Development of Predictive Models with Random Forest and XGBoost for Risk Stratification in Stroke Patients

Douglas Tranquilino[1], Marcos Vinicius[1], Rafael de Carli[2], Gustavo Callou[3], Eduardo Tavares[4], Thiago Bezerra[1,4]

[1] Instituto Federal de Pernambuco, Palmares, Brasil

[2]Faculdade de Ciências Médicas,
Universidade de Pernambuco, Recife, Brasil

[3]Departamento de Computação,
Universidade Federal Rural de Pernambuco, Recife, Brasil

[4]Centro de Informática,
Universidade Federal de Pernambuco, Recife, Brasil

{tvb,cbff,eagt}@cin.ufpe.br, {jdts1, mvvm}@discente.ifpe.edu.br

gustavo.callou@ufrpe.br, rafael.carli@upe.br

Abstract. *Machine learning (ML) is transforming clinical medicine by enabling the analysis of complex datasets to identify predictive patterns. This capability makes ML particularly valuable in neurology for prognosticating outcomes in complex conditions like stroke. This paper presents a prediction system for clinical risk stratification in stroke patients using ML and open-source data. The system utilizes a comprehensive panel of clinical and laboratory data, including hematological, metabolic, and inflammatory markers, to predict the patient's condition. Two ensemble algorithms, Random Forest and XGBoost, were developed and compared. The results demonstrate the feasibility of this approach in enhancing prognostic accuracy.*

1. Introduction

Stroke imposes a substantial burden on the Brazilian healthcare system, ranking among the leading causes of death and long-term functional disability in the country. Public health data confirm that stroke is consistently one of the top causes of mortality in Brazil, highlighting the urgent need to improve clinical approaches for this condition [dos Santos et al. 2025]. In this context, the ability to accurately and early stratify risk and predict clinical outcomes is a cornerstone for optimizing patient care. The successful implementation of such predictive systems relies on modern technological frameworks capable of collecting and processing patient data, such as those based on the Internet of Things (IoT) and machine learning for continuous patient monitoring [Bezerra et al. 2025].

Conventional prognostic assessment in stroke, while anchored in factors such as age and neurological scales, offers a limited view of the complex systemic dysregulation that follows acute brain injury. The present study proposes a data-driven approach,

analyzing a comprehensive panel of ten routine biomarkers that reflect critical pathophysiological axes. This panel includes markers from the complete blood count (hemoglobin, hematocrit, platelet and white blood cell counts), the metabolic and renal profiles (glucose, creatinine), the coagulation panel (PT, PTT), and electrolytes (sodium, potassium).

Each of these panels provides a window into a distinct physiological system. The complete blood count informs on oxygen-carrying capacity, inflammatory status, and thrombotic potential. Metabolic and renal markers reflect acute stress and vital organ function. Coagulation parameters are crucial for assessing hemorrhagic or thrombotic risk, while electrolytes are fundamental for neuronal and cardiac homeostasis. The interdependence among these ten markers creates an analytical challenge that transcends traditional regression models, justifying the use of machine learning (ML) methodologies like Random Forest and XGBoost, which are capable of modeling these complex interactions [Tam et al. 2019, Mitsios et al. 2018, Jiang et al. 2024].

This paper presents a predictive framework for prognostication in stroke patients, using machine learning to elucidate clinical risk from a set of serum biomarkers. The primary objective is to develop a system capable of stratifying patients into different risk levels while comparing the performance of two state-of-the-art ensemble algorithms, Random Forest and XGBoost, aiming to enhance prognostic accuracy and the quality of care. To achieve this, the framework operates in a two-stage process: it first tackles a regression task to forecast the values of the biomarkers, and subsequently applies a heuristic to these predictions to stratify patients into clinical risk levels. These two algorithms were selected as they represent the two leading and highest-performing ensemble learning approaches for tabular data: bagging (Random Forest) and boosting (XGBoost). The direct comparison between their distinct methodological philosophies provides a robust benchmark for the present prediction task. However, a contribution of this work extends beyond model development to include the creation of an interactive web application. This tool is designed to translate the model's complex predictions into a visual and easily interpretable format, serving as a prototype of a clinical decision support system for real-time use.

The remainder of this paper is organized as follows. Section 2 presents the related works available in the literature. Section 3 provides an overview of prominent concepts for understanding the proposed approach. In Section 4, the methodology is explained. Section 5 demonstrates the practical application of the proposed solution and presents the results of the model evaluation. Finally, Section 6 concludes this study and shows future research directions.

2. Related work

This section presents a review of recent research efforts that apply ML techniques to prognosis and risk stratification in patients with stroke. The studies are organized around key methodological approaches, such as the performance comparison between different algorithms, the optimization of clinical workflows, and the focus on specific populations or data sources. These works contribute to the understanding of current capabilities and limitations in the field and serve as a basis for identifying the distinctive contributions of the approach proposed in this study.

A significant line of research focuses on the empirical comparison of algorithms to determine the highest-performing architecture for specific tasks. [Wu and Fang 2020],

for example, focused on the challenge of predicting stroke with imbalanced data from an elderly population. They compared three models (Logistic Regression, SVM, and Random Forest) and demonstrated that the use of data balancing techniques was crucial for improving performance, achieving a maximum AUC of 0.72 with the Regularized Logistic Regression model.

Similarly, [Abujaber et al. 2024] sought to predict a long-term outcome, one-year mortality post-stroke, using data from a national registry. Upon comparing five algorithms, they identified that a more traditional model, Logistic Regression, outperformed the others, highlighting that more complex architectures do not always guarantee the best result. Seeking even greater optimization, [Huang et al. 2023b] explored a more complex ensemble methodology. Their objective was to predict post-stroke mortality using a stacked ensemble approach, which combines the predictions from multiple base algorithms to feed a final meta-model. Another research thrust directs ML toward the optimization of clinical workflows and large-scale risk identification. Focusing on real-time application in the emergency department, [Zheng et al. 2022] developed a model whose objective was the rapid diagnosis of ischemic stroke. Using 15 routine variables, their XGBoost-based model was able to identify the presence of stroke with very high accuracy (AUC > 0.90), demonstrating the potential of ML as a diagnostic support tool in emergency settings.

Finally, several studies delve into specific populations, outcomes, or data sources to generate more targeted insights. [Zhu et al. 2023] demonstrates a multifaceted application of ML, using data from the MIMIC-IV database not only to predict mortality with high precision but also to apply interpretability techniques (SHAP) to analyze how different patients respond to specific treatments, such as the use of warfarin. Aligned with open science, [Huang et al. 2023a] also utilized open-access databases (MIMIC-IV and eICU). Their objective was to develop an interpretable model to predict 28-day mortality in a specific population of hypertensive stroke patients in the ICU. By comparing five ML models, they found XGBoost to be the top performer and, through SHAP analysis, identified 11 important predictors, including RDW, glucose, and white blood cell count.

Table 1 presents a comparative summary of the reviewed literature on the application of machine learning for stroke prognostication. An analysis of the table highlights diverse methodological approaches. Several studies focus on the performance comparison of multiple algorithms, with findings indicating that simpler models like Logistic Regression can outperform more complex architectures in certain contexts [Wu and Fang 2020, Abujaber et al. 2024], while others demonstrate the superiority of advanced ensemble models like XGBoost for specific tasks [Zheng et al. 2022, Huang et al. 2023a]. A growing trend is the use of open-access data combined with interpretability techniques (SHAP) to predict outcomes and analyze treatment effects, as demonstrated by [Zhu et al. 2023] and [Huang et al. 2023a].

Unlike the aforementioned studies, our approach carves a distinct niche. While prior studies explore a wide range of variables, apply interpretability, or focus on specific outcomes such as rapid diagnosis, our work uniquely combines three core pillars in an integrated manner: (1) the direct performance comparison between Random Forest and XGBoost; (2) the exclusive use of a focused panel of ten low-cost, routine serum biomarkers; and (3) the application on open-source data to ensure maximum reproducibility.

Table 1. Summary of related work.

Work	Multiple Models	Open Data	Biomarkers	SHAP	App Visualization	ML Techniques
[Wu and Fang 2020]	√	-	√	-	-	RLR, SVM, RF
[Abujaber et al. 2024]	√	-	-	√	-	Logistic Regression
[Huang et al. 2023b]	-	-	√	√	-	Stacked Ensemble
[Zheng et al. 2022]	√	-	√	√	-	XGBoost
[Zhu et al. 2023]	√	√	√	√	-	Multiple ML models
[Huang et al. 2023a]	√	√	√	√	-	XGBoost
This Work	√	√	√	√	√	Random Forest, XGBoost

3. Background

This section introduces essential concepts to better understand this work.

3.1. Serum Biomarkers in Stroke Assessment

A modern prognostic assessment in stroke must transcend clinical and anatomical scores, incorporating biomarkers that reflect the body's pathophysiological response to cerebral insult. Certain routinely available serum biomarkers offer a window into this systemic dysregulation. This study, therefore, utilizes a comprehensive panel of ten routinely available biomarkers, selected to represent four critical pathophysiological axes: hematologic, metabolic/renal, coagulation, and electrolyte [Desai et al. 2023]. Each of these panels provides distinct information. The hematologic panel, which includes hemoglobin, hematocrit, platelet, and white blood cell counts, offers information into oxygen carrying capacity, inflammatory status, and thrombotic potential. Metabolic and renal markers [Hemmati et al. 2025] reflect the acute stress response and vital organ function, both strongly linked to patient outcomes [Zhang et al. 2023]. The coagulation panel, consisting of prothrombin time (PT) and partial thromboplastin time (PTT), is essential for assessing the underlying risk of bleeding or thrombotic events. Finally, electrolytes such as sodium and potassium are crucial for neuronal and cardiac homeostasis, with imbalances often signaling serious complications such as cerebral edema or cardiac arrhythmias [Pelouto et al. 2024]. The integrated analysis of these ten markers across four distinct physiological systems provides a comprehensive and robust overview of the patient's condition, forming a solid foundation for a data-driven prognostic model.

3.2. The Random Forest Algorithm

Random Forest is a robust and widely used ensemble learning method that operates on the principle of bootstrap aggregating (bagging) to generate a final prediction with high accuracy and stability [Breiman 2001]. Its architecture consists of constructing a "forest" composed of hundreds or thousands of decision trees. The process begins with the creation of multiple bootstrap samples from the training dataset, which are subsets of data generated by random sampling with replacement. A decision tree is then trained on each of these samples. A crucial element that differentiates Random Forest from traditional bagging is the introduction of a second layer of randomness: at each node of the tree, when evaluating a split, only a random subset of the predictor variables is considered. This process, known as feature subspace sampling, forces diversity among the trees, thereby decorrelating them. By preventing a few highly predictive features from dominating all the trees, the model is forced to explore a broader set of predictors, thus becoming

more robust. The final prediction for a new instance is obtained by aggregating the results from all trees in the forest, through a majority vote for classification problems or by averaging for regression tasks. Its main advantages are its strong resistance to overfitting and its intrinsic ability to provide an estimate of variable importance (feature importance), a feature of great value for clinical interpretability.

3.3. The XGBoost (Extreme Gradient Boosting) Algorithm

XGBoost (Extreme Gradient Boosting) is an optimized and scalable implementation of the gradient boosting algorithm, which has established itself as one of the leading methods for tabular data [Chen and Guestrin 2016]. Unlike Random Forest, which builds trees in parallel, XGBoost employs a sequential and additive approach. The process begins with a simple model (an initial tree), and at each iteration, a new tree is trained specifically to correct the residual errors of the preceding model. The "Gradient" in its name refers to the fact that this correction is guided by a gradient descent optimization process on a predefined loss function. The "Extreme" in its name is derived from several system and algorithmic optimizations that make it superior to traditional implementations. The most important of these is the inclusion of regularization (both L1 and L2 terms) in the objective function, which penalizes model complexity and provides sophisticated control over overfitting. Furthermore, XGBoost is designed for computational efficiency, with capabilities for parallelized tree construction and optimized routines to handle missing data natively. By virtue of its precision, efficiency, and flexibility in hyperparameter tuning, XGBoost frequently achieves state-of-the-art performance in data science competitions and is a primary choice for developing high-performance predictive models in complex clinical scenarios.

3.4. Evaluation Metrics

The performance of the developed predictive models, Random Forest and XGBoost, was rigorously evaluated using three key regression metrics. Each metric provides a complementary view of the model's goodness-of-fit and error magnitude. The Coefficient of Determination (R^2, Equation 1) quantifies the proportion of the variance in the target variable that is predictable from the input features. To assess the magnitude of error, the Mean Absolute Error (MAE, Equation 2) calculates the average absolute difference, while the Root Mean Squared Error (RMSE, Equation 3) gives higher weight to larger prediction errors.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

4. Methodology

The methodology of this study was structured into four distinct, sequential stages to ensure rigor and reproducibility, from data acquisition to final risk stratification. The work was conducted as a retrospective cohort study using the publicly available Medical Information Mart for Intensive Care (MIMIC-IV) database [Johnson et al. 2021].

The first step consisted of Cohort Definition and Variable Selection. In this phase, the final cohort of 2,306 adult patients with a primary diagnosis of stroke upon ICU admission was selected. Two sets of variables were defined: (1) the 10 predictor variables, comprising demographic and clinical patient data; and (2) the 10 target variables, represented by a panel of routine laboratory biomarkers (glucose, creatinine, hemoglobin, etc.) reflecting the metabolic, renal, hematological, and coagulation axes. In the second step, Preprocessing and Feature Engineering, the raw data was prepared for modeling. This phase included handling missing values via median imputation and encoding categorical variables. The key step here was feature engineering, where the temporal nature of the laboratory data was converted into a static feature vector by computing four descriptive statistics (mean, standard deviation, minimum, and maximum) for each biomarker over the observation period.

The third step involved Model Training and Comparative Evaluation. The dataset was partitioned into training (70%), validation (10%), and test (20%) sets at the patient level to prevent data leakage. The input data were then normalized using a StandardScaler. Subsequently, the Random Forest and XGBoost models were trained and optimized for each of the ten regression tasks. The performance of both was rigorously compared on the test set using R^2, MAE, and RMSE metrics, which allowed for the identification of Random Forest as the consistently superior algorithm. Finally, in the fourth and final step, the Clinical Risk Stratification System was implemented. Using exclusively the predictions from the Random Forest model (the winner of the previous step), a three-step heuristic system was applied: (1) a predicted value was classified as an "anomaly" if it exceeded the 85th percentile of the test's normal distribution; (2) a "Physiological Instability Index" (PII) was calculated for each patient by summing their anomalies; and (3) patients were classified as "High Risk" if their PII was 5 or greater.

5. Results

This section details the findings obtained from the methodological pipeline. The analysis initially explores the intrinsic relationships among the predictor variables through a collinearity analysis (Section 5.1). Subsequently, the quantitative performance of the Random Forest and XGBoost models is presented and comparatively evaluated (Section 5.2). This is followed by a visual inspection of the top-performing model to further characterize its predictive accuracy (Section 5.3). To elucidate the model's decision-making process, a global feature importance analysis is then detailed (Section 5.4). Finally, the section concludes by presenting the interactive tool developed as a proof-of-concept for the practical application of the predictive framework (Section 5.5).

5.1. Analysis of Predictor Collinearity

A Pearson correlation analysis was conducted to quantify the linear associations among the ten predictor biomarkers in the study. The results are visualized in the correlation matrix presented in Figure 1, where the magnitude and direction of the correlation for each

variable pair are represented. The analysis revealed an expected high collinearity between Hemoglobin and Hematocrit (R = 0.97), consistent with their physiological interdependence. Correlations of moderate magnitude were observed between the coagulation parameters PT and PTT (R = 0.38), and between Creatinine and Potassium (R = 0.32). In contrast, the majority of variable pairs representing distinct pathophysiological axes, as evidenced by the relationship between Glucose and Hemoglobin (R = 0.01), demonstrated low to no linear correlation.

The presence of multicollinearity, such as that identified between Hemoglobin and Hematocrit, can impact the coefficient stability and interpretability of conventional regression models. However, the tree-based ensemble algorithms selected for this work, Random Forest and XGBoost, are inherently robust to this phenomenon. Their recursive partitioning structure and, in the case of Random Forest, the random sampling of features at each node, mitigate the adverse effects of collinearity on the model's predictive performance. Additionally, the low correlations among the other biomarkers are methodologically advantageous, suggesting that each variable contributes a largely orthogonal source of information. This indicates that the selected panel is informative and non-redundant, providing a rich basis for the machine learning models to capture complex and non-linear interactions for risk stratification in stroke.

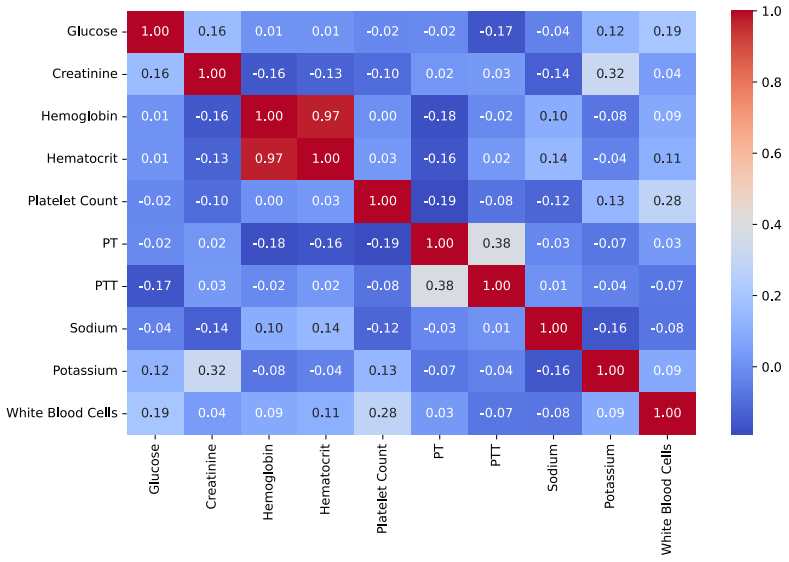

Figure 1. Collinearity Analysis among the Predictive Biomarkers.

5.2. Comparative Performance Analysis of Predictive Models

As detailed in Table 2, the comparative performance analysis between the ensemble models revealed that Random Forest consistently outperformed XGBoost in most of the prediction tasks. The Random Forest model achieved the highest Coefficients of Determination (R^2) for nine out of the ten variables analyzed. Performance was particularly strong in the prediction of hematological markers, with an R^2 of 0.930 for Hemoglobin and 0.926 for Hematocrit, indicating that the model was able to explain over 92% of the variance in this data. The prediction of PT ($R^2 = 0.901$) was also noteworthy. In contrast, the performance of Random Forest was relatively lower, although still robust, for the Potassium ($R^2 = 0.842$) and White Blood Cells ($R^2 = 0.843$) variables.

The XGBoost model also demonstrated high predictive capability, achieving competitive R^2 values greater than 0.89 for several variables, such as PT, Sodium, and Creatinine. For the PTT variable, XGBoost obtained a performance identical to that of Random Forest ($R^2 = 0.896$), suggesting that for this specific task, both models have similar efficacy. However, in contrast to the stability of Random Forest, XGBoost exhibited a sharp performance drop in the prediction of White Blood Cells, with an R^2 of only 0.688, which is considerably lower than the 0.843 achieved by Random Forest for the same variable. Regarding the error metrics (MAE and RMSE), both models exhibited similar magnitudes for most variables, indicating that while Random Forest better explains the data's variance (higher R^2), the average absolute error of the predictions was comparable between the two algorithms.

Table 2. Comparative Performance of Random Forest and XGBoost.

Clinical Variables	Random Forest			XGBoost		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Glucose	**0.870**	**9.762**	**29.55**	**0.836**	**10.50**	**33.16**
Creatinine	**0.899**	**0.140**	**0.600**	**0.892**	**0.145**	**0.619**
Hemoglobin	**0.930**	**0.265**	**0.627**	**0.925**	**0.277**	**0.648**
Hematocrit	**0.926**	**0.773**	**1.828**	**0.922**	**0.831**	**1.878**
Platelet Count	**0.893**	**13.94**	**37.95**	**0.889**	**14.35**	**38.63**
PT	**0.901**	**0.900**	**3.608**	**0.896**	**0.941**	**3.699**
PTT	**0.896**	**1.355**	**5.093**	**0.896**	**1.442**	**5.087**
Sodium	**0.894**	**0.693**	**1.721**	**0.891**	**0.719**	**1.745**
Potassium	**0.842**	**0.140**	**0.362**	**0.837**	**0.145**	**0.368**
White Blood Cells	**0.843**	**0.863**	**4.517**	**0.688**	**0.988**	**6.359**

5.3. Visual Performance Analysis of the Top-Performing Model

This visual analysis focuses exclusively on the Random Forest model, a decision prioritized by its consistent and statistically superior performance over the XGBoost model, as established in the preceding quantitative evaluation. To complement the aggregate metrics, a qualitative assessment of the model's predictive behavior was conducted via the scatter plots presented in Figure 2. Each plot maps the model's predicted values (y-axis) against the actual values from the test set (x-axis), with the line of identity (y=x) serving as the benchmark for a perfect prediction. The distribution and density of the data points relative to this diagonal offer a granular view of the model's performance.

Overall, the plots reveal a high degree of fidelity between the predicted and actual values, visually corroborating the high Coefficients of Determination (R^2) reported previously. The model's predictive accuracy was particularly pronounced for the hematological markers. In the cases of Hemoglobin and Hematocrit, the data points form a dense, homoscedastic cluster with minimal residual variance around the line of identity, indicating consistently low prediction error. In contrast, the predictions for White Blood Cells and PTT exhibit considerably greater dispersion, with a noticeable tendency for underprediction at higher leukocyte counts. This visual pattern is consistent with the lower R^2 values observed for these variables. Furthermore, the scatter plot for Glucose reveals a distinct pattern of heteroscedasticity, wherein the variance of the residuals increases proportionally with the magnitude of the actual value. This granular visual assessment complements the global performance metrics, offering a more nuanced understanding of the model's behavior and identifying specific areas of differential performance across the range of biomarker values.

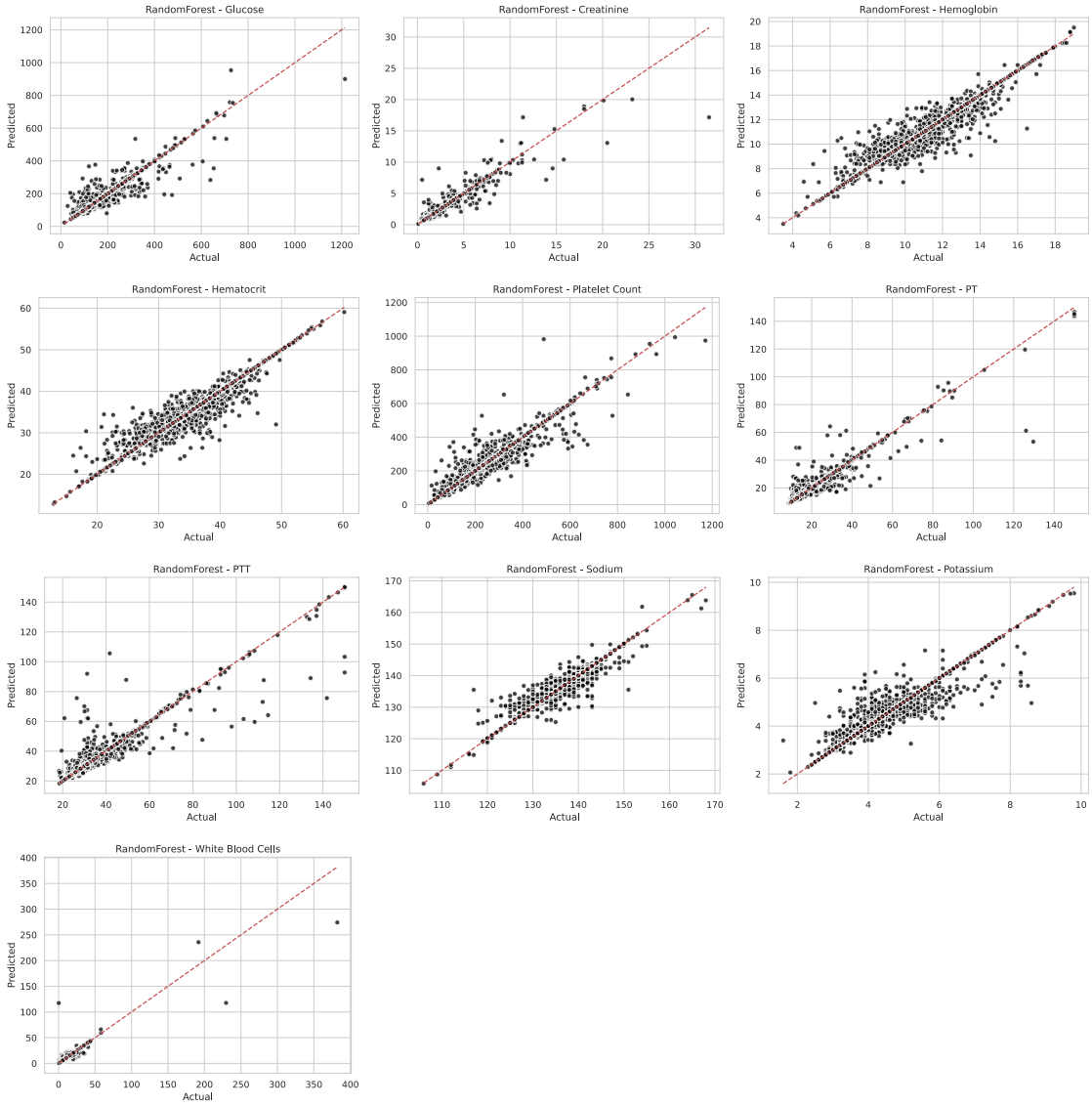

Figure 2. Random Forest Prediction Plots.

5.4. Global Feature Importance Analysis

SHAP (SHapley Additive exPlanations) is a model-agnostic, game-theoretic approach used to explain the output of any machine learning algorithm. The technique quantifies the marginal contribution of each feature to the prediction for a specific instance, ensuring a consistent and fair credit attribution. To assess the global importance and directional impact of each biomarker on the Random Forest model's predictions, a SHAP analysis was conducted. Figure 3 presents the SHAP summary plot, where features are ranked vertically by their mean absolute SHAP value.

The analysis identifies Hematocrit, Creatinine, and Glucose as the features with the highest global importance. The plot also reveals the directional effect of each feature; for example, it is observed that high values of Hematocrit (red instances) are predominantly associated with positive SHAP values, contributing to an increase in the model's output.

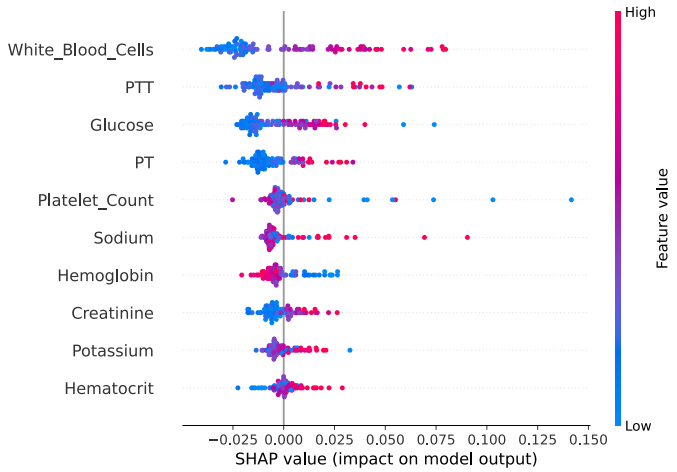

Figure 3. SHAP Feature Importance Summary Plot for the Random Forest Model.

5.5. Interactive Tool for Predictive Biomarker Analysis

To translate the modeling results into a practical solution, an interactive web application was developed, the prototype of which is presented in Figure 4, to serve as a clinical decision support system. The tool offers two complementary functionalities for patient analysis: individual biomarker trend forecasting and multisystem risk stratification. The first functionality, trend forecasting, allows a clinical user to select one of the ten analyzed biomarkers and define a time horizon. The application then uses the pre-trained model to generate and visualize a forecast of the marker's future trajectory, comparing it against the patient's historical data. This analysis provides a granular view of the likely evolution of a specific pathophysiological axis.

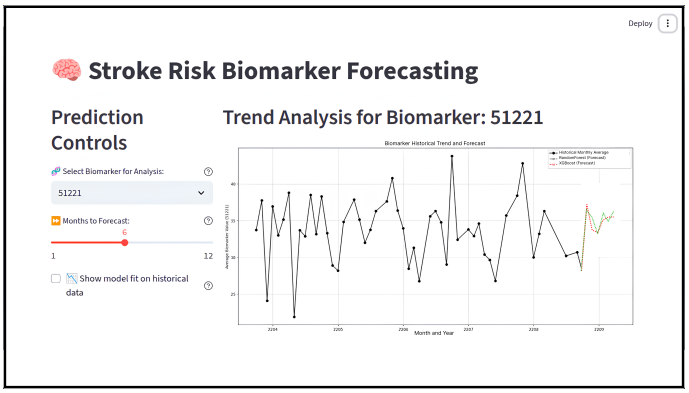

Figure 4. Clinical Decision Support Prototype.

Complementing this, the second functionality implements a holistic risk stratification system. This system utilizes the model's predictions for the entire biomarker panel through a three-step process. First, an "anomaly" is defined as a predicted value exceeding the 85th percentile of the test's normal distribution. Second, a PII is calculated for each patient by summing the total number of detected anomalies. Finally, patients are classified as "High Risk" if their PII is equal to or greater than 5. The application of this pipeline on the test set resulted in the identification of 44 patients (out of 462) who met

this criterion, demonstrating the system's ability to segment a subgroup with pronounced multisystem dysregulation. The development of this functional prototype constitutes one of the central translational contributions of this study, bridging the gap between predictive modeling and practical clinical application.

6. Conclusion

This paper presented a comparative benchmark analysis of the ML algorithms Random Forest and XGBoost for the task of predicting a panel of ten serum biomarkers in stroke patients using data from the open-source MIMIC-IV database. The results revealed a consistent performance superiority. For the prediction of most laboratory markers, the Random Forest model achieved lower error metrics and a higher Coefficient of Determination (R^2). The subsequent stratification analysis showed that this model's predictions, when translated into a Physiological Instability Index (PII), were able to effectively identify a high-risk subgroup ($n = 44$) with pronounced multisystem dysregulation, demonstrating that the accuracy of the base model is fundamental to the efficacy of derived risk systems.

This study demonstrates, therefore, that the selection of the ML algorithm is a critical factor in developing robust prognostic tools. The optimal choice lies in the balance between predictive performance and the ability to integrate predictions into a clinically interpretable framework, and the proposed methodology, which combines a regression model with the PII score, serves as a practical guide for this decision-making process. As future work, the analysis will be extended to validate the framework in Brazilian healthcare cohorts, to investigate the direct association between the PII and observed clinical outcomes, and to evolve the web application prototype into a fully integrated clinical decision, so that new studies are necessary in order to test the applicability of this model thought clinical trials.

References

Abujaber, A., Yaseen, S., Imam, Y., Nashwan, A., and Akhtar, N. (2024). Machine learning-based prediction of one-year mortality in ischemic stroke patients. *Oxford Open Neuroscience*, 3:kvae011.

Bezerra, T., Vinicius, M., Ciane, A., Callou, G., França, C., and Tavares, E. (2025). An approach based on iot and machine learning for monitoring patients on healthcare centers. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 260–271. SBC.

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

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.

Desai, A., Oh, D., Rao, E. M., Sahoo, S., Mahajan, U. V., Labak, C. M., Mauria, R., Shah, V. S., Nguyen, Q., Herring, E. Z., et al. (2023). Impact of anemia on acute ischemic stroke outcomes: a systematic review of the literature. *PLoS One*, 18(1):e0280025.

dos Santos, J. V., dos Santos Leopoldino, D. d. J., Silva, A. B. B., Lima, A. C. G., Teshima, I. E. N. S., de Oliveira Neto, E. B., Milones, M. E. d. S. V., de Santa Maria, K. C., Bomfim, L. C., de Albuquerque Maranhão, E. B., et al. (2025). Acidente vascular

cerebral no brasil: aspectos epidemiológicos da mortalidade no período de 2019 a 2023. *Brazilian Journal of Implantology and Health Sciences*, 7(3):1429–1439.

Hemmati, D., Eissazade, N., Eghdami, S., Mirzaasgari, Z., and Amouzegar, A. (2025). Three-month functional outcomes of acute ischemic stroke in patients with chronic renal function impairment. *PLoS One*, 20(5):e0323995.

Huang, J., Chen, H., Deng, J., Liu, X., Shu, T., Yin, C., Duan, M., Fu, L., Wang, K., and Zeng, S. (2023a). Interpretable machine learning for predicting 28-day all-cause in-hospital mortality for hypertensive ischemic or hemorrhagic stroke patients in the icu: a multi-center retrospective cohort study with internal and external cross-validation. *Frontiers in Neurology*, 14:1185447.

Huang, R., Liu, J., Wan, T. K., Siriwanna, D., Woo, Y. M. P., Vodencarevic, A., Wong, C. W., and Chan, K. H. K. (2023b). Stroke mortality prediction based on ensemble learning and the combination of structured and textual data. *Computers in Biology and Medicine*, 155:106176.

Jiang, Z., Wang, K., Duan, H., Du, H., Gao, S., Chen, J., and Fang, S. (2024). Association between stress hyperglycemia ratio and prognosis in acute ischemic stroke: a systematic review and meta-analysis. *BMC neurology*, 24(1):13.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark IV, R. (2021). Mimic-iv (version 1.0). physionet. 2021. *DOI: https://doi. org/10.13026/s6n6-xd98*.

Mitsios, J. P., Ekinci, E. I., Mitsios, G. P., Churilov, L., and Thijs, V. (2018). Relationship between glycated hemoglobin and stroke risk: a systematic review and meta-analysis. *Journal of the American Heart Association*, 7(11):e007858.

Pelouto, A., Reimer, J., Hoorn, E. J., Zandbergen, A. A., and den Hertog, H. M. (2024). Hyponatremia is associated with unfavorable outcomes after reperfusion treatment in acute ischemic stroke. *European Journal of Neurology*, 31(3):e16156.

Tam, C. W., Shum, H.-P., and Yan, W. (2019). Impact of dysnatremia and dyskalemia on prognosis in patients with aneurysmal subarachnoid hemorrhage: a retrospective study. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 23(12):562.

Wu, Y. and Fang, Y. (2020). Stroke prediction with machine learning methods among older chinese. *International journal of environmental research and public health*, 17(6):1828.

Zhang, H., Yue, K., Jiang, Z., Wu, X., Li, X., Luo, P., and Jiang, X. (2023). Incidence of stress-induced hyperglycemia in acute ischemic stroke: a systematic review and meta-analysis. *Brain Sciences*, 13(4):556.

Zheng, Y., Guo, Z., Zhang, Y., Shang, J., Yu, L., Fu, P., Liu, Y., Li, X., Wang, H., Ren, L., et al. (2022). Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine. *EPMA journal*, 13(2):285–298.

Zhu, E., Chen, Z., Ai, P., Wang, J., Zhu, M., Xu, Z., Liu, J., and Ai, Z. (2023). Analyzing and predicting the risk of death in stroke patients using machine learning. *Frontiers in Neurology*, 14:1096153.