HIC
Fundación Cardiovascular de Colombia FCV
Fundación Universitaria FCV
Cl. 155A No. 23 - 58, Floridablanca, Santander, Colombia

[3]Universidade Federal de São Carlos – UFSCar
13565-905 – São Carlos – SP, Brazil

[4]Universidad de Santander – UDES
Cl. 70 No. 55-210, Bucaramanga, Santander, Colombia

henrique_parro@usp.br, alexandrajhop@gmail.com, alanvalejo@ufscar.br

l.torcoroma@udes.edu.co, linagarces@usp.br

Abstract. *This work investigates machine learning for risk stratification in Chagas cardiomyopathy. We evaluate a novel set of 10 derived clinical variables and propose an interpretable decision-support framework based on calibrated decision trees, threshold optimization, and abstention mechanisms. Three clinical representations (4 established variables, 10 derived variables, and 97 clinical variables) are compared under binary and ternary prediction settings. Results show that the comprehensive dataset achieves the best predictive performance, while the derived variables provide a compact alternative, outperforming the 4-variable panel only in the ternary setting. Furthermore, the proposed framework improves the sensitivity–specificity trade-off while preserving coverage, supporting safer clinical triage.*

1. Introduction

Chagas disease, caused by *Trypanosoma cruzi*, remains a major public health problem, especially in Latin America. Transmission occurs mainly through the triatomine vector (commonly known as the "kissing bug"), but also through non-vector routes such as congenital transmission, blood transfusions and organ transplants, laboratory accidents, and oral outbreaks associated with contaminated food. Historically endemic in rural areas, the disease has followed migration flows and urbanization, becoming a challenge in urban centers and in non-endemic countries as well. In developing countries, the burden is aggravated by socioeconomic inequalities and barriers to diagnosis and specialized care, leading to underdiagnosis and delayed treatment [World Health Organization 2025]. In the chronic phase, many individuals remain asymptomatic for years, while a fraction progresses to Chagas cardiomyopathy with heart failure, arrhythmias, and risk of sudden death [Rassi et al. 2010].

In this context, machine learning (ML) techniques can support clinical decision-making in low-resource settings by assisting triage and risk stratification, prioritizing patients with a higher probability of adverse outcomes. For models to be useful in practice, they must produce well-calibrated probabilities and rely on explicitly defined decision points, improving transparency, reproducibility, and interpretability, which are critical in clinical applications [Guo et al. 2017].

Although several studies use large collections of clinical variables, it remains unclear whether compact derived representations can preserve predictive information while improving interpretability and reducing data acquisition costs for cardiovascular risk prediction. A second challenge concerns translating predictions into clinical decisions. Most machine learning studies in Chagas disease focus on predictive performance metrics. At the same time, calibrated probabilities, explicit decision thresholds, and abstention mechanisms remain underexplored despite their potential to enable safer clinical triage [Guo et al. 2017, Kompa et al. 2021, Kelly et al. 2019].

To address these gaps, first, we evaluate a novel set of 10 derived clinical variables. While these variables were designed as a compact representation of patient status, their usefulness for predicting cardiovascular complications in Chagas disease has not yet been systematically investigated. Second, we propose an interpretable decision-support framework based on calibrated decision trees, probability calibration, threshold optimization, and abstention mechanisms. Unlike conventional classification pipelines that rely solely on predicted labels, the proposed framework explicitly models decision confidence and supports safer triage by deferring low-confidence cases for human review.

Specifically, we investigate whether: (i) the proposed derived variables effectively condense clinically relevant information; (ii) compact and comprehensive clinical representations provide different trade-offs between simplicity, interpretability, and predictive performance; (iii) calibrated decision trees produce reliable probability estimates; and (iv) threshold-based abstention policies improve the safety of automated clinical triage.

To answer these questions, we compare three clinical representations: 4 established variables, the novel proposed set of 10 derived variables, and a comprehensive dataset containing 97 variables. Evaluations are conducted under both binary (“healthy” versus “complication”) and ternary (“healthy”, “complication”, and “death”) prediction settings. The main contributions of this work are:

- A first systematic evaluation of ten derived clinical variables for cardiovascular risk prediction in patients with Chagas disease.
- An interpretable decision-support framework based on calibrated decision trees, explicit thresholds, and abstention mechanisms for safer clinical triage.
- A comparative analysis of compact and comprehensive clinical representations, investigating the trade-offs between predictive performance, interpretability, and data requirements.

The 10-variable representation offered a relevant compromise between compactness and predictive performance, reducing the number of variables by nearly 90%. In the ternary setting, it improved macro-F1 from 0.377 to 0.489 over the 4-variable panel, indicating that it preserves useful information for distinguishing clinically heterogeneous outcomes.

2. Related Work

The application of ML to Chagas disease has expanded in recent years, ranging from the prediction of clinical outcomes to population screening and support for etiological decision-making[1]. Overall, studies differ in the type of data they use (clinical variables, biomarkers, structured electrocardiogram (ECG) features, or raw signals), in model complexity (trees, *ensembles*, maximum-margin methods, and deep learning), and in validation rigor. This section summarizes recent contributions, highlighting methodological convergences and recurring gaps.

In outcome prediction, Ferreira et al. [Ferreira 2022] explored multiple families of variables (clinical interview, ECG, and biomarkers), comparing supervised algorithms under stratified cross-validation and class-imbalance mitigation strategies. Their study highlights the discriminative role of biomarkers such as NT-proBNP and shows how preprocessing and feature-selection choices can substantially affect performance. Similarly, Weber et al. [Weber de Santana Teles 2021] built curated pipelines with feature selection and overfitting control for identifying Chagas cardiomyopathy, reporting competitive performance for tree-based models and *ensembles*. Both works emphasize the importance of experimental design and governance of the analytical workflow.

In low-cost screening settings, Ghilardi et al. [Ghilardi 2024] showed that demographic and exposure variables, when properly encoded and imputed, can support useful models, especially with *boosting* methods and resampling techniques such as SMOTE. The study also highlights interpretability via SHAP, improving decision transparency, an aspect not always made explicit in earlier work.

Deep learning applied to ECG signals represents another relevant line of research. Oliveira-da-Silva et al. [Oliveira 2021] applied an *AI-ECG* model to detect systolic dysfunction in the SaMi-Trop cohort, achieving high discriminative power, although without external validation in the reported study. Building on this, Jidling et al. [Jidling 2023] used large cohorts (CODE and SaMi-Trop) with external validation on REDS-II and ELSA-Brasil, explicitly addressing generalization and domain shift challenges. These findings highlight both the scalability potential of automated ECG-based screening and the need for robust evaluation across populations.

Finally, Abreu et al. [Abreu 2021] investigated stroke etiology in patients with Chagas disease, demonstrating the feasibility of decision support in a multicenter retrospective setting. The study also underscores the importance of inter-institutional standardization and consistent validation to enable methodological comparability.

Against this background, the main differentiator of this work is a consolidated pipeline centered on the interpretability of a calibrated decision tree, with explicit threshold selection guided by operational targets, and a triage scheme with abstention that reports conditional PPV/NPV and coverage. In contrast to much of the related literature, which prioritizes *ensembles* or deep learning and does not always make the operating point explicit, our approach makes the decision policy clear (calibration outside the training fold, selection of positive and negative decision thresholds, and *trade-offs* between precision and sensitivity) and keeps the leaf-level decision rules traceable.

[1]Etiology refers to the scientific study of the causes, origins, or determining factors of diseases, medical conditions, or phenomena.

In addition, we adopt stratified cross-validation, class-imbalance control, and permutation feature-importance analysis (macro and per class), preserving transparency and reproducibility without sacrificing clinical utility. In summary, while related work reinforces the potential of ML for Chagas, this study contributes by operationalizing safer and interpretable decisions through calibrated trees and abstention-based triage, offering a pragmatic path toward deployment in low-resource settings.

Accordingly, this study complements previous Chagas-related ML research by jointly examining representation size, probability reliability, and decision coverage. Rather than treating interpretability solely as feature attribution, we use traceable tree rules and explicit operating thresholds to determine when the model should predict and when it should defer the case for clinical review.

3. Materials and Methods

For developing the ML models in this study, we adopted CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [Chapman 2000, Domingues 2020], a widely used process in data science that provides a methodological framework to organize the study into five iterative phases, as described as follows.

3.1. Business Understanding

This phase consists of identifying and scoping the business problem to be addressed, mapping stakeholders, clarifying the decision the model should support, recording operational and regulatory constraints, and formulating verifiable objectives. In this project, business understanding was conducted with support from a medical team specialized in infectious diseases. The problem was defined as the limited clarity around the clinical progression of patients with Chagas disease. The main stakeholders are clinicians and cardiologists, for whom the model may support care and the diagnosis of Chagas cardiomyopathy, as well as patients, who may benefit from improved diagnosis and follow-up.

3.2. Data Understanding

Data understanding begins with mapping the available sources and documenting how data are produced and managed, including ownership, update frequency, granularity, linkage keys, and the domains and units of each variable. This also involves recording data lineage, describing how information flows from the source through transformations to the analytical dataset [Chapman 2000, Domingues 2020]. In parallel, we perform exploratory analyses and basic data-quality checks to assess completeness, validity, consistency, uniqueness, and timeliness, summarize distributions and skewness, identify rare categories, and inspect temporal patterns such as seasonality and delays between event occurrence and data capture [Provost and Fawcett 2013, Kuhn and Johnson 2013]. These steps guide subsequent decisions on which variables to prioritize and which preprocessing procedures are required [Chapman 2000].

In this study, the anonymized data were extracted from the electronic health record system of a specialized cardiovascular clinic. Patients positive for *Trypanosoma cruzi* were labeled into three mutually exclusive outcomes: healthy, complication without death, and death; patients who later died were assigned to the death class.

3.3. Data Preparation

Data preparation was supervised and reviewed by medical specialists. This phase turns raw sources into a reproducible analytical dataset ready for modeling [Chapman 2000]. It involves selecting relevant variables and records, standardizing formats and units, assigning appropriate data types, and encoding categorical variables; when suitable for the learning algorithm, normalization or standardization is applied [Kuhn and Johnson 2013]. Missing values and outliers are handled through explicit strategies, high-cardinality categories are reduced, and rare levels are consolidated, with all decisions documented to preserve traceability between the processed data and the evaluation metrics [Provost and Fawcett 2013, Kuhn and Johnson 2013].

The complete clinical dataset contains 97 variables from 1,037 patients diagnosed between 2023 and 2025. Preparation followed a common procedure across experiments. Continuous variables were discretized into clinically meaningful ranges when this improved interpretability, and normalized when needed. Missing values were imputed using the median for numeric variables and the mode for categorical variables, with explicit missingness indicators added when clinically relevant. To mitigate information leakage, we identified fields that could reveal the outcome directly, particularly date-related variables and records occurring after the event, and removed them to retain only information plausibly available at the time of clinical decision-making. Finally, categorical variables were encoded via one-hot encoding with consolidation of rare levels, and the resulting dataset was stratified by class for validation.

3.3.1. Preparing Three Datasets

We used three datasets derived from the complete clinical database. The first is a compact set of four variables that are well established in the literature for this cohort: age, sex, left ventricular ejection fraction (LVEF), and N-terminal pro-B-type natriuretic peptide (NT-proBNP). The second is an intermediate set of ten variables (age, sex, LVEF, NT-proBNP, NYHA functional class, heart rate, cardiovascular-risk comorbidities, creatinine, blood urea nitrogen (BUN), and TAPSE), selected by the medical team in prior work based on a linear regression applied to the original dataset. Finally, we considered the full clinical dataset, with 97 variables for 1,037 patients. Because the 4- and 10-variable sets are subsets derived from the full database, all three datasets share the same sample of 1,037 patients. The data used in this study are available in a public repository [Parro et al. 2026]. Figure 1 shows the class distribution for this dataset.

3.4. Modeling

After the dataset validation by the specialists, we proceeded to model development. In the modeling phase, we translated the analytical formulation into a reproducible experimental protocol by defining the model space, the transformation pipeline, and a validation scheme designed to prevent information leakage, as recommended in [Chapman 2000, Provost and Fawcett 2013].

We used a decision tree classifier with probabilistic outputs calibrated via sigmoid calibration using `CalibratedClassifierCV` with three-fold internal cross-validation, aiming to improve the reliability of predicted probabilities

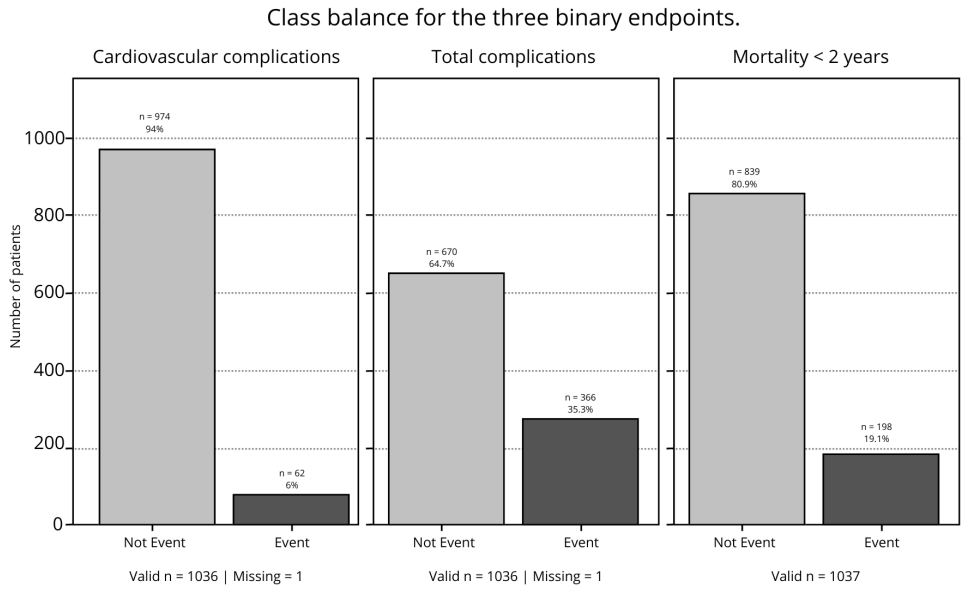


Figure 1. Class distribution.

[Niculescu-Mizil and Caruana 2005]. The code was implemented in *Python* 3.14; the main libraries were *scikit-learn*[2] and *imblearn*[3], which provide, respectively, a mature and reproducible framework for building end-to-end ML pipelines and a standardized set of methods for handling class imbalance. All experiments were run on Google Colab, ensuring a consistent environment for dependency management and reproducible execution of the notebooks.

Hyperparameters were tuned using grid search with five-fold stratified cross-validation and a fixed random seed (42). We evaluated combinations of impurity criterion, maximum depth, minimum split and leaf sizes, cost-complexity pruning, and imbalance-handling strategies, including class weighting and resampling methods (*oversampling*, *undersampling*, and hybrid approaches). The *sampler* stage compared no resampling against alternatives such as *RandomOverSampler*, *SMOTE*, *SMOTE-Tomek*, *SMOTE-ENN*, and *RandomUnderSampler*. Model selection was guided by multiple metrics, using *precision macro* as the *refit* criterion while tracking *F1 macro*, *F1 weighted*, macro sensitivity, balanced accuracy, and the Matthews correlation coefficient, providing a robust assessment under class imbalance [Kuhn and Johnson 2013, Provost and Fawcett 2013]. The full experimental design, including seeds, dataset versions, and parameter settings, was documented to ensure reproducibility and traceability [Chapman 2000].

Given the large number of combinations explored in the grid, the runtime was substantial. To enable faster iterations and local comparisons, we implemented a "fast" version of the code that fixes the *best_params* found in the full search (or uses a reduced parameter grid), while keeping the same preprocessing, calibration, and validation pipeline to preserve comparability. The code for this study is publicly available in [Parro et al. 2026].

[2]`https://scikit-learn.org/stable/`
[3]`https://imbalanced-learn.org/stable/`

4. Model Evaluation Results

This section presents the experimental results. We trained six models based on the same calibrated decision-tree architecture, split across the binary and ternary classification tasks. For each task, we evaluated the three feature sets defined earlier: the reduced panel (4 variables), the intermediate set (10 variables), and the comprehensive clinical dataset. Figure 2 summarizes this experimental design.

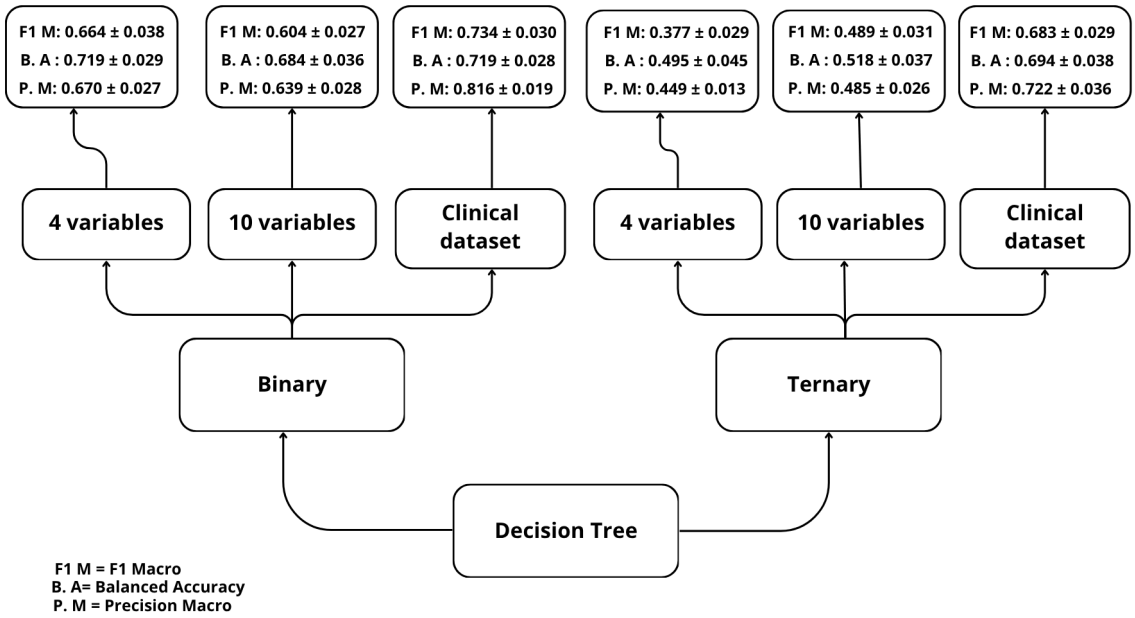


Figure 2. Organization of results.

The evaluation stage assessed whether the models met the intended triage goals under the constraints of an imbalanced clinical setting [Chapman 2000]. We used stratified cross-validation and reported both average performance and fold-to-fold variability across a set of complementary metrics, including balanced accuracy, macro-*F1*, positive-class *F1*, macro precision, macro sensitivity, ROC AUC, and negative log-loss. These metrics were chosen to reflect the operational trade-offs of screening, where errors have different clinical costs.

To define an explicit operating policy, we computed calibrated *out-of-fold* probabilities and selected class-specific decision thresholds using precision–sensitivity curves in a one-vs-rest framework [Provost and Fawcett 2013]. For comparison, we evaluated the standard *argmax* rule alongside the threshold-based regime, producing confusion matrices and macro/per-class summaries in both cases.

The reference confusion matrices intentionally use different rules: an optimized threshold in the binary task to reflect the clinical sensitivity–false-negative trade-off, and standard *argmax* in the ternary task as a direct multiclass reference; abstention is evaluated separately in both settings.

In the binary task, we defined two thresholds from the *out-of-fold* probabilities: T_{neg} for safe rule-out (healthy) and T_{pos} for detecting complications. Given a calibrated probability $p(\text{complication} \mid x)$, cases with $p(\text{complication} \mid x) \geq T_{\text{pos}}$ are labeled as complication, while cases with $p(\text{complication} \mid x) \leq 1 - T_{\text{neg}}$ are labeled as healthy; intermediate cases trigger abstention and are deferred for human review. This design allows safety and coverage to be tuned explicitly. The selected values of T_{neg}, T_{pos}, and T_c are stored as part of the experimental artifacts.

In the ternary task, we applied one-vs-rest thresholds T_c for each class in the abstention-based triage analysis. A case is automatically decided when at least one class reaches its threshold; if multiple thresholds are met, we assign the class with the highest calibrated probability. If no threshold is met, the model abstains. This setup supports asymmetric targets for minority classes and reduces the tendency of the majority class to dominate decisions.

Finally, metrics were interpreted in combination. Since plain accuracy can be misleading under imbalance, we relied on balanced accuracy and macro-averaged measures to weight classes equally [Kuhn and Johnson 2013, Provost and Fawcett 2013]. We also used PR-based criteria for threshold selection, which are typically more informative when positive classes are rare, and inspected confusion matrices to understand class-specific failure modes [Davis and Goadrich 2006]. Under abstention, coverage captures how often the system decides automatically, while conditional metrics on the decided subset quantify decision quality when the model does decide.

4.1. Binary-Target Model Results

Table 1 reports the average performance under stratified cross-validation for the three feature configurations. Among them, the full clinical dataset achieved the strongest overall results (*F1 macro* = 0.734) and the highest *precision macro* (0.816). In contrast, the reduced panels tended to be more selective, but at the cost of sensitivity under some decision regimes.

Table 1. Performance summary (mean over stratified cross-validation) for the binary target.

Feature set	*F1 macro*	*Bal. Acc.*	*Prec. macro*
4 variables	0.664	0.719	0.670
10 variables	0.604	0.684	0.639
Clinical dataset	0.734	0.719	0.816

To illustrate real-world behavior, we report the *out-of-fold* confusion matrix for the best-performing model (clinical dataset) under the optimized-threshold policy.

Table 2. *OOF* confusion matrix for the binary target (clinical dataset) under the optimized-threshold policy.

	Pred. HEALTHY	Pred. COMPLICATION
True HEALTHY	540	131
True COMPLICATION	96	270

To increase clinical safety, we used an abstention-based triage scheme in which the model automatically decides only when confidence is sufficiently high. For the best

binary model (clinical dataset), this policy achieved 70.0% coverage, with *NPV* = 0.944 and sensitivity = 0.985 on the decided subset (Table 3).

Table 3. Abstention-based triage for the binary target (clinical dataset): metrics on the decided subset.

Metric	Value
Coverage (decided)	70.0%
Sensitivity (decided)	0.985
Specificity (decided)	0.211
PPV	0.500
NPV	0.944

Unlike Table 2, which covers the full cohort without abstention, Table 3 reports metrics only for automatically decided cases; low-confidence cases are deferred for human review.

4.2. Ternary-Target Model Results

In the ternary setting (healthy, complication, death), the clinical dataset delivered a clear improvement over the reduced feature panels, reaching *precision macro* = 0.722, *F1 macro* = 0.683, and *balanced accuracy* = 0.694 (Table 4). The 4- and 10-variable configurations performed worse, which is consistent with the challenge of separating rare classes with limited information.

Table 4. Performance summary (mean over stratified cross-validation) for the ternary target.

Feature set	*Prec. macro*	*F1 macro*	*Bal. Acc.*
4 variables	0.448	0.377	0.495
10 variables	0.485	0.489	0.518
Clinical dataset	0.722	0.683	0.694

Next, we report the *out-of-fold* confusion matrix for the best-performing model (clinical dataset) under the standard *argmax* rule, which helps highlight where misclassifications concentrate across classes.

Table 5. *OOF* confusion matrix for the ternary target (clinical dataset) under the *argmax* rule.

	Pred. HEALTHY	Pred. COMPLICATION	Pred. DEATH
True HEALTHY	627	0	44
True COMPLICATION	11	48	3
True DEATH	165	25	114

Under the abstention-based policy, the best ternary model (clinical dataset) automatically decided 86.9% of cases, achieving *F1 macro* = 0.714 and *accuracy* = 0.728 on the decided subset (Table 6). This indicates cleaner separation between classes once low-confidence cases are deferred for human review.

Table 6. Abstention-based triage for the ternary target (clinical dataset): metrics on the decided subset.

Metric	Value
Coverage (decided)	86.9%
Accuracy (decided)	0.728
F1 macro (decided)	0.714
Precision macro (decided)	0.699
Recall macro (decided)	0.756

5. Discussion

The results suggest that feature composition is a key driver not only of overall performance but also of decision safety in triage settings. Across both targets, the full clinical dataset achieved the strongest aggregate indicators, outperforming the 4- and 10-variable panels in *precision macro* and *F1 macro*. This is consistent with the idea that, especially when outcomes are rare, separating classes relies on variables that capture different dimensions of clinical status, reducing ambiguity that arises when information is limited.

For the binary target, there is a clear tension between selectivity and sensitivity. Reduced panels can lead to more "conservative" decisions, but often at the expense of sensitivity, increasing the risk of missing patients who will experience a cardiovascular outcome. This is particularly problematic in triage, where false negatives can delay referral and reduce the window for effective intervention. In this context, optimized thresholds and probabilistic calibration matter because they let the operating point be aligned with the clinical goal, rather than relying on a default cutoff or the *argmax* rule.

For the ternary target, the task is harder because complication and death are minority and clinically heterogeneous classes. Their small sample sizes may increase metric variability and threshold sensitivity to fold composition. Even so, the clinical dataset produced substantially better macro metrics, suggesting cleaner overall separation between the three classes. The confusion matrices indicate that a portion of errors concentrates in confusions between the healthy class and higher-risk classes, reinforcing the value of decision regimes with class-specific thresholds and abstention to avoid automatic decisions in low-confidence regions.

Prior work also suggests that reported performance in Chagas-related ML is sensitive to the prediction task, the data modality (structured clinical variables, biomarkers, ECG), and outcome prevalence, so direct comparisons across studies should be interpreted cautiously. Ferreira et al. [Ferreira 2022] reported a *ROC AUC* around 0.74 for 2-year mortality, with better performance when laboratory exams were available and with NT-*proBNP* playing a prominent role. In our setting, NT-*proBNP* and related features consistently appear among the top permutation importances in both reduced panels and the full clinical dataset, which aligns with that evidence. In the 10-variable configuration, clinically meaningful markers of severity and cardiac function also emerge, such as NYHA class and TAPSE. In the full clinical dataset, structural and laboratory parameters remain relevant, and in the ternary task, the importance of variables such as end-diastolic volume index and creatinine is consistent with the expectation that cardiac remodeling and systemic dysfunction contribute to discriminating severity and mortality risk.

In ECG-based deep learning studies such as Oliveira-da-Silva et al.

[Oliveira 2021], higher performance reflects a different modality with richer signal content and a different task formulation. Work such as Jidling et al. [Jidling 2023] and Ghilardi et al. [Ghilardi 2024] shows that performance can drop in broader population screening due to domain shift and prevalence differences, which reinforces the need for external validation before any operational use. Still, what is often less explicit in these studies is the decision policy itself. Here, calibrated probabilities, explicit thresholds, and abstention allow us to report how often the system decides automatically (coverage) and how reliable it is when it does decide (conditional metrics on the decided subset), which is central for safer deployment in low-resource settings.

6. Conclusion

This study investigated cardiovascular risk stratification in patients with Chagas disease through two complementary perspectives: the evaluation of a novel set of 10 derived clinical variables and an interpretable decision-support framework based on calibrated decision trees, threshold optimization, and abstention mechanisms. The results showed that the comprehensive clinical dataset achieved the best predictive performance, while the derived variables provided a compact representation, outperforming the 4-variable panel only in the ternary setting. Furthermore, the proposed framework improved the sensitivity–specificity trade-off while preserving decision coverage, supporting safer clinical triage. A limitation of this study is the lack of external validation, which may affect the generalizability of the findings. Future work includes validating the framework in independent cohorts and real clinical environments, as well as assessing its impact on routine decision-making and patient outcomes.

Acknowledgments

This study was financed in part by the PRPI-USP (Grant number: 22.1.09345.01.2), the São Paulo Research Foundation (FAPESP) (Grant number 22/03090-0), the Brazilian National Council for Scientific and Technological Development (CNPq) (Grants 420025/2023-5 and 406941/2025-4), and the ICMC-GRANT. The authors also acknowledge the use of AI-based tools to support text revision.

References

Abreu, V. V. (2021). Artificial intelligence-based decision for the prediction of cardioembolism in patients with chagas disease and ischemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, 30(10):106034.

Chapman, P. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Technical report, CRISP-DM Consortium.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pages 233–240.

Domingues, P. (2020). Você sabe o que é CRISP-DM? CRISP-DM é a abreviação de Cross Industry Standard Process for Data Mining. Bexs.IO.

Ferreira, A. M. (2022). Two-year death prediction models among patients with chagas disease using machine learning-based methods. *PLOS Neglected Tropical Diseases*, 16(4):e0010356.

Ghilardi, J. (2024). Machine learning for predicting chagas disease infection in rural areas of brazil. *PLOS Neglected Tropical Diseases*, 18(9):e0012026.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Jidling, C. (2023). Screening for chagas disease from the electrocardiogram using a deep neural network. *PLOS Neglected Tropical Diseases*, 17(7):e0011118.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(195).

Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(4).

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, pages 625–632, New York, NY, USA. ACM.

Oliveira, B. (2021). Left ventricular systolic dysfunction predicted by artificial intelligence using the electrocardiogram in chagas disease patients – the sami-trop cohort. *PLOS Neglected Tropical Diseases*, 15(7):e0009974.

Parro, H., Ortiz, A., Valejo, A., García, L., and Garcés, L. (2026). Chagas decision tree: Data and code repository. `https://github.com/blind/chagas-decision-tree/`. GitHub repository, accessed February 25, 2026.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Rassi, A., Rassi, A., and Marin-Neto, J. A. (2010). Chagas disease. *The Lancet*, 375(9723):1388–1402.

Weber de Santana Teles, S. (2021). Machine learning and automatic selection of attributes for the identification of chagas disease from clinical and sociodemographic data. *Research, Society and Development*, 10(4):e19310413879.

World Health Organization (2025). Chagas disease (also known as american trypanosomiasis). `https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)`. WHO Fact Sheet, updated 2 April 2025.