Machine Learning Models for Predicting Mortality in Hemodialysis Patients

L. L. Lantmann, F. G. Gauer, I. C. Reinheimer, D. C. de Souza,
M. M. Bernardes, C. E. Poli-de-Figueiredo and S. R. Musse
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre, RS, Brazil
Emails: {lucas.langer, felipe.gattelli, isabel.reinheimer, souza.diego, marina.musse}@edu.pucrs.br,
cepolif@pucrs.br, soraia.musse@pucrs.br

***Abstract*—In Brazil, over 133,464 individuals with Chronic Kidney Disease (CKD) undergo hemodialysis, facing significant mortality risks. The mandatory biomarkers for monitoring these patients are specified by the 2014 Clinical Guideline of the Ministry of Health. Annually, more than thirty biomarkers are periodically evaluated; however, no critical evaluation of the predictive value of these biomarkers using machine learning (ML) has been conducted in Brazil to date. This paper aims to develop ML models to predict mortality outcomes in hemodialysis patients based on routine biomarkers. The goal is to investigate technologies that can assess the predictive effectiveness of clinical tests, ultimately improving patient quality of life and contributing to cost management within the Brazilian Unified Health System (SUS). This study utilizes data from hemodialysis patients in a retrospective cohort study conducted between 2012 and 2016 across 23 dialysis units in five Brazilian states. The features used in model development include biomarkers, patient profile variables, and clinical outcomes. Various ML approaches and algorithms are tested, including Decision Tree, Random Forest, Logistic Regression, XGBoost and TabPFN to identify and compare the most accurate predictive model. Among the tested models, TabPFN exhibited the best overall predictive performance, notably benefiting from balanced training data. Furthermore, the application of SHAP (SHapley Additive ExPlanations) provided clear and interpretable insights into the most influential biomarkers, which contributed to understanding the clinical plausibility of the results.**

***Index Terms*—Biomarkers, Prognostic Models, Artificial Intelligence, Renal Replacement Therapy, Chronic Kidney Disease**

I. Introduction

Chronic Kidney Disease (CKD) is a progressive condition that often requires Renal Replacement Therapy (RRT), such as peritoneal dialysis, hemodiafiltration, or hemodialysis, to sustain life [1]. Among these modalities, hemodialysis is the most prevalent, with approximately 92% of Brazilian patients undergoing this form of treatment. Despite advances in the medical and technological fields, mortality rates among hemodialysis patients remain critically high. In Brazil, the crude mortality rate for this population reached 24.5% in 2020, with certain subgroups experiencing rates as high as 10 deaths per 100 individuals [2].

The economic burden associated with hemodialysis is equally significant, with cumulative costs rising from 1.5 billion Brazilian reals in 2009 to 2.9 billion in 2018, totaling 22.4 billion over the decade [3]. This scenario underscores the urgent need for tools that can improve clinical decision making and optimize healthcare resource allocation.

Early identification of patients at heightened risk of mortality is critical to enable timely and personalized interventions that may improve survival outcomes. Machine Learning (ML) has emerged as a promising approach in this context, capable of handling complex, high-dimensional datasets and detecting hidden patterns not readily accessible through traditional statistical methods [4]. ML models have demonstrated substantial predictive power for adverse hospital outcomes, including in nephrology [5].

However, despite the growing application of ML in healthcare worldwide, there remains a lack of large-scale, validated studies evaluating the predictive value of routine biomarkers for mortality in Brazilian hemodialysis patients. In prior studies, such as the work by [6], various ML algorithms, including random forests, logistic regression, and neural networks, achieved high accuracy (ranging from 95% to 99%) in CKD diagnosis prediction. Nonetheless, to the best of our knowledge, this analysis has not been conducted using Brazilian data yet.

Given these challenges, the primary objective of this study is to develop and evaluate Machine Learning models capable of predicting mortality outcomes in hemodialysis patients based on routine clinical biomarkers. Data for this research were collected from a retrospective cohort of patients treated between 2012 and 2016 at 23 dialysis units across five Brazilian states. The specific objectives are twofold: *i)* to identify which biomarkers have the most significant impact on mortality predictions, and *ii)* to determine which ML model among those tested demonstrates the best predictive performance.

By combining clinical expertise with state-of-the-art computational methods, this study aims to contribute to the development of effective decision-support tools that can help healthcare professionals improve patient outcomes while optimizing the management of public healthcare resources.

II. Related Work

Several studies have explored the use of ML techniques for disease classification and risk prediction in patients with CKD. [6] evaluated the performance of classification algorithms

(Decision Tree, Logistic Regression, and Naive Bayes) on a dataset containing 24 biomarkers. The models achieved high accuracy levels, ranging from 95% to 99%, in predicting CKD status.

Similarly, Random Forest, Logistic Regression, and Neural Networks have been widely applied in international studies for early diagnosis and risk stratification of CKD patients, demonstrating strong predictive capabilities across various datasets [4]. These approaches have shown particular strength in handling structured clinical data, especially when large, high-dimensional datasets are involved.

Despite these promising results, most of the existing studies have focused on CKD detection rather than mortality prediction, and typically rely on datasets from outside Brazil. To date, there is a lack of large-scale, validated analyses using routine biomarkers for mortality prediction specifically in Brazilian hemodialysis patients.

III. Data Preparation

The dataset used in this study consists of monthly clinical records of hemodialysis patients. The data were collected from 23 dialysis centers located in five Brazilian states, spanning the period from 2012 to 2016. The records contained 147 parameters, including laboratory test results, demographic information, and patient outcomes. However, within the scope of this work, only 22 biomarkers, as determined by clinical guidelines, were analyzed, covering a total of 9,367 patients (DBRaw). All data were anonymized prior to the study.

To identify the biomarkers with the greatest impact and determine the most effective machine learning predictive model, a thorough understanding of the data was essential. First, we observed that the original data had missing values due to two main reasons: 1) patients who occasionally miss scheduled exams and 2) the inherent variability in the recommended frequency for certain laboratory tests, which can be monthly, quarterly, semiannual, or annual.

ML models require complete sets of variables to ensure consistent training and evaluation. Given that our clinical records are organized monthly, we adopted a systematic approach to missing data, avoiding missing data being filled with zeros, which could distort patients' true clinical status. The procedure followed two complementary approaches.

First, for laboratory tests with longer frequency (quarterly or semiannual), we used the propagation of the most recent available value, limited to a window corresponding to the Clinical Guideline. For quarterly tests, the value was propagated for a maximum of two subsequent months; for semiannual tests, for up to five months. This approach ensured that each monthly record contained plausible information, respecting the recommended time limits.

Second, we incorporated tolerance for delays in testing, reflecting the real variability in patient compliance. Missing data for up to one month was allowed for monthly tests, three months for quarterly tests, and six months for semiannual tests. Absences beyond these limits resulted in the record being classified as incomplete.

Based on the frequency of each biomarker and in consensus with the medical team, three completeness rules were defined: (i) no intervals longer than two months in the sequence of records; (ii) no gaps in monthly exams longer than two consecutive months; (iii) compliance with the quarterly, semiannual, or annual exam frequency, with a grace period of one additional period.

The definition of these rules incorporated flexibility to avoid unnecessary exclusions, given that occasional missed appointments are common. Furthermore, minimum criteria for inclusion in the analysis were established:

1) having at least three clinical records;
2) being 18 years of age or older.

This methodology ensured the integrity and clinical validity of the data, reconciling statistical rigor with the operational reality of monitoring patients on renal replacement therapy.

A. Applying the Completeness Rules

To perform the patient assignment process, we implemented the completeness rules described in the previous section using the Python programming language. The methodology for analyzing each patient is summarized below.

Figure 1 shows the flowchart used to classify patients into subsets based on criteria such as age, number of exam records, record continuity, and allowable gaps in follow-up. This flowchart details the logical steps followed to assess the quality and completeness of each patient's exam records for study inclusion.

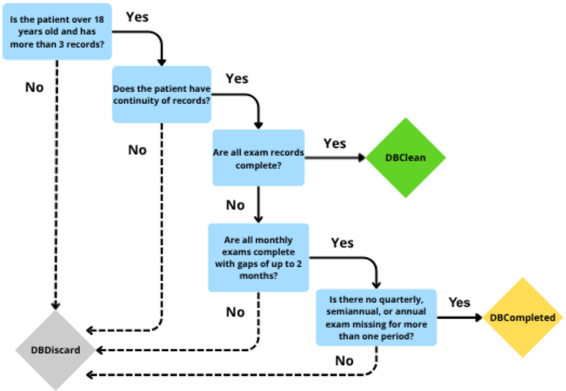
Is the patient over 18 years old and has more than 3 records?
Yes
No
Does the patient have continuity of records?
Yes
No
Are all exam records complete?
Yes
DBClean
No
Are all monthly exams complete with gaps of up to 2 months?
Yes
No
Is there no quarterly, semiannual, or annual exam missing for more than one period?
Yes
DBCompleted
No
DBDiscard

Fig. 1. Flowchart of patient assignment based on data completeness and record continuity.

Patients with fewer than three records or under 18 years of age were excluded **DBDiscard**. For the remaining patients, the continuity of the records and the completeness of the exams were verified. Completely complete records were classified as **DBClean**. In cases with absences, absence of monthly exams for up to two months was accepted; beyond that, the patient was discarded. Those who met this criterion were evaluated for quarterly, semiannual, and annual exams, accepting absence of a maximum of one additional period. If this condition was met, the patient was classified as **DBCompleted**; otherwise, they were discarded.

After preprocessing, four groups were formed:

- DBClean: complete monthly records, without the need for imputation;
- DBCompleted: records completed through clinically consistent imputation, respecting the frequency of each exam;
- DBDiscard: records unsuitable for analysis, even after considering tolerated intervals;
- DBRaw: Original records with missing data and inconsistencies, used to test model performance without filtering or imputation.

This categorization allowed us to generate databases with different levels of completeness, maintaining the rigor required for robust model development.

After running the patient assignment process, the distribution of patients across datasets is presented in Table I.

TABLE I
NUMBER OF PATIENTS IN EACH DATASET.

Dataset	Patients	Alive	Deceased
DBClean	373	354	19
DBCompleted	3,333	3,145	188
DBDiscard	5,661	4,953	708
DBRaw	9,367	8,452	915

IV. Methodology

In this section, we outline the methodology employed to develop and evaluate mortality prediction models for hemodialysis patients. First, we introduce the selected Machine Learning models. Then, we explain the partitioning scheme used to divide the data into training and testing sets. Finally, we present the metrics chosen to assess the models' performance, emphasizing their relevance in the clinical context.

To effectively represent each patient's clinical trajectory, we selected the most recent examination before the clinical outcome as input for the predictive models. This standardizes the temporal reference and ensures predictions use the latest clinical data. Records were chronologically ordered, and the last entry was extracted without restricting the time interval to the outcome, which may introduce variability in the prediction window.

A. Machine Learning Models

In this work, we employed five known algorithms, which will be described below. These models were chosen due to their proven effectiveness in handling structured, high-dimensional data—common characteristics in clinical datasets.

- Decision Tree: This algorithm uses a tree-based structure, in which internal nodes represent conditions on an attribute, branches indicate the outcome of these conditions, and leaf nodes represent the predictions (classes). Decision Trees are easy to interpret and can manage both categorical and numerical data, making them a popular choice for medical problems [7].
- Logistic Regression: Widely used for binary classification tasks, Logistic Regression models the probability of an event occurring as a logistic function of the input features. In the context of this study, it estimates the probability of death among patients based on their biomarkers. Logistic Regression is known for its simplicity and interpretability, as well as being robust for relatively small datasets or when the relationship between variables is approximately linear [8].
- Random Forest: This method consists of an ensemble of independent decision trees that vote to determine the final class of a sample. It improves accuracy compared to a single tree by reducing overfitting through multiple trees and bagging (bootstrap aggregating). Random Forest is frequently used in clinical data analysis, as it can handle unbalanced datasets and identify high-importance variables for prediction [9].
- XGBoost (Extreme Gradient Boosting): A boosting algorithm that builds sequential decision trees, where each subsequent tree corrects the errors of the previous one. XGBoost is known for its computational efficiency and strong performance in machine learning competitions, particularly when dealing with complex data and nonlinear interactions. In this study, it was selected for its ability to handle high-dimensional clinical data and its excellent predictive performance in medical contexts [10].
- TabPFN (Tabular Prior-Data Fitted Network): A model based on pre-trained *transformers*, designed for classification in tabular data. TabPFN is pre-trained on millions of synthetic tasks, learning to generalize to new datasets without the need for traditional hyperparameter tuning. At inference time, it makes predictions as if solving the problem in a *Bayesian* manner, considering a wide range of distributions generated during pre-training. This approach has demonstrated strong performance on small and medium-sized datasets, motivating its inclusion in this work. Additionally, it offers significantly reduced training time [11].

To evaluate model performance under different preprocessing and class balance scenarios, the algorithms were executed in five distinct configurations:

1) **DBRaw** (unbalanced)
2) **DBCompleted** (unbalanced)
3) **DBCompleted** (balanced)
4) **DBClean** (unbalanced)
5) **DBClean** (balanced)

These five conditions allowed for a comprehensive evaluation of how data quality and class imbalance affect model performance.

For the balanced experiments, we selected all deceased patients available in each dataset and randomly sampled the same number of survivors to ensure class parity. In the case of *DBCompleted*, where the total number of deceased patients was 188, an equal number of living patients was sampled to form a balanced subset.

B. Metrics

The algorithms were applied to the patient data in each dataset, and their performance was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. All experiments were conducted in a Python environment and run in Jupyter Notebook, providing a robust and scalable setting for model execution.

The metrics employed in this study are commonly used in ML methods and are described as follows:

- **Accuracy:** Measures the proportion of correct predictions (both positive and negative) relative to the total number of predictions made by the model. Although simple, this metric can be misleading in unbalanced datasets where one class may dominate.
- **Precision:** Measures how many of the instances predicted as positive are actually correct. In the context of this study, precision indicates the model's ability to correctly predict deaths among patients. High precision is essential to minimize false positives, which could cause unnecessary alarm for patients.
- **Recall:** Measures the proportion of actual positive instances that were correctly identified by the model. Here, recall reflects the model's ability to correctly detect patients who are genuinely at risk of death. A high recall value is crucial for reducing false negatives, which might fail to flag patients in real danger.
- **F1-Score:** Represents the harmonic mean between precision and recall, thus balancing these two metrics. F1-Score is particularly useful when there is class imbalance, such as in the case of living vs. deceased hemodialysis patients, as it provides a more robust measure of the model's effectiveness.

The use of the SHAP (SHapley Additive exPlanations) method was also essential for interpreting the results and identifying which biomarkers most strongly influenced the death predictions. By combining standard metrics with SHAP analysis, we gained a comprehensive understanding of model performance and were able to make fine-tuned adjustments to improve predictive accuracy.

V. Results

Using five different dataset configurations varying in class balance and data completeness, we evaluated the performance of several machine learning models and identified the most accurate approaches for predicting mortality in hemodialysis patients. Additionally, we applied SHAP to determine the most influential features contributing to the predictions. The tables below summarize the performance metrics for each model across all scenarios.

Several models reported zero precision, recall, and F1 on unbalanced datasets, predicting only the majority class (Tables II, III, V), highlighting the importance of class-balancing strategies for meaningful predictive performance.

A. DBRaw

Performance on the raw dataset (Table II) was heavily affected by class imbalance, particularly for minority-class recall and precision.

TABLE II
Comparison of performance metrics for Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), TabPFN (TPFN), and XGBoost (XGB) on the raw dataset containing 9,367 patients.

Metric	DT	LR	RF	TPFN	XGB
Accuracy Score	0.83	0.95	0.90	0.90	0.92
Precision	0.16	0.0	0.2	0.0	0.0
Recall	0.19	0.0	0.004	0.0	0.0
F1-Score	0.17	0.0	0.007	0.0	0.0

B. DBCompleted

1) Unbalanced: Table III presents the performance metrics for the unbalanced version of the DBCompleted dataset.

TABLE III
Comparison of performance metrics on the completed unbalanced dataset containing 3,333 patients.

Metric	DT	LR	RF	TPFN	XGB
Accuracy Score	0.88	0.94	0.94	0.94	0.94
Precision	0.09	0.0	0.0	0.0	0.33
Recall	0.12	0.0	0.0	0.0	0.03
F1-Score	0.10	0.0	0.0	0.0	0.06

2) Balanced: Table IV shows the results after balancing the classes in DBCompleted. All models exhibited improved recall and F1-score, with TabPFN and Random Forest standing out as the top performers.

TABLE IV
Comparison of performance metrics on the completed balanced dataset containing 376 patients.

Metric	DT	LR	RF	TPFN	XGB
Accuracy Score	0.59	0.66	0.71	0.75	0.63
Precision	0.55	0.63	0.69	0.71	0.58
Recall	0.59	0.65	0.69	0.77	0.69
F1-Score	0.57	0.64	0.69	0.74	0.63

C. DBClean

Note that the DBClean dataset is relatively small, so the models are likely prone to overfitting

1) Unbalanced: This version contains high-quality data, with records meeting all predefined consistency and validity rules. Despite the improved data quality, the severe class imbalance still impacted results (Table V).

2) Balanced: With both high-quality data and balanced classes, models achieved stronger overall performance, especially in the XGBoost model (Table VI).

Observing the metrics presented in Table VII, we can see that TabPFN achieved the best performance overall, followed by XGBoost.

TABLE V
COMPARISON OF PERFORMANCE METRICS ON THE CLEAN UNBALANCED DATASET CONTAINING 373 PATIENTS.

Metric	DT	LR	RF	TPFN	XGB
Accuracy Score	0.90	0.95	0.95	0.95	0.92
Precision	0.0	0.0	0.0	0.0	0.20
Recall	0.0	0.0	0.0	0.0	0.20
F1-Score	0.0	0.0	0.0	0.0	0.20

TABLE VI
COMPARISON OF PERFORMANCE METRICS ON THE CLEAN BALANCED DATASET CONTAINING 38 PATIENTS.

Metric	DT	LR	RF	TPFN	XGB
Accuracy Score	0.50	0.33	0.42	0.33	0.67
Precision	0.75	0.0	0.67	0.0	1.0
Recall	0.37	0.0	0.25	0.0	0.50
F1-Score	0.50	0.0	0.36	0.0	0.67

TABLE VII
TOP-PERFORMING MODEL IN EACH CASE.

Dataset	Model	Accu	Pre	Rec	F1
Raw un	DT	0.82	0.16	0.18	0.17
Completed un	DT	0.87	0.08	0.12	0.10
Completed	TPFN	0.75	0.71	0.77	0.74
Clean un	XGB	0.92	0.2	0.2	0.2
Clean	XGB	0.67	1.0	0.5	0.67

TabPFN achieved the highest F1 score (0.74) and recall (0.77), maintaining a solid balance with precision (0.71), which indicates its ability to correctly identify most positive cases while minimizing false positives. XGBoost also benefited from balanced datasets, improving recall from 0.20 to 0.50 and F1 from 0.20 to 0.67, highlighting the positive effect of class balancing. In contrast, tree-based models showed modest performance on unbalanced datasets, reflecting their limited ability to detect minority-class cases under severe class imbalance. Overall, these results emphasize the critical role of dataset balancing in improving predictive performance, with TabPFN consistently providing the most reliable results across metrics.

Figure 2 presents SHAP feature importance for the Completed Balanced dataset, illustrating how input variables contribute to model predictions. Due to its architecture, TabPFN does not support SHAP explainability in the same manner as the other models and is therefore excluded from this analysis.

Furthermore, we include a confusion matrix in Figure 3 that details the model's performance in terms of correct and incorrect classifications. This matrix shows how many cases were correctly identified (true positives and true negatives) and how many were misclassified (false positives and false negatives), providing a more comprehensive view of potential areas for improvement.

In summary, TabPFN stands out as the most promising model due to its superior combination of accuracy, precision, recall, and F1-Score.

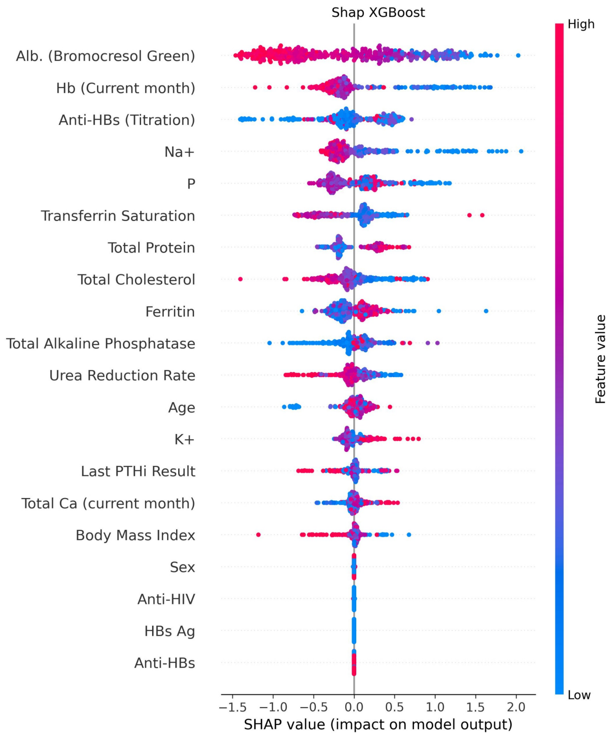

Fig. 2. SHAP chart showing the importance of features for predictions in the Completed Balanced model.

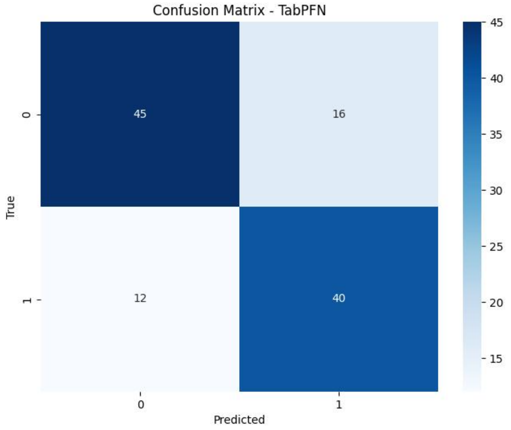

Fig. 3. Confusion matrix of the TabPFN model on the Completed Balanced dataset.

VI. FINAL CONSIDERATIONS

In this study, Machine Learning (ML) models were developed and evaluated with the goal of predicting mortality in hemodialysis patients, using routine biomarkers obtained from a large set of real clinical data encompassing 9,367 patients from 23 dialysis units across five Brazilian states. After rigorous data analysis, the XGBoost and TabPFN models stood out, with the latter showing the best performance, achieving notable values of recall (0.77), Accuracy (0.75), precision (0.71), and F1-Score (0.74).

Beyond predictive performance, the application of SHAP (SHapley Additive exPlanations) provided clear insights into the most influential biomarkers, enabling a clinically meaningful interpretation of the models' predictions. This interpretability is crucial in healthcare, as it allows clinicians to understand how predictions are influenced by key biomarkers, potentially guiding personalized interventions for high-risk patients.

These results demonstrate that advanced ML techniques can effectively support healthcare professionals in early identification of patients at high mortality risk, optimizing resource allocation and potentially improving patient outcomes. The study also emphasizes the importance of addressing class imbalance and data completeness, which are common challenges in real-world clinical datasets.

These findings highlight the value of integrating data-driven tools into the Brazilian Unified Health System (SUS), where improved risk stratification may optimize the allocation of limited healthcare resources and support more sustainable cost management.

VII. Limitations and Future Work

This study has several limitations, including reliance on a single-country dataset and the lack of external validation, limiting model generalizability. Additionally, while random undersampling was employed to address class imbalance (a simple and practical approach for exploratory analysis) it may introduce biases and potential overfitting, especially in balanced subsets.

Future work will investigate more advanced imbalance-handling techniques, including SMOTE and ADASYN, and strengthen statistical robustness through cross-validation and confidence intervals. Finally, expanding the dataset to include more centers and a larger patient population would improve generalizability and allow for a more rigorous assessment of advanced models, while preserving interpretability via SHAP.

Acknowledgment

This study was partly financed by the Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil (CNPq) and PUCRS.

Availability Statement

Due to data protection restrictions, the dataset cannot be shared. Nevertheless, we plan to make the code developed in this study publicly available in the future.

References

[1] W. G. Couser, G. Remuzzi, S. Mendis, and M. Tonelli, "The contribution of chronic kidney disease to the global burden of major noncommunicable diseases," *Kidney international*, vol. 80, no. 12, pp. 1258–1270, 2011.

[2] P. D. M. d. M. Neves, R. d. C. C. Sesso, F. S. Thomé, J. R. Lugon, and M. M. Nasicmento, "Brazilian dialysis census: analysis of data from the 2009-2018 decade," *Brazilian Journal of Nephrology*, vol. 42, pp. 191–200, 2020.

[3] "Valor apresentado de produção ambulatorial do sus para procedimento hemodiálise no período 2009-2018 no brasil." 2021.

[4] H. Habehh and S. Gohel, "Machine learning in healthcare," *Curr Genomics*, vol. 22, no. 4, pp. 291–300, Dec. 2021.

[5] V. Mahalingasivam, G. Su, M. Iwagami, M. R. Davids, J. B. Wetmore, and D. Nitsch, "COVID-19 and kidney disease: insights from epidemiology to inform clinical practice," *Nat Rev Nephrol*, vol. 18, no. 8, pp. 485–498, Apr. 2022.

[6] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," in *2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE)*. IEEE, 2017, pp. 291–296.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[8] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.

[9] L. Breiman and A. Cutler, *Random Forests*. CRC press, 2001.

[10] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[11] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, 2025.