# Building a Comprehensive Robust Framework for Predictive Machine Learning Models Development Using Real-World Clinical Data

Andrey Ermak, Denis Gavrilov, Roman Novitskiy, Alexander Gusev, Anna Andreychenko
K-SkAI LLC
Petrozavodsk, Russia
aermak, dgavrilov, roman, agusev@webiomed.ru, andreychenko.a@gmail.com

***Abstract*—This study presents a comprehensive and considerably automated framework for development, evaluation, and validation of prediction models using machine learning (ML) algorithms and real-world clinical data. Specifically, the framework was designed to predict preventable hospitalizations in patients with arterial hypertension (AH) and its complications, a critical clinical task given the significant economic and social costs associated with inpatient treatment of these patients. The field of cardiology is currently faced with the challenge of developing widely accepted prognostic scales for patients with arterial hypertension, and ML methods offer promising solutions to this issue. The framework was tested on a large dataset of 1,165,770 depersonalized electronic health records of 151,492 patients with AH, with 43 potential predictors considered. The framework includes essential steps such as preprocessing (including missing value imputation, scaling, and class imbalance correction), optimal model selection and testing, and external validation with a clear and an unified approach to selection of the best model. The XGBoost algorithm with Random Undersampling showed the best results and stability to external data with an area under the receiver operating characteristic curve (AUROC) of 0.815 (95% CI 0.797-0.835), demonstrating its potential for close monitoring of high-risk patients, early preventive interventions, and optimized medical care.**

## I. Introduction

A wide usage of predictive machine learning (ML) models in clinical practice is hindered by their acceptance and trust of the domain experts and end users that are not ML specialists. A typical process of modelling is complicated for non-technical specialists and consists of several stages [1]. These stages in turn have multiple choices and pathways (methods for data imputation, curation, algorithms, training approaches etc.). As a result, utilizing state-of-the-art ML and related techniques at each stage of the process and adhering to strict criteria for model selection requires expertise to understand fully ML methodology. Next to it, in the field of predictive modeling in medicine there are own requirements to the model development and evaluation based on biostatistics that are often not addressed by the ML community (e.g. definition of outliers, guidelines for external validation, confidence intervals). It is important to note that the interaction of domain-specific knowledge and simplified transparent tools for ML model development can be an important step towards harnessing of all the benefits of ML technology for the predictive modelling in medicine. Therefore, in this work we focused on the development of a framework to build ML models for clinical tasks based on real world data that allows for a rapid prototyping and testing novel state-of-the arts ML-related methods and at the same time includes familiar for domain experts' steps of the predictive model evaluation.

The framework was developed and tested for a clinical task to predict outcomes of complex disease like arterial hypertension (AH). Elevated blood pressure is a leading modifiable risk factor for cardiovascular-related deaths and disability worldwide [2]. However, according to epidemiological studies, less than half (46.5%) of adults with hypertension are aware of their condition, 36.9% are treated with anti-hypertensive medication, and only 13.8% have their blood pressure controlled [3], [4]. Treatment absence or nonadherence may increase rates of hospitalization due to unrealized medication benefits and a corresponding decline in health status. Thus, prevalence of AH and its complications, medical care expenses, particularly hospitalization costs, and the loss of patients' productivity result in significant economic losses and social damage.

The field of cardiology is currently faced with the challenge of developing a widely accepted prognostic scale for patients with arterial hypertension. The SCORE scale, which is commonly used, relies on systolic blood pressure as a prognostic risk factor to determine the overall 10-year risk for fatal cardiovascular events [5]. However, this tool may not be optimal for short-term decision-making. In light of the importance of this task and the rapid advancements in artificial intelligence, ML methods can be effectively applied to address this issue.

Several publications have been focused on predicting the progression of the AH with ML, utilizing different complications as the main event of interest [6]–[13]. However, these studies have important limitations that restrict their applicability in routine clinical practice. The size of the datasets used for ML modeling varied from 3,395 to 2,037,027 records, and the number of features considered for modeling varied from 8 to 555 depending on the dataset used. Although some publications described internal validation with AUROC ranged from 0.607 to 0.932 and an average value of 0.772 [6], [8], [12], [13], only one study conducted external validation demonstrating the following metrics: Accuracy of 0.744, Re-

call of 0.779, Precision of 0.644, and F1-score of 0.705 [11]; however, the model used 555 features, limiting its usability in real-world practice. Additionally, few studies addressed the challenges of class imbalance, and socio-demographic or treatment-related factors were often omitted, further reducing the generalizability and clinical relevance of these models.

In this regard, research in the field of prediction of AH and its complications is essential due to the numerous input parameters involved, lack of established models for predicting preventable hospitalizations, and the need for its external validation using real-world data. By utilizing a unified and simplified framework for model development, evaluation, and validation, the process can be significantly streamlined, bridging the gap between healthcare providers and downplaying the complexities of ML.

## II. Webiomed data

### A. Dataset formation

The dataset used for this real-world multicenter retrospective observational study encompasses anthropometric measurements, physical examination results, laboratory, instrumental, anamnestic, and socio-demographic data for patients of tertiary hospitals and primary care ambulatories across 11 regions of the Russian Federation, sourced from the Webiomed predictive analytics platform database (https://webiomed.ru/). The data base contains depersonalized patient records collected between 2000 and 2023. Patients were eligible for inclusion if their age was older than 18 and medical history contained an ICD-10 code linked to AH (I10-I15, including all subcodes). Given the retrospective and depersonalized nature of the dataset no informed consent from the patients was needed. A total of 1,165,770 records associated with 151,492 patients with AH were included in this study. The study design is presented in Fig. 1.

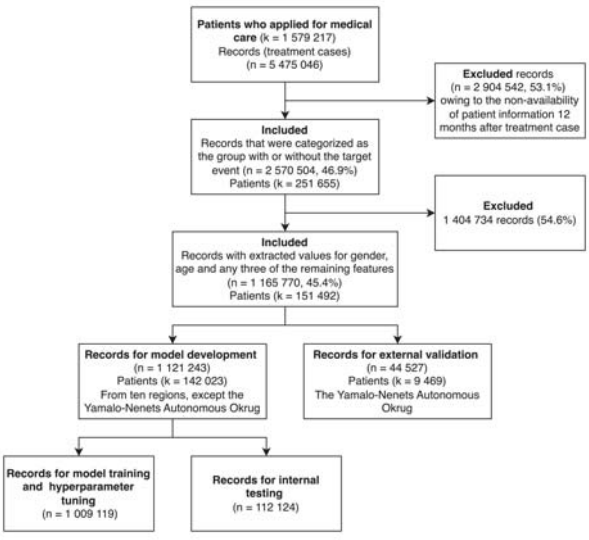


Fig. 1. Study design

For each record, the target variable was set to 1 if the patient was admitted to an inpatient treatment facility with a primary diagnosis of AH or AH-related complications within the 12 months following the given medical record. Planned or elective hospitalizations, such as routine check-ups, scheduled procedures, or follow-up visits, were excluded from the outcome definition. Therefore, the target variable reflects only unplanned or acute hospitalizations, and hospitalizations for uncomplicated, stable hypertension were not included. ICD-10 diagnosis codes used to identify AH and its complications are provided in Supplemental Table I.

Following TRIPOD guidelines [14], 44,527 records from a separate region were selected for external validation. Then, before any preprocessing steps, the remaining data for development was randomly split into a training (80%), hyperparameters tuning (10%) and internal test (10%) sets. The data splitting was performed so that class frequencies were approximately preserved in each split and all records pertaining to any given patient were in the same split, to avoid data leakage [15].

### B. Feature variable analysis

Statistical analysis and development of ML models were performed using Python programming language, version 3.9. The Kolmogorov-Smirnov test was used to evaluate the normality of the distribution of quantitative variables. Quantitative variables are presented as median with interquartile range (IQR), while categorical variables are presented as proportions (N, %). The Mann-Whitney test was used to compare quantitative variables between groups with and without the target event, and the $\chi^2$ test was used to compare categorical variables. A p-value $\leq 0.05$ was considered statistically significant.

As features for primary analysis, 61 variables were selected by a medical expert (Chief Medical Officer, Cardiologist, 20 years of clinical practice) based on their clinical relevance and frequency of occurrence. These variables were further reviewed and validated by several clinicians to ensure appropriateness for predicting preventable hospitalizations, adding an additional layer of clinical quality control.

Supplemental Table II reports the descriptive statistics for patients in the development set. Of the 1,107,672 records, 781,790 (70.6%) were associated with female patients. The median age of the cohort was 63 years old (IQR 54-70), and the observation time span ranged from 1 month to 21 years. The dataset was highly imbalanced, with only 4% of records associated with admission to an inpatient treatment facility for AH or its complications within the next 12 months. Weakness, cough, headache, dizziness, chest pain, and dyspnea were the most commonly reported complaints among the patients. The prevalent comorbidities and complications of AH included cerebrovascular disease (CVD), various types of arrhythmias, coronary artery disease (CAD), and heart failure (HF). Significant differences were observed between the two target categories in terms of clinical and demographic features, including weight, age, number of hospitalizations

and outpatient visits in the previous 12 months, glucose and cholesterol levels, blood pressure, as well as symptoms such as chest pain and shortness of breath. Notably, patients in class 1 exhibited a higher prevalence of risk factors, including family history of heart disease, obesity, and diabetes.

It should be noted that certain socio-demographic factors (e.g., income, education) and detailed medication regimen data were not consistently available across all records and were therefore not included. Despite these limitations, the dataset encompasses multicenter real-world records from 11 regions across the Russian Federation over a period of 23 years, ensuring diverse patient representation and enhancing the generalizability of the model findings.

The external validation set comprised 44,527 records from a separate region, of which 4,335 (9.7%) belonged to class 1 and 40,192 (90.3%) belonged to class 0. Notable differences were observed between the development and external validation sets in the occurrence rates of diabetes, dyslipidemia, iron deficiency anemia, CAD, renal insufficiency, atrial fibrillation, CVD, and HF. The comparison of feature distributions in both sets is provided in Supplemental Table II.

## III. Overview of our framework

The overall structure of our framework, along with the main steps, is illustrated in Fig. 2.

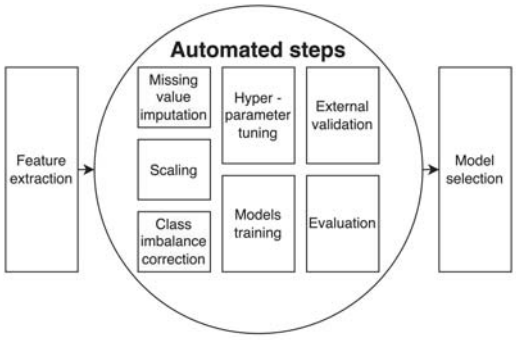


Fig. 2. Framework structure

### A. Data preprocessing

For imputation, we used a constant value of ”-10000” [16] for quantitative features, while binary features with missing values were set to False. This strategy was primarily chosen to optimize tree-based algorithms, which are robust to such values. Any quantitative feature exceeding acceptable clinical limits was removed and treated as a missing value. Alternative imputation methods (left unfilled, mean, median, KNN) were tested; however, they did not improve performance or stability of the final models. This choice is explicitly reported to clarify potential biases in non-tree models.

For feature scaling, various techniques such as robust scaling [17], standardization [18], and retaining original dimensionality were utilized. Multiple algorithms were employed and compared for class imbalance correction, including Random Undersampling (RUS), Random Oversampling (ROS), a combination of RUS and ROS (Combined Sampling), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling Approach (ADASYN) [19]. The preprocessing pipeline was iteratively performed to determine the most effective combination of missing value imputation, data scaling, and class imbalance correction before developing the ML model.

### B. Modeling

Several different families of ML algorithms were considered for primary analysis: logistic regression (LR), discriminant analysis (LinearDiscriminant, QuadraticDiscriminant), naive Bayes classifier (GaussianNB), Multi-layer Perceptron and decision tree-based ensemble methods, including gradient boosting (AdaBoost, LightGBM, XGBoost, CatBoost), and bagging (RandomForest and ExtraTrees). Hyper-parameter optimization was performed for all these algorithms for each preprocessing pipeline using the RandomGridSearch [20] optimizing for AUROC on hyperparameters tuning set. After this procedure, the models were re-trained using the selected hyperparameters and calibrated (using isotonic regression) on the training set and then evaluated on the separate test set with using a maximum of the Youden index as threshold in terms of discrimination metrics: AUROC, AUPRC, sensitivity, specificity, accuracy, balanced accuracy, geometric mean (GMC), Matthew’s correlation coefficient, positive predictive value (PPV), negative predictive value (NPV), F1 score and likelihood-ratio test [21]–[24]. The bootstrap method with 1000 re-samplings was utilized to estimate 95% confidence intervals for these metrics [25].

After this, the optimal preprocessing pipeline was chosen for each algorithm based on the highest AUROC, PPV and AUPRC on internal test with a prerequisite that a confidence interval around the difference between AUROC means on training and test sets includes zero [26]. For seven algorithms proven their stability on test data (LightGBM, XGBoost, CatBoost, RandomForest, ExtraTrees, LR and Multi-layer Perceptron) final feature selection was made based on Shapley method [27] by selecting the ones with the highest vector lengths, which accounted for 95% of the total Shapley vector’s length. During the second stage of development, only selected features and previously defined preprocessing pipelines and hyperparameters for each of the remaining algorithms were utilized for re-training and calibration. Internal test and external validation were performed for these resulting models, after that its discrimination, utility [28] and calibration [29] were also evaluated. The final model selection was based on several criteria, including the highest AUROC value during external validation and overlap of the 95% confidence intervals of this metric for the internal test and external validation sets. Based on the approach proposed in recent works with the formation of three risk groups [30]–[32], for this model we additionally calculated two activation thresholds and all metrics in the data for the internal test, depending on the target NPV - 0.999 and PPV - 0.5.

## IV. Experimental results

### A. Impact of data preprocessing

After testing various methods for processing input data and optimizing hyperparameters through multiple iterations, we identified the most effective strategies for handling missing values, scaling, and correcting class imbalance for all ten algorithms (Table I). These strategies were chosen based on their ability to achieve the highest values of AUROC, PPV, and AUPRC.

TABLE I. The most effective preprocessing strategies for algorithms

| Model | Imputation | Scaling | Resampling |
|---|---|---|---|
| MLP | Constant | Standartization | - |
| LR | Constant | Robust | ADASYN |
| CatBoost | - | - | ROS |
| XGB | - | - | RUS |
| LGMB | - | - | - |
| ET | Constant | - | - |
| RF | Constant | - | ROS |

We placed particular emphasis on evaluating the impact of class imbalance correction on discrimination metrics that prioritize the minority class, namely, GMC and PPV (Fig. 3).

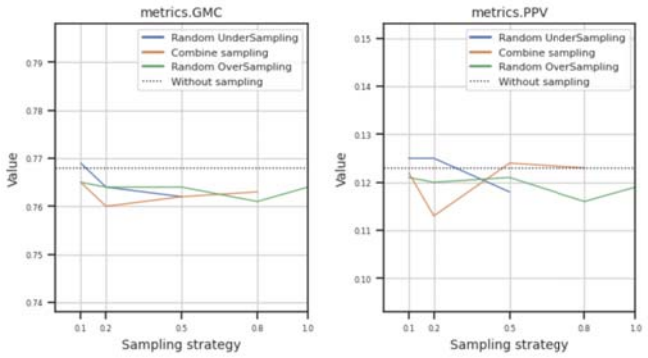


Fig. 3. Impact of class imbalance correction on metrics

However, no significant changes were observed in PPV or GMC across most imbalance correction methods. Additionally, no discernible pattern was observed for each method based on the proportion of class 1 records after correction. Nevertheless, internal testing of the XGBoost model revealed that RUS with a "sampling strategy" of 0.1 provided a slight numerical improvement in PPV (0.125 vs 0.123) and GMC (0.769 vs 0.768) compared to the uncorrected approach. More importantly, RUS consistently produced stable performance across multiple bootstrap resamples and internal/external validation sets, particularly for minority-class metrics. Based on both the observed performance and its stability, RUS with a "sampling strategy" of 0.1 was selected as the final pipeline.

### B. Model performance

The study found that the XGBoost model is effective in predicting hospitalizations for AH patients using final selected 43 features (Supplemental Table II), outperforming other algorithms with the highest AUROC values for internal test (0.849, 95% CI 0.825-0.873) and external validation (0.815, 95% CI 0.797-0.835) with the minimal difference between the calculated Youden indexes for both sets. The full range of evaluated hyperparameter settings for final model is reported in Table II.

TABLE II. Range of evaluated hyper-parameters for the final model

| Hyper-parameter | Value range |
|---|---|
| learning_rate | [1e-06, 0.5] **(0.15)** |
| n_estimators | [10, 300] **(100)** |
| subsample | [0.2, 1] **(0.7)** |
| max_depth | [1, 11] **(5)** |
| colsample_bytree | [0.5, 1] **(1)** |
| min_child_weight | [1, 4] **(3)** |
| reg_alpha | [1e-10, 10] **(2)** |
| reg_lambda | [1e-10, 10] **(0.7)** |
| scale_pos_weight | [1, 50] **(37.1)** |

During external validation, the XGBoost model was also evaluated using two additional thresholds to identify a population at increased risk (rule-in), or at decreased risk (rule-out), and to change care regimen accordingly [30]–[32]. The specificity of the model with a classification threshold of 0.001, which achieved the target NPV (0.999) on internal test, was 0.163 (95% CI 0.152-0.173), with a sensitivity of 0.996 (95% CI 0.986-1). Using the second threshold (0.265) with an expected PPV of 0.5, the quality metrics were: sensitivity – 0.083 (95% CI 0.047-0.012), and specificity – 0.996 (95% CI 0.994-0.998). The results of the final model on the separate internal test and external validation sets after applying the aforementioned thresholds and the maximum Youden index calculated on the test set are reported in Supplemental Table III. Furthermore, the decision, calibration and ROC curves for this model are reported in Figs. 4, 5, and 6, respectively.

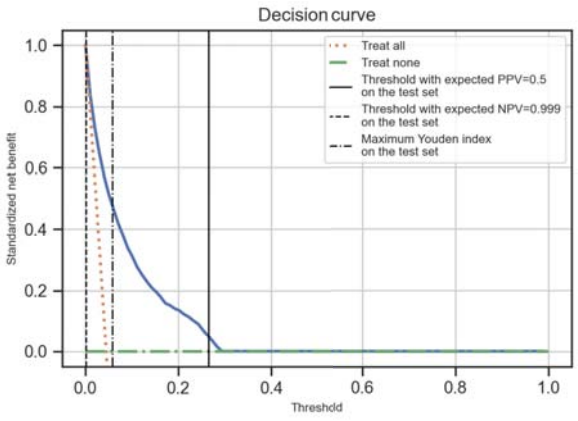


Fig. 4. The decision curve of the final model

The Shapley vectors in Fig. 7 have identified the top 10 most important features for the model, which include dyspnea, irregular heart sounds, fever and cyanosis of the skin during physical examination, left ventricular ejection fraction and

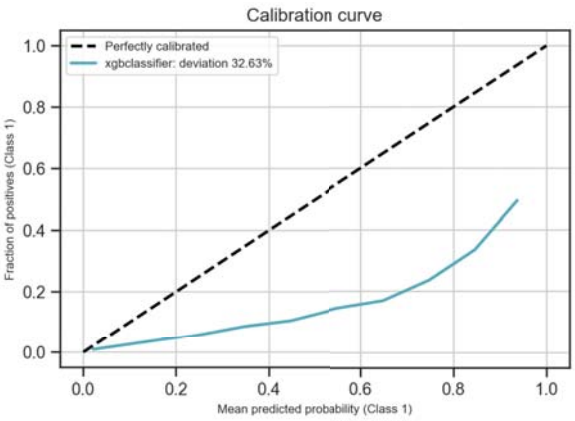


Fig. 5. The calibration curve of the final model

SBP at the time of prognosis, age, gender, hospitalizations and outpatient visits in the last 12 months. Fig. 8 displays the AUROC values obtained on both sets for all models after second stage of development. As the last stage of our study, we carried out a meta-validation of final model using a graphical representation [33] (Supplemental Fig. 1). Both data sets for internal test and external validation had sufficient sample sizes and were significantly dissimilar from the training set, as illustrated in Supplemental Fig. 1. Despite these observable differences in feature distributions, the validated model maintained strong performance across three complementary dimensions—discrimination, utility, and calibration—demonstrating robust generalizability to heterogeneous patient populations.

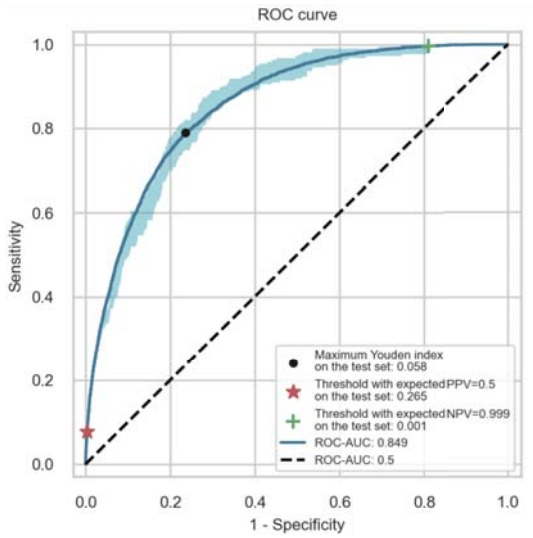


Fig. 6. The ROC curve of the final model with 95% CI during internal test

## V. Discussion

AH is a chronic medical condition characterized by elevated blood pressure levels, and it has become a major public health concern worldwide due to its increasing prevalence. In Russia, over 40% of women and nearly 50% of men aged 30-79 are affected by the disease [3]. Hypertensive patients are at a higher risk of developing various complications, which can lead to hospitalization and place a significant burden on healthcare systems. To reduce this burden, it is crucial to identify high-risk hypertensive patients and intervene with timely management strategies. Predicting the risk of hospitalization for hypertensive patients is a crucial step towards effective disease management and reducing its impact on public health.

### A. Key Advantages of the Framework

Our framework offers several notable strengths. First, the use of an external validation set, held out prior to any preprocessing, provides a rigorous assessment of model generalizability. Second, multiple strategies for class imbalance correction were systematically compared, ensuring stable performance on the minority class. Third, patient stratification into three risk groups—rule-out, intermediate, and rule-in—supports clinically relevant decision-making. Fourth, model interpretability is ensured through Shapley values (SHAP), highlighting the most influential predictors. Finally, the framework represents a complete, reproducible ML pipeline from preprocessing to calibration and external validation, adhering to TRIPOD guidelines.

### B. Limitations and comparison with other publications

After a thorough review of the available literature, it is evident that hospitalization resulting from complications of arterial hypertension has not been a primary focus of previous studies. Instead, researchers have concentrated on developing models for cardio-cerebrovascular events or kidney diseases prediction [6]–[13]. Some of these studies, just like ours, have suggested an algorithm for identifying the target event by using specific ICD-10 codes in the patient's electronic medical record and registration dates [7], [8]. However, this approach is not without limitations - it may result in the exclusion of patients who received treatment at medical facilities not included in the set and is susceptible to errors in diagnosis coding in medical practice.

Numerous studies, including our own, have observed a notable class imbalance in ML sets. To address this problem, some researchers have utilized data balancing techniques during model creation [6], [8], [10]. Nevertheless, these investigations have not conducted parallel analyses to ours, which aim to assess the impact of selecting a data balancing algorithm and its parameters on the metric values. Previous studies have identified disease duration and the use of antihypertensive medication as important predictors for effective hypertension management. Furthermore, socio-demographic factors like place of residence and patient income have been shown to impact treatment adherence and disease outcomes [7], [8]. However, our study had limitations as we did not include these socio-demographic indicators and drug therapy in our input features due to difficulties in accurately extracting dosages over a 21-year period. It is important to note, however, that drug therapy can potentially influence on the developed model.

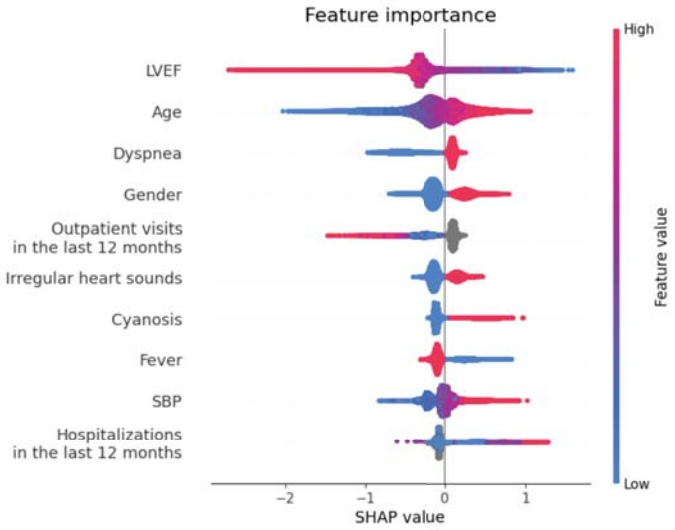


Fig. 7. The significance of the top 10 features according to the Shapley values. Grey represent missed values.

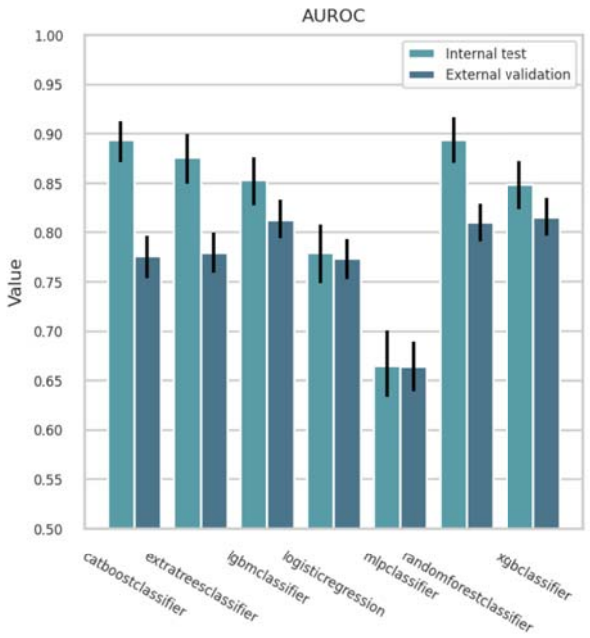


Fig. 8. The AUROC values on the internal test and the external validation sets. The black vertical lines indicate the 95% confidence intervals.

### C. Implications

Building on the outlined strengths, this study presents a comprehensive and robust framework for developing, evaluating, and validating personalized predictive models using ML algorithms and real-world clinical data to forecast preventable hospitalizations in patients with AH and its complications. Our final model has successfully completed all stages of the standard ML project lifecycle, including data collection, preprocessing (with missing value imputation, scaling, and class imbalance correction), optimal model selection and testing, feature interpretability analysis, and deployment within the Webiomed platform.

Our approach involves leveraging state-of-the-art techniques and methodologies at each stage of the modeling process. We recognize the importance of domain-specific knowledge and the need for external validation to ensure the discrimination and robustness of our model. We propose to categorize patients into three distinct risk groups for the purpose of managing AH, using probability estimates generated by the XGBoost model and calculated two thresholds [30]–[32]. The first threshold (0.001) was selected to optimize the model's sensitivity and ensure high accuracy in identifying patients at low risk of hospitalization who may not require active treatment. The second threshold (0.265), on the other hand, was chosen to maximize the model's specificity and reliably predict hospitalization for patients at high risk of adverse outcomes, thus enabling healthcare providers to implement more careful monitoring strategies.

The model exhibits strong potential as an additional monitoring tool for AH patients, with robust performance in both internal and external validation. The approach represents a significant contribution to AH management and has important implications for improving patient outcomes and reducing healthcare costs. Prospective studies can further validate the model for practical use.

### D. Future research

The issue of data imbalance in clinical practice has been widely discussed in the literature, with several proposed solutions [34]–[36]. One such approach is the use of activation thresholds with target levels for negative and positive predictive values, as employed in our study. Further research is needed to improve classification in cases that fall between these thresholds. Class imbalance correction algorithms, such as ROS, SMOTE and ADASYN, have also been suggested as a solution. However, these synthetic algorithms require the filling in of missing values before use and may generate records that do not align with clinical practice. On the other hand, ROS only duplicates existing records, failing to offer additional information about target events or improve model performance. While undersampling is a promising alternative, it comes at the cost of records loss. Therefore, further research is needed to develop effective solutions for addressing data imbalance in healthcare ML applications while minimizing the loss of data.

## VI. Conclusion

Our research highlights the potential of ML techniques in developing accurate healthcare models, particularly for predicting outcomes of complex diseases. The incorporation of routine clinical and laboratory parameters as factors in the model makes it easily applicable in clinical practice. However, further research is necessary to address the issue of data imbalance and minimize data loss while developing effective solutions. By utilizing domain-specific knowledge and simplified tools for ML model development, such as our

framework, we can improve prediction outcomes, streamline the process, and make it accessible for narrow specialists without computational expertise. Overall, our study contributes to the growing body of evidence supporting the integration of ML in healthcare decision-making, which has the potential to enhance patient outcomes and reduce healthcare costs.

## REFERENCES


[1] S. H. A. Harbi, L. N. Tidjon, and F. Khomh, "Responsible design patterns for machine learning pipelines," 2023.

[2] T. R. Frieden and M. G. Jaffe, "Saving 100 million lives by improving global treatment of hypertension and reducing cardiovascular disease risk factors." *Journal of clinical hypertension (Greenwich, Conn.)*, vol. 20, pp. 208–211, 2 2018.

[3] B. Zhou, R. M. Carrillo-Larco, G. Danaei, L. M. Riley, and P. C. J. Paciorek, "Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants." *Lancet (London, England)*, vol. 398, pp. 957–980, 9 2021.

[4] K. T. Mills, J. D. Bundy, T. N. Kelly, J. E. Reed, and P. M. Kearney, "Global disparities of hypertension prevalence and control: A systematic analysis of population-based studies from 90 countries." *Circulation*, vol. 134, pp. 441–450, 8 2016.

[5] M. F. Piepoli, A. W. Hoes, S. Agewall, C. Albus, and C. Brotons, "2016 european guidelines on cardiovascular disease prevention in clinical practice the sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts) developed with the special contribution of the european association for cardiovascular prevention rehabilitation (eacpr)," *European Journal of Preventive Cardiology*, vol. 23, pp. NP1–NP96, 7 2016.

[6] W. Lee, J. Lee, H. Lee, C. H. Jun, and I. S. Park, "Prediction of hypertension complications risk using classification techniques," *Industrial Engineering and Management Systems*, vol. 13, pp. 449–453, 2014.

[7] Y. Feng, A. A. Leung, X. Lu, Z. Liang, and H. Quan, "Personalized prediction of incident hospitalization for cardiovascular disease in patients with hypertension using machine learning." *BMC medical research methodology*, vol. 22, p. 325, 12 2022.

[8] S.-J. Lee, S.-H. Lee, H.-I. Choi, J.-Y. Lee, and Y.-W. Jeong, "Deep learning improves prediction of cardiovascular disease-related mortality and admission in patients with hypertension: Analysis of the korean national health information database." *Journal of clinical medicine*, vol. 11, 11 2022.

[9] X. Wu, X. Yuan, W. Wang, K. Liu, and Y. Qin, "Value of a machine learning approach for predicting clinical outcomes in young patients with hypertension." *Hypertension (Dallas, Tex. : 1979)*, vol. 75, pp. 1271–1278, 5 2020.

[10] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records." *BMC medical informatics and decision making*, vol. 19, p. 51, 4 2019.

[11] J. Park, J.-W. Kim, B. Ryu, E. Heo, and S. Y. Jung, "Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data." *Journal of medical Internet research*, vol. 21, p. e11757, 2 2019.

[12] R. C. Lacson, B. Baker, H. Suresh, K. Andriole, and P. Szolovits, "Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients." *Clinical kidney journal*, vol. 12, pp. 206–212, 4 2019.

[13] R. Chen, Y. Yang, F. Miao, Y. Cai, and D. Lin, "3-year risk prediction of coronary heart disease in hypertension patients: A preliminary study." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2017, pp. 1182–1185, 7 2017.

[14] K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, and P. Macaskill, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration." *Annals of internal medicine*, vol. 162, p. W1–73, 1 2015.

[16] Y. Ding and J. S. Simonoff, "An investigation of missing data methods for classification trees applied to binary response data," *Journal of Machine Learning Research*, vol. 11, pp. 131–170, 2010.

[15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, 9 2023.

[17] X. H. Cao, I. Stojkovic, and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BMC Bioinformatics*, vol. 17, 9 2016.

[18] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109924, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622009735

[19] G. Weiss, *Foundations of Imbalanced Learning*, 06 2013, pp. 13–41.

[20] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[21] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 7 2009.

[22] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:11557689

[23] R. Barandela, J. S. S. A. B, V. Garcã, and E. Rangel, "Rapid and brief communication strategies for learning in class imbalance problems," pp. 849–851, 2003. [Online]. Available: www.elsevier.com/locate/patcog

[24] D. Chicco and G. Jurman, "The matthews correlation coefficient (mcc) should replace the roc-auc as the standard metric for assessing binary classification." *BioData mining*, vol. 16, p. 4, 2 2023.

[25] A. M. Zoubir and D. R. Iskander, "Bootstrap methods and applications : A tutorial for the signal processing practitioner," *IEEE - Signal Processing Magazine*, vol. 24, pp. 10–19, 2007.

[26] D. B. Wester, "Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis," *Biometrics Biostatistics International Journal*, vol. 7, pp. 73–85, 2 2018.

[27] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, and J. M. Prutkin, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9

[28] A. Campagner, F. Sternini, and F. Cabitza, "Decisions are not all equal—introducing a utility metric based on case-wise raters' perceptions," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106930, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722003121

[29] B. V. Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the achilles heel of predictive analytics." *BMC medicine*, vol. 17, p. 230, 12 2019.

[30] B. G. Fischer and A. T. Evans, "Sppin and snnout are not enough. it's time to fully embrace likelihood ratios and probabilistic reasoning to achieve diagnostic excellence," *Journal of General Internal Medicine*, vol. 38, pp. 2202–2204, 2023. [Online]. Available: https://doi.org/10.1007/s11606-023-08177-5

[31] G. Thomas, L. C. Kenny, P. N. Baker, and R. Tuytten, "A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests," *Diagnostic and Prognostic Research*, vol. 1, 12 2017.

[32] A. Baduashvili, G. Guyatt, and A. T. Evans, "Roc anatomy—getting the most out of your diagnostic test," *Journal of General Internal Medicine*, vol. 34, pp. 1892–1898, 9 2019.

[33] F. Cabitza, A. Campagner, F. Soares, L. G. de Guadiana-Romualdo, and F. Challa, "The importance of being external. methodological insights for the external validation of machine learning models in medicine," *Computer Methods and Programs in Biomedicine*, vol. 208, 9 2021.

[34] M. Kim and K.-B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data." *PloS one*, vol. 17, p. e0271260, 2022.

[35] J. Albuquerque, A. M. Medeiros, A. C. Alves, M. Bourbon, and M. Antunes, "Comparative study on the performance of different classification algorithms, combined with pre- and post-processing techniques to handle imbalanced data, in the diagnosis of adult patients with familial hypercholesterolemia." *PloS one*, vol. 17, p. e0269713, 2022.

[36] K. Fujiwara, Y. Huang, K. Hori, K. Nishioji, and M. Kobayashi, "Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis." *Frontiers in public health*, vol. 8, p. 178, 2020.