Machine Learning-Based Clinical Triage for Leptospirosis Using SINAN/SUS Data

Ana Carolyne Pereira[1], **Diego Pereira Lima**[1], **Lilian Berton**[1]

Specialization Course in Data Science – Instituto Tecnológico de Aeronáutica (ITA)
Praça Mal. Eduardo Gomes 50 – São José dos Campos – SP – Brasil

{Ana.pereira.20818, diego.lima.20829}@ga.ita.br, lilian.berton@gp.ita.br

Abstract. *This study proposes a predictive model to support the clinical triage of suspected leptospirosis cases using data from the Notifiable Diseases Information System (SINAN), based on clinical, demographic, and epidemiological variables. The LightGBM model achieved the best performance in classifying confirmed versus discarded cases (AUC $\approx$ 0.93) and moderate performance in predicting clinical severity (AUC $\approx$ 0.70). Threshold analysis indicated that values below 0.5 are more appropriate for prioritizing sensitivity and reducing false negatives in triage settings.*

1. Introduction

Leptospirosis is a globally distributed zoonosis and a relevant public health problem, especially in tropical countries and in contexts of socio-environmental vulnerability [Pappas et al. 2008]. In Brazil, it is endemic, with seasonal increases associated with floods, inadequate sanitation, the presence of rodents, and occupational or environmental exposure to contaminated water and soil [Brasil. Ministério da Saúde 2024]. Its clinical presentation ranges from mild and nonspecific cases to severe forms, such as acute kidney injury, hemorrhagic manifestations, respiratory impairment, and Weil's syndrome, which hinders early diagnosis and increases the risk of morbidity and mortality when treatment is delayed [Costa et al. 2015, Li et al. 2022].

In Brazil, suspected cases are recorded in the Notifiable Diseases Information System (SINAN), which gathers clinical, demographic, epidemiological, and outcome-related variables [Brasil. Ministério da Saúde 2016]. Despite this availability, the systematic use of these data to support triage and prioritization remains limited within the Brazilian Unified Health System (SUS), especially in scenarios of high demand, resource constraints, and regional heterogeneity [Brasil. Ministério da Saúde 2023, da Silva et al. 2014, de Araujo Lima et al. 2009]. Thus, risk assessment tends to depend on the individual experience of healthcare professionals, increasing variability in case prioritization [Brasil. Ministério da Saúde 2014].

From a data science perspective, the challenge lies in transforming public surveillance databases, characterized by missing values, noise, heterogeneity, and class imbalance, into useful evidence for decision support. Supervised models, especially *ensemble learning* and *gradient boosting* methods, can capture nonlinear relationships and interactions among variables in clinical and operational public health contexts [Chen and Guestrin 2016, Reinheimer et al. 2025]. However, their application requires

careful experimental design, metrics suitable for imbalanced data, and a critical analysis of the limits of their use in triage.

Given this scenario, this study seeks to answer the following research question: to what extent can supervised machine learning models, trained on SINAN/SUS data from 2015 to 2024, support the prioritization of suspected leptospirosis cases and the stratification of clinical severity? To this end, we propose a supervised classification pipeline including preprocessing, imbalance handling, feature selection, hyperparameter optimization, cross-validation, and decision-threshold adjustment. As a contribution, this paper organizes the problem into two complementary tasks, compares classifiers using metrics suitable for imbalanced scenarios, and discusses the applicability, interpretability, and limitations of predictive models as support for triage in the SUS.

2. Related Work

Machine learning in healthcare has expanded in clinical prediction, diagnosis, and decision support, including applications in public health. However, challenges related to bias, missing values, data heterogeneity, and interpretability in real-world settings persist [Mooney and Pejaver 2018].

In leptospirosis, recent studies report promising performance of *machine learning* and *deep learning* models, but also highlight limitations such as the predominant use of hospital-based or private datasets, limited external validation, and restricted population-level generalizability [Sawesi et al. 2025]. Studies on severity or mortality follow a similar pattern, showing relevant results but relying on specific contexts, which limits their transferability to population-level triage [Galdino et al. 2023, Reagan et al. 2022].

In Brazil, applications using SINAN data have already been explored for predicting final outcomes, such as recovery or death, and for classifying suspected cases in regional datasets. Although these studies used algorithms such as *Random Forest*, AdaBoost, SVM, and neural networks, limitations remain regarding data quality, missing values, inconsistencies, and limited emphasis on clinical triage and explicit decision-threshold adjustment [da Silva et al. 2024, Bohm 2024].

Thus, gaps persist regarding the operational application of predictive models for leptospirosis in clinical triage, especially using national SINAN/SUS data, robust validation, and *threshold* adjustment aimed at reducing false negatives. This work addresses this gap by proposing a supervised model to support the confirmation of suspected cases and the stratification of clinical severity, incorporating feature engineering, feature selection, and analysis of the *trade-off* between sensitivity and specificity.

3. Methodology

3.1. Data and Initial Preparation

This study is characterized as a retrospective observational study based on public and anonymized SINAN/SUS data, comprising approximately 170,000 records of suspected leptospirosis cases in Brazil between 2015 and 2024 [Brasil. Ministério da Saúde 2016, Brasil. Ministério da Saúde 2025]. The dataset includes clinical, demographic, epidemiological, and outcome-related variables from the Individual Notification Form (FIN), an

instrument used in national epidemiological surveillance to record compulsory notifiable diseases.

The selection of predictive variables was guided by clinical and epidemiological evidence, covering demographic, socio-environmental, seasonal, and exposure-related attributes, as well as signs and symptoms available at the time of notification. These included age, sex, education level, race/color, occupation, probable exposure environment, epidemiological history, rainy season, and initial clinical manifestations such as fever, myalgia, headache, prostration, conjunctival suffusion, calf pain, vomiting, and diarrhea [Brasil. Ministério da Saúde 2014, Li et al. 2022, Bohm 2024, Costa et al. 2015]. These variables correspond exclusively to attributes available at the time of notification, distinguishing them from variables dependent on subsequent clinical progression, such as systemic complications and final outcome, which were used only to construct the target variables. Preprocessing included missing-value handling, standardization, creation of derived socio-environmental and seasonal attributes, and encoding of categorical variables, aiming to preserve as many records as possible and mitigate sample bias caused by excessive exclusions. The complete variable dictionary is provided in the supplementary material.

The initial exploratory analysis was used only as a step to verify data consistency, identify missingness patterns, inspect marginal distributions, and support preliminary feature selection. The epidemiological interpretation of the observed patterns is presented in Section 4.1, together with the results of dimensionality reduction and the analysis of the selected attributes.

3.2. Classification Task Formulation

The problem was formulated as two supervised binary classification tasks. The first task consisted of distinguishing between confirmed and discarded leptospirosis cases, using as reference the final classification recorded in SINAN [Brasil. Ministério da Saúde 2016, Pongpan et al. 2023, da Silva et al. 2014]. This formulation aims to estimate the probability of confirmation among suspected cases, potentially supporting prioritization processes in scenarios of diagnostic uncertainty.

The second task consisted of stratifying clinical severity through a derived binary variable. Records with markers compatible with systemic forms of leptospirosis were classified as severe, including respiratory impairment, pulmonary hemorrhage, hemorrhagic manifestations, cardiac impairment, meningeal involvement, Weil's syndrome, represented by the combination of jaundice and renal dysfunction, or the co-occurrence of renal dysfunction/bleeding with critical organ involvement [Bohm 2024, Brasil. Ministério da Saúde 2014]. The remaining records were classified as non-severe. To avoid information leakage, the variables directly used to construct this outcome were not included in the set of explanatory variables used to train the models. Thus, the severity task was evaluated using demographic, epidemiological, socio-environmental, seasonal, and sign/symptom attributes that were not used in the label definition.

3.3. Predictive Modeling Strategy

Modeling was structured as a unified pipeline integrating preprocessing, feature selection, imbalance handling, and classifier training, in order to preserve the validity

of the evaluation and reduce the risk of information leakage. Preprocessing included median imputation and standardization for numerical variables, and mode imputation and encoding through indicator variables (*dummy variables*) for categorical attributes [Kutner et al. 2005]. In both cases, imputation parameters were estimated exclusively from the training data of each fold and applied to the respective test sets, ensuring the absence of information leakage between partitions, a procedure adopted in all data-dependent steps.

Considering the imbalance of the main task, with a predominance of discarded cases (65%–70%) over confirmed cases (30%–35%), resampling strategies were evaluated during optimization, including random *oversampling*, random *undersampling* [He and Garcia 2009], and the *Synthetic Minority Over-sampling Technique* (SMOTE) [Chawla et al. 2002]. These techniques were applied only to the training data in each cross-validation iteration. Internal imbalance-adjustment mechanisms were also considered, such as the `scale_pos_weight` parameter in XGBoost, aiming to favor the identification of the minority class and increase sensitivity, a central metric in clinical triage due to its association with reducing false negatives.

Feature selection was incorporated into the pipeline to reduce dimensionality, improve computational efficiency, and favor interpretability [Guyon and Elisseeff 2003]. After preprocessing, filter methods were combined, including *Variance Threshold*, *Mutual Information*, and ANOVA/F-regression-based tests, capable of capturing low variance, nonlinear associations, and statistical relationships between attributes and the target variable [Kraskov et al. 2004, Kutner et al. 2005]. Next, an *embedded* method based on Random Forest was used to estimate feature importance, followed by cross-validation evaluation to compose the final subset of attributes. This strategy allowed us to compare gains in efficiency and interpretability against possible losses in predictive performance.

Four supervised algorithms consolidated for heterogeneous tabular data were evaluated: Random Forest [Breiman 2001], Support Vector Machines (SVM) [Cortes and Vapnik 1995], XGBoost [Chen and Guestrin 2016], and LightGBM [Ke et al. 2017]. The choice of these models combined traditional baselines and competitive methods based on *gradient boosting*, covering different learning paradigms, such as trees, margins, and ensembles. SVM received probabilistic calibration using a sigmoid method to enable probability-based metrics, while the *boosting* models had their complexity and regularization hyperparameters optimized [Friedman 2001]. In the severity task, variables directly used to construct the label were excluded from the predictive set before modeling, preserving the independence between explanatory attributes and the target variable. The complete search space for hyperparameters and balancing strategies is described in the supplementary material.

3.4. Validation, Optimization, and Decision-Threshold Adjustment

Model evaluation was conducted using *Nested Cross-Validation*, a strategy recommended in scenarios involving hyperparameter optimization because it separates the tuning and evaluation stages, reducing the risk of overfitting and optimistic performance estimates [Berrar 2019]. Validation was structured at two levels: an outer loop, responsible for estimating generalizable performance, using stratified cross-validation with four folds; and an inner loop, applied only to the training data of each outer fold, using stratified cross-validation with three folds for model, hyperparameter, and balancing-strategy selection.

Optimization was performed with the Optuna library, based on Bayesian search, allowing simultaneous selection of the algorithm, its hyperparameters, and the balancing technique most suitable for the data [Akiba et al. 2019]. The objective function was defined as the mean AUC-ROC in the inner validation, as it assesses the overall discriminative capacity of the models. Although AUC-ROC guided optimization, the final analysis included metrics more sensitive to class imbalance and to the triage context, such as sensitivity, F1-score, and area under the Precision–Recall curve (PR-AUC). To reduce computational cost, early interruption mechanisms (*pruning*) were employed, concentrating resources on the most promising configurations [Li et al. 2018]. At the end of each outer loop, the best hyperparameter set was trained on the complete training set and evaluated on the corresponding test set.

Performance was analyzed using metrics aligned with the clinical triage context, including sensitivity (*recall*), specificity, AUC-ROC, and PR-AUC. Since false negatives have a higher clinical and social cost in leptospirosis, the decision threshold was also evaluated, analyzing the behavior of false positives (FP) and false negatives (FN) at different cutoff points [Fawcett 2006, Reagan et al. 2022]. The equivalence between FP and FN was used as an operational reference, but the final choice of thresholds prioritized sensitivity and F1-score, aiming to reduce false negatives without an operationally unfeasible increase in false positives.

4. Analysis and Discussion of Results

4.1. Dimensionality Reduction and Computational Efficiency

The exploratory analysis of SINAN demographic, clinical, and epidemiological variables indicated patterns consistent with the literature on leptospirosis and served as a plausibility check for the attributes used in the modeling process. Records were predominantly associated with male individuals and economically active age groups, with increased lethality observed among older groups. Clinically, a high frequency of fever, myalgia, and headache was observed among suspected cases, while records classified as severe concentrated a higher occurrence of systemic manifestations.

From an epidemiological perspective, exposures associated with synanthropic animals, mud, garbage, contaminated water, and animal breeding reinforced the relevance of socio-environmental determinants in characterizing suspected cases. These findings supported the retention of clinical, demographic, and environmental variables in the set of attributes evaluated by the models.

Feature selection substantially reduced the dimensionality of the problem. For diagnostic confirmation, the set decreased from 43 to 13 variables, a reduction of approximately 69.8%; for clinical severity, it decreased from 36 to 16 variables, a reduction of around 55.6%. In both tasks, the selected subset retained socio-environmental, demographic, and clinical attributes, preserving dimensions relevant to the modeling process.

Dimensionality reduction also decreased computational cost in the same experimental environment. In LightGBM, the average training time decreased from 29.3s to 13.8s for diagnostic confirmation, a reduction of approximately 52.9%, and from 15.1s to 5.7s for clinical severity, a reduction of around 62.3%. SVM showed the highest computational cost, a behavior consistent with the sensitivity of margin-based models to the dimensionality of the feature space.

Overall, the findings indicate that dimensionality reduction simplified the feature space and improved the efficiency of the pipeline, although it represents a trade-off in relation to the predictive performance obtained with the complete set of variables. This balance between performance, efficiency, and interpretability is relevant for triage applications based on epidemiological surveillance data.

4.2. Predictive Performance and Discriminative Capacity

Internal validation indicated consistently superior performance of LightGBM compared with XGBoost, SVM, and Random Forest in most configurations, suggesting a greater ability of the algorithm to capture nonlinear relationships and interactions among variables in a heterogeneous and noisy epidemiological dataset. For diagnostic confirmation, all models achieved an AUC above 0.89, whereas for clinical severity the performance was moderate, with AUC ≈ 0.70, indicating greater overlap between the clinical profiles of the classes. The complete results by algorithm and experimental configuration are provided in the supporting material and appendices.

In the outer cross-validation, LightGBM maintained the best average performance, reinforcing its consistency and generalization capacity. For diagnostic confirmation, the approach without feature selection achieved a mean AUC close to 0.93, with gains in precision and F1-score, suggesting that the complete set of attributes captures complementary interactions among clinical, demographic, and epidemiological variables. For clinical severity, performance was moderate (AUC ≈ 0.70), with small differences across scenarios, indicating greater intrinsic complexity of the task and lower separability between classes. Part of this limitation can be attributed to the static and retrospective nature of the Individual Notification Form: since severity is a dynamic process in which the patient may deteriorate within 48 to 72 hours, the model relies only on a single point-in-time record, without temporal progression data or laboratory results, representing a structural limitation of static tabular models in capturing this dimension of the disease.

Table 1. Mean LightGBM performance in the outer cross-validation for diagnostic confirmation and clinical severity, with and without feature selection, considering decision thresholds adjusted for each task.

Task	Approach	Threshold	Accuracy	Precision	Recall	F1-score	AUC
Confirmation	With selection	0.40	0.904 ± 0.002	0.761 ± 0.005	0.762 ± 0.005	0.761 ± 0.005	0.910 ± 0.004
Confirmation	Without selection	0.40	0.911 ± 0.001	0.781 ± 0.002	0.779 ± 0.003	0.779 ± 0.002	0.931 ± 0.003
Severity	With selection	0.48	0.652 ± 0.007	0.54 ± 0.08	0.55 ± 0.02	0.54 ± 0.01	0.69 ± 0.01
Severity	Without selection	0.48	0.66 ± 0.01	0.56 ± 0.02	0.56 ± 0.01	0.56 ± 0.02	0.70 ± 0.02

The ROC curves confirmed high discriminative performance for diagnostic confirmation and moderate performance for severity. The Precision–Recall curves, which are more informative in imbalanced scenarios, showed AUPRC values around 0.81–0.82 for confirmation and 0.55–0.59 for severity, reinforcing the greater difficulty of the second task.

Although the approach without feature selection showed slightly superior performance, the subsequent analyses were conducted using LightGBM integrated into the feature-selection pipeline. This choice represents a trade-off between predictive performance, computational efficiency, reduced complexity, and greater interpretability, all of

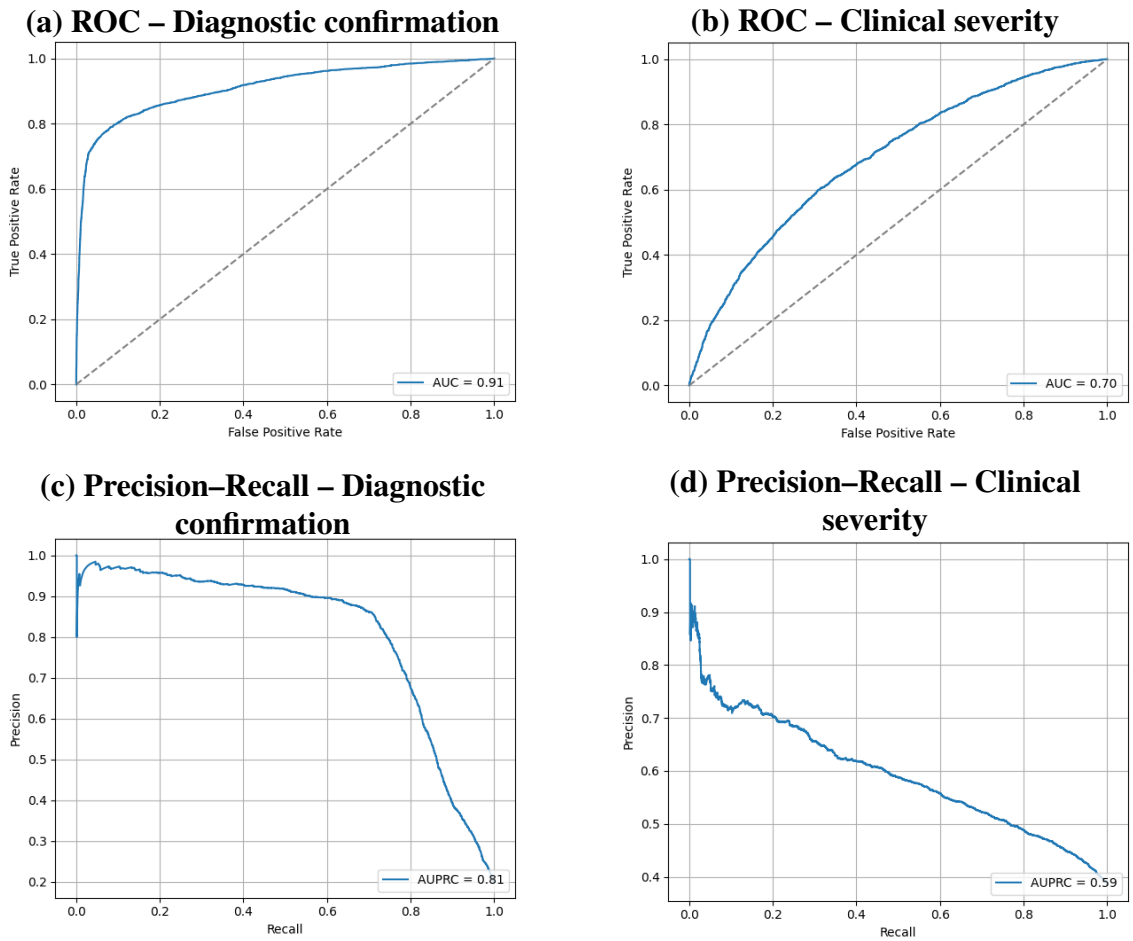


Figure 1. ROC and Precision–Recall curves of the LightGBM model in the outer cross-validation: (a) ROC for diagnostic confirmation; (b) ROC for clinical severity; (c) Precision–Recall for diagnostic confirmation; (d) Precision–Recall for clinical severity.

which are relevant aspects for public health applications. The fold-level results and complementary curves are available in the supporting material and appendices.

4.3. Decision-Threshold Analysis

The definition of the decision threshold is relevant in clinical triage because the costs of errors are asymmetric. In leptospirosis, false negatives may delay diagnosis and treatment initiation, increasing the risk of progression to severe forms. Therefore, the behavior of false positives (FP) and false negatives (FN) was evaluated across different *thresholds*.

For diagnostic confirmation, lowering the threshold reduced false negatives with a gradual increase in false positives, with a consistent intersection point across folds around 0.39–0.40. A *threshold* of 0.40 was adopted, below the conventional value of 0.5, prioritizing sensitivity in the identification of confirmed cases. The resulting increase in false positives is epidemiologically justifiable, given that the cost of initial treatment with antibiotic therapy is substantially lower than the human and financial cost of progression to severe disease requiring ICU admission.

For clinical severity, the balance point occurred around 0.47–0.49, indicating a

smaller shift relative to the standard threshold and a greater balance between sensitivity and specificity. A *threshold* of 0.48 was adopted, preserving the reduction of false negatives without excessively compromising overall precision. These results indicate that task-adjusted thresholds are more appropriate for the epidemiological triage context than the fixed use of 0.5.

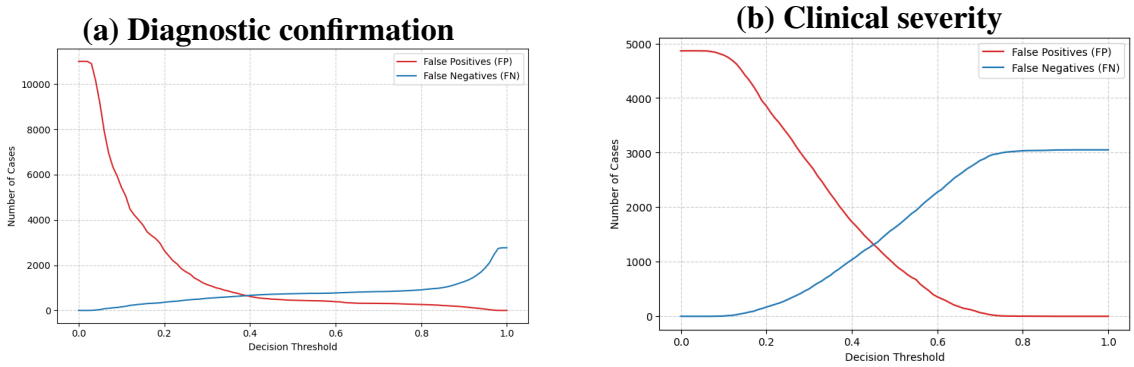


Figure 2. False positive (FP) and false negative (FN) curves as a function of the decision threshold for: (a) classification between discarded and confirmed cases; (b) classification of clinical severity.

4.4. Calibration, Interpretability, and Applicability in Triage

Probabilistic calibration was evaluated because *boosting* models may show good discrimination but poorly calibrated probabilities [Niculescu-Mizil and Caruana 2007]. The outer-validation curves indicated reasonable agreement between predicted probabilities and observed frequencies, suggesting that the model estimates may support case prioritization in triage and epidemiological surveillance.

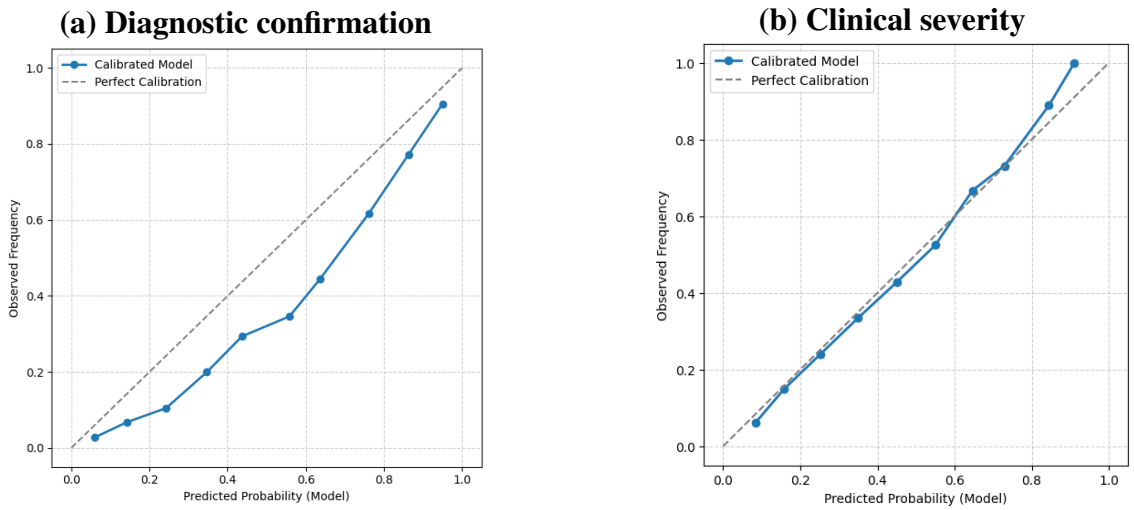


Figure 3. Calibration curves of the LightGBM model in the outer cross-validation for: (a) classification between discarded and confirmed cases; (b) classification of clinical severity.

Interpretability was analyzed based on the mean feature importance in the outer cross-validation, considering its stability across folds [Berrar 2019]. For diagnostic confirmation, sociodemographic, clinical, and environmental variables stood out, such as

occupation, education level, age, jaundice, calf pain, and exposure conditions, a pattern consistent with the multifactorial nature of the disease [Brasil. Ministério da Saúde 2014, Bohm 2024, Brasil. Ministério da Saúde 2016].

For clinical severity, the most relevant variables included state of notification, prostration, conjunctival suffusion, calf pain, age, and indicators of environmental exposure, suggesting that severity is conditioned not only by isolated clinical signs but also by regional heterogeneity, sanitation infrastructure, and differences in access to healthcare [Costa et al. 2015, Brasil. Ministério da Saúde 2023, Li et al. 2022].

The comparison between tasks highlights distinctions in the mechanisms captured by the model: diagnostic confirmation was predominantly associated with exposure factors, socioeconomic conditions, and clinical manifestations compatible with the disease, whereas severity showed a stronger relationship with systemic manifestations, regional context, and healthcare-related determinants. This differentiation is consistent with the exploratory analysis, in which variables such as occupation, education level, contact with mud, age, and jaundice reflect patterns of vulnerability and environmental exposure, while state of notification, prostration, and conjunctival suffusion capture the clinical and regional heterogeneity associated with severity. Feature importance reinforces that the patterns learned by LightGBM are consistent with the clinical, demographic, and socio-environmental dimensions of leptospirosis. The complete feature-importance values and fold-level analyses are provided in the supporting material and appendices.

5. Conclusion

This study investigated the use of machine learning on SINAN/SUS data to support diagnostic confirmation and clinical severity stratification in suspected leptospirosis cases. LightGBM showed high discriminative capacity in classifying confirmed and discarded cases, with an AUC close to 0.93 in outer cross-validation, and moderate performance in clinical severity stratification, with an AUC around 0.70. The ROC and Precision–Recall curves indicated consistency across folds, and the threshold analysis showed that values below 0.5 are more suitable for triage, as they prioritize the reduction of false negatives.

From an applied perspective, the results indicate greater potential for supporting diagnostic confirmation, the task in which the model showed more robust performance. For clinical severity, the findings should be interpreted as preliminary evidence of risk stratification, given the greater overlap between severe and non-severe profiles. The main limitations include incompleteness and inconsistencies in secondary data, regional heterogeneity, and the absence of independent external validation.

As future work, we highlight the incorporation of environmental, climatic, laboratory, and territorial variables; external and prospective validation in different epidemiological scenarios; and cross-validation by geographic blocks to assess the model's generalization capacity in states underrepresented in SINAN. We also consider integration with multi-agent decision-support architectures, capable of coordinating specialized modules for the interpretation of unstructured clinical reports and support for triage prioritization in the SUS.

Statement: The linguistic and stylistic refinement of this text was supported by a large language model (LLM), while the technical, scientific, and factual content remains the sole responsibility of the authors.

6. Acknowledgments

The authors thank the Data Science Specialization Program at Instituto Tecnológico de Aeronáutica (CEDS-ITA) for the academic and institutional support provided throughout this work. The authors also thank the Brazilian National Council for Scientific and Technological Development (CNPq) for its support.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Berrar, D. (2019). Cross-validation.

Bohm, B. C. (2024). Avaliação de dados do sistema de saúde e elaboração de ferramentas de triagem. pages 91–105.

Brasil. Ministério da Saúde (2014). Leptospirose: diagnóstico e manejo clínico. Technical report, Ministério da Saúde.

Brasil. Ministério da Saúde (2016). Sistema de informação de agravos de notificação (sinan): normas e rotinas. Technical report, Ministério da Saúde.

Brasil. Ministério da Saúde (2023). Sistema Único de saúde (sus): estrutura, princípios e funcionamento. Technical report, Ministério da Saúde.

Brasil. Ministério da Saúde (2024). Guia de vigilância em saúde: volume 3. Technical report, Ministério da Saúde.

Brasil. Ministério da Saúde (2025). Datasus-tabnet. Acesso em: 2025.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Costa, F., Hagan, J. E., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M. S., Stein, C., Abela-Ridder, B., and Ko, A. I. (2015). Global morbidity and mortality of leptospirosis: a systematic review. *PLoS neglected tropical diseases*, 9:e0003898.

da Silva, A. F., Figueiredo, K., Falcão, I. W. S., Costa, F. A. R., da Rocha Seruffo, M. C., and de Moraes, C. C. G. (2024). Study of machine learning techniques for outcome assessment of leptospirosis patients. *Scientific Reports*, 14:13929.

da Silva, L. P., Moreira, C. M. M., Amorim, M. H. C., de Castro, D. S., and Zandonade, E. (2014). Avaliação da qualidade dos dados do sistema de informações sobre nascidos

vivos e do sistema de informações sobre mortalidade no período neonatal, espírito santo, brasil, de 2007 a 2009. *Ciência & Saúde Coletiva*, 19:2011–2020.

de Araujo Lima, C. R., de Andrade Schramm, J. M., Coeli, C. M., and da Silva, M. E. M. (2009). Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. *Cadernos de saúde pública*, 25:2095–2109.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27:861–874.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Galdino, G. S., de Sandes-Freitas, T. V., de Andrade, L. G. M., Adamian, C. M. C., Meneses, G. C., da Silva Junior, G. B., and de Francesco Daher, E. (2023). Development and validation of a simple machine learning tool to predict mortality in leptospirosis. *Scientific Reports*, 13:4506.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21:1263–1284.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69:66138.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, english edition.

Li, D., Liang, H., Yi, R., Xiao, Q., Zhu, Y., Chang, Q., Zhou, L., Liu, B., He, J., Liu, T., et al. (2022). Clinical characteristics and prognosis of patient with leptospirosis: A multicenter retrospective analysis in south of china. *Frontiers in Cellular and Infection Microbiology*, 12:1014530.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of machine learning research*, 18:1–52.

Mooney, S. J. and Pejaver, V. (2018). Big data in public health: Terminology, machine learning, and privacy. *Annual Review of Public Health*, 39:95–112.

Niculescu-Mizil, A. and Caruana, R. (2007). Obtaining calibrated probabilities from boosting. volume 5, pages 413–420.

Pappas, G., Papadimitriou, P., Siozopoulou, V., Christou, L., and Akritidis, N. (2008). The globalization of leptospirosis: worldwide incidence trends. *International journal of infectious diseases*, 12:351–357.

Pongpan, S., Thanatrakolsri, P., Vittaporn, S., Khamnuan, P., and Daraswang, P. (2023). Prognostic factors for leptospirosis infection severity. *Tropical Medicine and Infectious Disease*, 8:112.

Reagan, K. L., Deng, S., Sheng, J., Sebastian, J., Wang, Z., Huebner, S. N., Wenke, L. A., Michalak, S. R., Strohmer, T., and Sykes, J. E. (2022). Use of machine-learning algorithms to aid in the early detection of leptospirosis in dogs. *Journal of Veterinary Diagnostic Investigation*, 34:612–621. PMID: 35603565.

Reinheimer, I. C., Corrêa-Dos-Santos, A., Carlan, B. D., Telöken, I. B., da Silva Lessa-De-Oliveira, V., Braga-Da-Silveira, J., Escouto, D. C., Barra, A. B. L., Strogoff-De-Matos, J. P., and Poli-De-Figueiredo, C. E. (2025). Impacto do balanceamento de dados na performance de modelos de machine learning para predição de óbito de pacientes em hemodiálise no sus. *Braz. J. Nephrol.*, 47:S501.

Sawesi, S., Jadhav, A., and Rashrash, B. (2025). Machine learning and deep learning techniques for prediction and diagnosis of leptospirosis: Systematic literature review. *JMIR Med Inform*, 13:e67859.

Supplementary Materials

The appendices include the dictionary of selected variables, the hyperparameter search space, the exploratory analysis of SINAN variables, and additional validation and feature-importance results, available at:

`https://anonymous.4open.science/r/leptospirosis-ml-sinan-sus-6C96/`