

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Proceedings of the Eighth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’24), Volume 2	
Series Title		
Chapter Title	Decision Tree Modification Based on Multi-level Data Categorization	
Copyright Year	2024	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Author	Family Name	Shakhgeldyan
	Particle	
	Given Name	Karina I.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Division	
	Organization	Vladivostok State University
	Address	Vladivostok, Russia
	Email	
	ORCID	http://orcid.org/0000-0002-4539-685X
Corresponding Author	Family Name	Kuksin
	Particle	
	Given Name	Nikita S.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Email	kuksin.ns@dvfu.ru
	ORCID	http://orcid.org/0009-0005-9106-0117
Author	Family Name	Domzhalov
	Particle	
	Given Name	Igor G.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Email	

	ORCID	http://orcid.org/0000-0002-6722-2535
Author	Family Name	Pak
	Particle	
	Given Name	Regina L.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Email	
	ORCID	http://orcid.org/0009-0004-3745-5399
Author	Family Name	Geltser
	Particle	
	Given Name	Boris I.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Division	
	Organization	Vladivostok State University
	Address	Vladivostok, Russia
	Email	
	ORCID	http://orcid.org/0000-0002-9250-557X
Author	Family Name	Rublev
	Particle	
	Given Name	Vladislav Yu.
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Far Eastern Federal University
	Address	Vladivostok, Russia
	Division	
	Organization	Vladivostok State University
	Address	Vladivostok, Russia
	Email	
	ORCID	http://orcid.org/0000-0001-7620-4454

Abstract The study is aimed on modifying a decision tree (DT) for predicting adverse events in clinical medicine by including risk factors (RF) in its structure, identified using multi-level categorization of predictors. A retrospective cohort study was conducted using data from 4,673 electronic medical records of patients with a diagnosis of ST-segment elevation myocardial infarction (STEMI) who underwent percutaneous coronary intervention (PCI). Patients were divided into two groups; the first group consisted of 318 (6.8%) patients who died in the hospital, the second group included 4,359 (93.2%) patients with a favorable

outcome of PCI. DT method and multimetric categorization of predictors were used to create prognostic models for in-hospital mortality (IHM). The performance of the models was assessed using 6 quality metrics. The study endpoint was the all-cause IHM in patients with STEMI after PCI.

A modified DT method has been developed on the basis of multi-level categorization of predictors and identification of RF for IHM. A comparative analysis of the quality of models based on the CART and modified DT algorithms showed higher performance of the second (AUC 0.813 vs 0.765, p-value = 0.003). The advantage of this method is the ability to extract production rules that ensure transparency of the generated predictive solutions.





Conclusions. A model based on a modified DT algorithm is an effective prognostic tool allowing high performance estimation of IHM probability and clinical interpretation of the prognostic results.

Keywords
(separated by '-')

decision tree - risk factors - categorizing continuous variables - Shapley additive explanation (SHAP) - explainable artificial intelligence (XAI)



Decision Tree Modification Based on Multi-level Data Categorization

Karina I. Shakhgeldyan^{1,2} , Nikita S. Kuksin¹ , Igor G. Domzhalov¹ ,
Regina L. Pak¹ , Boris I. Geltser^{1,2} , and Vladislav Yu. Rublev^{1,2} 

¹ Far Eastern Federal University, Vladivostok, Russia
kuksin.ns@dvfu.ru

² Vladivostok State University, Vladivostok, Russia

Abstract. The study is aimed on modifying a decision tree (DT) for predicting adverse events in clinical medicine by including risk factors (RF) in its structure, identified using multi-level categorization of predictors.

A retrospective cohort study was conducted using data from 4,673 electronic medical records of patients with a diagnosis of ST-segment elevation myocardial infarction (STEMI) who underwent percutaneous coronary intervention (PCI). Patients were divided into two groups; the first group consisted of 318 (6.8%) patients who died in the hospital, the second group included 4,359 (93.2%) patients with a favorable outcome of PCI. DT method and multimetric categorization of predictors were used to create prognostic models for in-hospital mortality (IHM). The performance of the models was assessed using 6 quality metrics. The study endpoint was the all-cause IHM in patients with STEMI after PCI.

A modified DT method has been developed on the basis of multi-level categorization of predictors and identification of RF for IHM. A comparative analysis of the quality of models based on the CART and modified DT algorithms showed higher performance of the second (AUC 0.813 vs 0.765, p-value = 0.003). The advantage of this method is the ability to extract production rules that ensure transparency of the generated predictive solutions.

Conclusions. A model based on a modified DT algorithm is an effective prognostic tool allowing high performance estimation of IHM probability and clinical interpretation of the prognostic results.

Keywords: decision tree · risk factors · categorizing continuous variables · Shapley additive explanation (SHAP) · explainable artificial intelligence (XAI)

1 Introduction

Ischemic heart disease (IHD) leads in the mortality structure of the population from cardiovascular diseases [14]. Among the most dangerous clinical types of IHD is acute myocardial infarction with ST segment elevation on the electrocardiogram (STEMI). One of the effective methods for treating STEMI is myocardial revascularization using percutaneous coronary intervention (PCI) [4]. Despite the improvement of PCI technologies, in-hospital mortality (IHM) after its performance for emergency indications remains high, ranging from 4 to 7%, highlighting the need for predicting adverse events [11].

Traditionally, prognostic scales are used to assess IHM risks, the performance of which is often insufficient for making necessary decisions to reduce the risks of adverse outcomes. Improving the quality of prognosis can be achieved by applying machine learning (ML) methods and developing prognostic models based on them that take into account nonlinear relationships between predictors and the endpoint. However, the widespread implementation of ML models in clinical practice is limited by the complexity of interpreting the generated conclusions. Algorithms of explainable artificial intelligence (XAI) are promising tools to solve this problem.

Decision trees (DT) are one of the popular XAI methods used to develop prognostic models [8]. The essence of the method lies in constructing an acyclic tree-like graph, and its strength is the interpretability of the conclusions generated by the model. It is important to note that DTs tend to overfit and lose the advantage of forecast interpretability as the number of predictors used increases. This is primarily because popular DT construction methods such as Classification and Regression Trees (CART) are “greedy” algorithms that form data splitting rules independently of each other [1]. Research aimed at modifying DTs addresses these issues by various methods: finding and excluding insignificant predictors [7], using the “African Buffalo Optimization” method [10], applying mixed integer linear programming methods [1], regularization, and pruning of DTs [5]. Nevertheless, the direction aimed at improving the quality and interpretability of forecasts using DTs remains relevant and is the subject of active discussion.

This study proposes a new approach to improve the performance of prognostic models based on DTs by using multi-level categorization of predictors. Its implementation involves identifying risk factors for adverse outcomes, which are subsequently used as predictors in the DT model.

The aim of the study is to modify DTs for predicting adverse events in clinical medicine by including risk factors obtained through multi-level categorization of predictors in its structure.

2 Methods

The modification of DT was based on the results of previously conducted studies in which predictors of IHM were identified and validated [13]. In the present study, the DT modification algorithm consisted of two main stages:

- 1) determination of IHM risk factors by multi-level categorization of previously selected predictors;
- 2) development of the DT model in which only the risk factors are used in data splitting.

These stages are presented in Sects. 2.1 and 2.2. The data and methods related to the validation of the new approach in the prognostic model of IHM in patients with STEMI are presented in Sects. 2.3 and 2.4.

2.1 Identification of Risk Factors

In the first stage of the study, risk factors (RF) are determined by categorizing continuous predictors of adverse events. The goal of this stage is to identify threshold values of

predictors that provide the best data separation between the IHM group and patients with favorable outcomes of cardiovascular disease. Identifying binary risk factors allows for the identification of patients at high or low risk of IHM. Unique variable values were considered as potential thresholds, and four methods were used to assess their prognostic abilities: the minimum p-value – Min(p-value), determined using the χ^2 test, the maximum area under the ROC curve (AUC) – Max(AUC) in the single-factor logistic regression model, the equidistant distance between the centroids of comparison groups [15], analysis of SHAP values obtained on the basis of the principles of operation of a single-factor stochastic gradient boosting (SGB) model [6]. It is worth noting that potential risk factors of adverse outcomes included both values greater than and less than the threshold value.

Since SHAP values reflect the degree of influence of a predictor on the decisions made by the forecasting model, analyzing the curve formed by them allows for assessing the dependence of the resulting variable on the predictor's influence. The process of analyzing such a curve (Fig. 1) is associated with the search for points of interest: inflection and intersection of the SHAP graph with the x-axis.

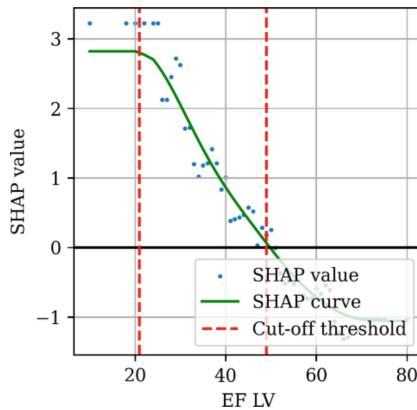


Fig. 1. Example of graph with SHAP values.

The analysis algorithm consists of the following steps: averaging SHAP values for each unique variable value; smoothing the obtained one-dimensional array using a Gaussian filter [3]; finding points of intersection of the obtained graph with the x-axis; finding inflection points of the resulting curve.

To find inflection points, the gradient was calculated for each known point on the obtained graph. Since the available data do not cover the entire possible range of predictor values, inflection points were considered not only where the gradient was equal to 0, but also the average values of neighboring points with different signs.

2.2 Development of Decision Tree Models

The process of building decision trees involves a recursive algorithm for dividing a dataset using the most optimal rules. The algorithm proposed in the study is similar to

well-established methods such as ID3, C4.5, or CART [12]. These methods split the data by selecting the best cut-off thresholds and recursively build a decision tree to maximize information gain. A modification in the developed decision tree model is that the search for thresholds during data partitioning is done not based on unique variable values, but from a list of previously identified important features.

To determine the optimal rule, the information gain criterion (entropy criterion) (1) [12] and the Gini index (2) [2] are used, which evaluate the quality of data separation considering both classes.

$$\text{Gain}(A') = - \sum_{j=1}^n P_j(A) * \log(P_j(A)) + \sum_{j=1}^n P_j(A') * \log(P_j(A')), \quad (1)$$

where P_j – probability of the i -class in the dataset; A' – subset of data obtained using the rule; A – total dataset; n – number of predicted classes.

$$\text{Gini}(A') = 1 - \sum_{j=1}^n P_j(A')^2, \quad (2)$$

where P_j – probability of i -class in dataset; A' – subset of data obtained using the rule; A – total dataset; n – number of predicted classes.

Due to the fact that in the process of building the tree, when the data is split at a node, 2 subsets are formed, it is necessary to average the quality metrics taking into account the scale of the subsets:

$$\text{Mean}(\text{Met}(A')) = \left(\text{len}(A_t) / \text{len}(A) * \text{Met}(A_t) + \text{len}(A_f) / \text{len}(A) * \text{Met}(A_f) \right) / 2, \quad (3)$$

where len – number of records in dataset; A_t – subset of data satisfying the rule; A_f – subset of data not satisfying the rule; A – total dataset; Met – quality metric for dividing the dataset.

Simplified versions of these characteristics are also proposed, which consider only one of the classes. In this case, the information gain criterion will correspond to formula (3), and the Gini index will correspond to formula (4).

$$\text{Gain}(A') = -P_j(A) * \log(P_j(A)) + P_j(A') * \log(P_j(A')), \quad (4)$$

where P_j – probability of i -class in dataset; A' – subset of data obtained using the rule; A – total dataset.

$$\text{Gini}(A') = 1 - P_j(A')^2, \quad (5)$$

where P_j – probability of i -class in dataset; A' – subset of data obtained using the rule; A – total dataset.

2.3 Machine Learning

Training, cross-validation, and final model testing were performed according to the following algorithm. The predictors used in the model were features that were selected and validated by the authors earlier in the development of the predictive model of in-hospital mortality in patients with STEMI after PCI [13]. The dataset was divided into 2 samples: for training and cross-validation (80%) and for final testing (20%). The training and cross-validation procedure was conducted using the stratified k-Folders method with 10 samples. Averaged quality metrics were used: AUC, sensitivity (Sen), specificity (Spec), F1-score, positive predictive value (PPV), negative predictive value (NPV). The AUC metric was applied for selecting the best model, feature selection, and hyperparameter tuning. The cutoff threshold for calculating Sen and Spec was determined by finding a balance between them. For the final testing, the best models with optimal parameters and hyperparameters were trained on 80% and tested on a subset for final testing. To provide a confidence estimate for the quality metrics, the procedure was repeated 50 times, initially randomizing the split.

Data analysis and model building were conducted in Python with open-source code, version 3.9.16.

2.4 Data Collecting

A single-center retrospective cohort study was conducted, during which data from the medical records of patients treated at the Regional Vascular Center of “Primorsky Krai Clinical Hospital No. 1” in Vladivostok from 2015 to 2021 were analyzed. The Ethical Committee of the School of Medicine of the Far Eastern Federal University supported the study. The study included the medical histories of 4,673 patients who underwent invasive coronary angiography with subsequent transluminal balloon angioplasty with stenting of infarct-related arteries within the first day of hospitalization. 30-day in-hospital mortality (IHM) after PCI was recorded in 318 (6.8%) patients. The results of IHM predictors analysis are presented in Table 1. In addition to demographic data (Age), clinical blood analysis parameters were analyzed: the relative number of neutrophils (NEUT), eosinophils (EOS), and the plateletcrit (PCT), as well as the levels of creatinine (Cr) and glucose (Glu) in the blood serum. Furthermore, the model included postoperative echocardiographic parameters – left ventricular ejection fraction (EF), and objective assessment results: heart rate (HR), systolic blood pressure (SBP), and Killip class of acute heart failure.

The study endpoint was the occurrence of IHM for all causes represented as a categorical binary feature (“absence” or “development”).

3 Results

3.1 Final Included Cohort

The study included 4,673 patients aged 26 to 93 years with a median age of 63 years and a 95% confidence interval [62; 63], of whom 318 (6.8%) died within 30 days after PCI. The majority (90%) of fatal outcomes occurred within the first 7 days after the operation, 6% died between days 10 and 20, and 4% between days 20 and 30.

Intergroup analysis of demographic, clinical, laboratory, and instrumental parameters demonstrated that most of them have statistically significant differences (Table 1).

Table 1. Baseline characteristics of study population

Predictor	Group 1 (n = 318)	Group 2 (n = 4357)	OR [(95% CI)]	p-value
Female, n (%)	142 (44.65)	1332 (30.5)	1.8 [1.5; 2.3]	< 0.000001
Age (y)	71 (63; 78)	62 (55; 69)	-	< 0.000001
SBP (mmHg)	110 (90; 130)	130 (120; 150)	-	< 0.000001
HR (bpm)	86 (72; 100)	72 (65; 80)	-	< 0.000001
Cr (umol/l)	130 (96; 193.3)	97 (81; 114.8)	-	< 0.000001
Killip class, n(%) III-IV	189 (59.4)	748 (17.2)	7.1 [5.6; 9]	< 0.000001
LVEF (%)	46.5 (38; 54.8)	56 (50; 61)	-	< 0.000001
NEUT (%)	81.3 (75.75; 86.5)	66.7 (59.1; 74.9)	-	< 0.0001
PCT (%)	0.22 (0.17; 0.28)	0.2 (0.16; 0.24)	-	0.0012
EOS (%)	0.1 (0.00; 0.3)	0.9 (0.3; 1.9)	-	< 0.000001
Glu (mmol/l)	7.9 (6.3; 10.31)	5.8 (5.1; 7)	-	< 0.000001

Abbreviations: LVEF – left ventricle ejection fraction; Glu – serum glucose; SBP – systolic blood pressure; HR – heart rate; Cr – serum creatinine; PCT – plateletcrit; NEUT – neutrophil count in %; EOS – eosinophil count in %.

Among the deceased, older individuals and females predominated (OR = 1.8, p-value < 0.00001). The presence of Killip class 3 and 4 was characteristic for the first group of patients (OR = 7.1), with lower values of systolic blood pressure (SBP) and left ventricular ejection fraction (LVEF), an increase in heart rate (HR), higher levels of creatinine (Cr), neutrophils (NEUT), eosinophils (ESO), and plateletcrit (PCT).

3.2 Training and Validation of Models

Based on predictors such as age, SBP, HR, Killip class, LVEF, Cr, NEUT, EOS, PCT and Glu prognostic models were developed, and the quality indicators are presented in Table 2. The “PyEntropy” and “PyGini” algorithms were developed based on decision trees using information entropy and Gini coefficient metrics, respectively. To build the remaining models, the modified decision tree method proposed by the authors of this study was applied.

The analysis showed that the model developed using the modified decision tree method, utilizing the Gain metric and considering only data from patients with favorable treatment outcomes, has the highest predictive potential (AUC – 0.813). Models that only consider data from patients with unfavorable outcomes or include information from both classes of patients have lower predictive potential (AUC – 0.799).

Table 2. Evaluation of the performance of prognostic IHM models.

Model	AUC	Se	Sp	PPV	NPV	F1
PyGini	0.765 [0.730; 0.800]	0.726 [0.696; 0.76]	0.737 [0.713; 0.76]	0.167 [0.16; 0.174]	0.726 [0.696; 0.76]	0.272 [0.263; 0.28]
PyEntropy	0.733 [0.699; 0.767]	0.700 [0.651; 0.75]	0.695 [0.655; 0.74]	0.080 [0.068; 0.09]	0.700 [0.651; 0.75]	0.144 [0.125; 0.16]
Gini total	0.741 [0.701; 0.781]	0.727 [0.695; 0.76]	0.717 [0.675; 0.76]	0.157 [0.144; 0.17]	0.727 [0.695; 0.76]	0.258 [0.244;0.27]
Gini 0	0.780 [0.754; 0.807]	0.750 [0.725; 0.78]	0.760 [0.731; 0.79]	0.107 [0.096; 0.12]	0.747 [0.722; 0.77]	0.187 [0.168; 0.20]
Gini 1	0.764 [0.737; 0.791]	0.726 [0.696; 0.76]	0.737 [0.713; 0.76]	0.167 [0.16; 0.174]	0.726 [0.696; 0.78]	0.272 [0.263; 0.28]
Gain total	0.799 [0.777; 0.820]	0.778 [0.753; 0.80]	0.756 [0.741; 0.77]	0.104 [0.09; 0.116]	0.778 [0.753; 0.80]	0.184 [0.168; 0.20]
Gain 0	0.813 [0.795; 0.831]	0.778 [0.758; 0.78]	0.806 [0.793; 0.82]	0.127 [0.12; 0.137]	0.778 [0.76; 0.797]	0.219 [0.205; 0.23]
Gain 1	0.799 [0.775;0.823]	0.778 [0.753;0.80]	0.756 [0.741;0.77]	0.104 [0.093;0.12]	0.778 [0.753;0.80]	0.184 [0.168;0.20]

Comparative analysis of statistically significant differences in the prognostic performance of models based on the modified decision tree, “PyEntropy,” and “PyGini” indicates pvalues ranging from 0.001 to 0.026 (Table 3).

Table 3. Evaluation of statistical differences in the AUC metric of the analyzed models.

	PyEntropy	PyGini	Gain total	Gain 0	Gain 1	Gini total	Gini 0	Gini 1
PyEntropy	1.000	0.6223	0.007	0.001	0.009	0.511	0.088	0.115
PyGini	-	1.000	0.026	0.003	0.022	0.293	0.237	0.293
Gain total	-	-	1.000	0.237	0.792	0.003	0.043	0.048
Gain 0	-	-	-	1.000	0.325	0.0002	0.033	0.036
Gain 1	-	-	-	-	1.000	0.003	0.043	0.047
Gini total	-	-	-	-	-	1.000	0.022	0.026
Gini 0	-	-	-	-	-	-	1.000	0.921
Gini 1	-	-	-	-	-	-	-	1.000

Models developed using the authors’ proposed algorithm but using the Gini metric do not differ from models obtained using existing solutions like PyEntropy and PyGini (p-values range from 0.088 to 0.511). The ROC curve also demonstrates the higher quality of the model developed based on the modified DT, utilizing the Gain metric and considering only data from patients with favorable outcomes of PCI.

During cross-validation process it was determined that modified DT algorithm allows to construct small sized trees compared to CART: 13 [12; 14] vs 17 [17; 18] leaves, respectively. The statistical significance of these differences is indicated by p-value < 0.000001. The smaller resulting trees size is proposed as algorithm advantage, because of production rules extraction simplification. DT cutoff thresholds were determined before their construction utilizing p-value minimization, AUC maximization, comparison groups centrolides equidistant distance calculation and based on GBS models shap-values analysis.

Model analysis (Fig. 2), developed by modified DT method, allows us to identify 6 RF production rules (Table 4).

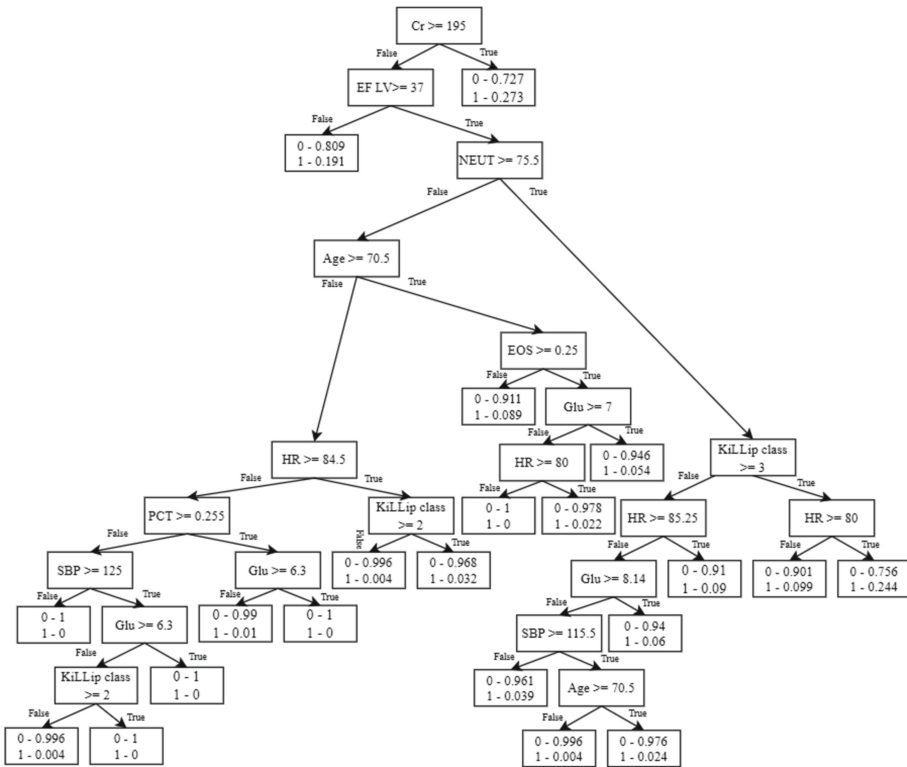


Fig. 2. Structure of decision tree

Patient classification was performed by 0.0543 threshold probability cutoff to balance sensitivity and specificity. The model is an unbalanced tree, which in some cases allow to predict IHM by partly predictors utilization. For example, if the patient’s blood creatinine level is above 195 umol/l, the model will not use other parameters.

Patient classification is performed using a probability threshold of 0.0543, providing a balance between sensitivity and specificity.

Table 4. Ranking the production rules of IHM risk factors after PCI in patients with STEMI

Nº	Production rules of risk factors	OR [95% CI]	Probability
1	Cr \geq 195 μ mol/l	13.34 [9.5; 18.6]	38.4%
2	NEUT \geq 75.5% and Glu \geq 8.15 mmol/l	10.82 [7.74; 15.12]	23.3%
3	LVEF $<$ 37%	10.75 [7.2; 16.06]	26.8%
4	NEUT \geq 75.5% and Killip \geq 3	8.48 [6.21; 11.55]	23.6%
5	NEUT \geq 75.5% and HR $>$ 82 bpm	7.61 [5.67; 10.22]	23.5%
6	EOS $<$ 0.25% and Age $>$ 70 years	5.34 [3.95; 7.22]	19.5%

4 Discussion

Despite the fact that machine learning methods demonstrate high performance in solving tasks of predicting adverse events, their application in clinical practice is currently limited. The main obstacle to their more active implementation is the opacity of ML models, and therefore, the mistrust of doctors towards the conclusions they generate. Among the methods that allow explaining the predicted probability of developing an adverse event, the most well-known is the decision tree, the main drawback of which is associated with insufficient forecast performance. Among the promising technologies of explainable artificial intelligence, the Shapley additive explanation method can be highlighted, with the help of which it becomes possible not only to assess the degree of influence of predictors on the final outcome but also to identify their threshold values that have the highest predictive value. The combination of these methods allows for the modification of decision trees, providing transparency on one hand in the decision-making process, and on the other hand, high model efficiency, which is explained by the use of knowledge extracted from clinical data in the process of identifying risk factors. The process of gaining knowledge involves searching for optimal values based on minimizing or maximizing target functions Min(p-value) and Max(AUC), as well as analyzing the shap-value of a single-factor model of stochastic gradient boosting. Assessing the dynamics of changes in shap-values allows explaining the relationship between different predictor values and the final study endpoint, which is the basis for using this method in multi-level categorization procedures. The set of risk factors thus formed, taking into account the context of the problem being addressed, allows avoiding overfitting and the algorithm thinning process of decision trees. At the same time, it is still possible to apply structural constraints to the developed model [9].

The dataset used for the validation of the modified decision tree was significantly unbalanced. As a result, the best results were achieved when using Gain index of 0, allowing for the exclusion of a large number of objects in the sample belonging to the majority class. It can be assumed that with a balanced dataset, the more effective prognostic tool will be a model based on the modified decision tree with the characteristic of "Gain total".

In the present study, the model with the best quality metrics was developed based on the modified decision tree proposed by the authors and the Gain metric, considering

only the majority class. A comparison of this model with models based on the CART method implemented in the `DecisionTreeClassifier` class in the `sklearn` module shows the superiority of the proposed solution (AUC 0.813 vs 0.765) in conditions of class imbalance.

An important advantage of the modified decision tree is the ability to interpret the forecast by extracting production rules (Table 4). This allows determining the degree of risk of in-hospital mortality and making necessary decisions to limit it. For example, with a blood Cr concentration exceeding 195 $\mu\text{mol/l}$, the odds of adverse outcomes increase by 13.3 times, which is due to progressive renal insufficiency. For patients with a NEUT blood level exceeding 75% and Glu concentration exceeding 8.15 mmol/l , the odds of in-hospital mortality increase by 10.8 times, due to pronounced inflammatory reactions and carbohydrate metabolism disorders. Other examples of production rules characterizing patients at high risk of in-hospital mortality include LVEF less than 37% and a combination of NEUT blood content over 75% with Killip class 3 or 4 indicating severe heart failure.

5 Conclusion

In the present study presents the modified DR method developed by the authors. The method is based on replacing the threshold values determined during the construction of the DR with the FR obtained by the methods of Shapley's additive explanation, minimizing p-value, maximizing AUC and calculating the equidistant distance between the centroids of features in comparison groups. The modified DR was tested on a dataset of patients with STEMI after PCI. Predictive models based on it demonstrated higher accuracy compared to the CART method (AUC 0.813 vs 0.765). The modified DR allows you to build trees of a smaller size than CART (the number of leaves is 13 vs 17), which simplifies the extraction of production rules that provide interpretation of forecast results.

Acknowledgments. This research was funded by the Ministry of Science and Higher Education of the Russian Federation (the project # FZNS-2023–0010 of the State Assignment of the Far Eastern Federal University (FEFU)).

Declaration of Competing Interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Bertsimas, D., Dunn, J., Gibson, E., et al.: Optimal survival trees. *Mach. Learn.* **111**, 2951–3023 (2022). <https://doi.org/10.1007/s10994-021-06117-0>
2. Daniya, T., Geetha, M., Suresh, K.: Classification and regression trees with Gini index. *Adv. Math. Sci. J.* **9** (2020). <https://doi.org/10.37418/amsj.9.10.53>
3. Deisenroth, M., Ohlsson, H.: A general perspective on Gaussian filtering and smoothing: explaining current and deriving new algorithms, pp. 1807–1812 (2011). <https://doi.org/10.1109/ACC.2011.5990871>

4. Ibáñez, B., James, S., Agewall, S., et al: 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Rev. Esp. Cardiol. (Engl. Ed.)* **70**(12) (2017). <https://doi.org/10.1016/j.rec.2017.11.010>
5. Lin, J., Zhong, C., Hu, D., Rudin, C., Seltzer, M.: Generalized and scalable optimal sparse decision trees. In: Proceedings of the 37th International Conference on Machine Learning, pp. 6150–6160. Vienna, Austria. (2020)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Proceedings of the 31st Annual Conference on Neural Information Processing Systems, Long Beach, USA (2017). <https://doi.org/10.48550/arXiv.1705.07874>
7. McTavish, H., Zhong, C., Achermann, R., et al: Fast sparse decision tree optimization via reference ensembles. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 9, pp. 9604–9613 (2022). <https://doi.org/10.1609/aaai.v36i9.21194>
8. Molnar, C.: Interpretable machine learning: a guide for making black box models explainable (2023). <https://christophm.github.io/interpretable-ml-book/>
9. Nanfack, G., Temple, P., Frénay, B.: Constraint enforcement on decision trees: a survey. *ACM Comput. Surv.* **54** (2022). <https://doi.org/10.1145/3506734>
10. Panhalkar, A.R., Doye, D.D.: Optimization of decision trees using modified African buffalo algorithm. *J. King Saud Univ. Comput. Inf. Sci.* (2021). <https://doi.org/10.1016/j.jksuci.2021.01.011>
11. Pfunter, A., Wier, L.M., Stocks, C.: Most frequent procedures performed in U.S. hospitals. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD): Agency for Healthcare Research and Quality (US) (2013)
12. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986). <https://doi.org/10.1007/BF00116251>
13. Shakhgeldyan, K.I., Kuksin, N.S., Domzhalov I.G., Rublev, V.Yu., Geltser, B.I.: Interpretable machine learning for in-hospital mortality risk prediction in patients with ST-elevation myocardial infarction after percutaneous coronary interventions. *Comput. Biol. Med.* **170** (2024). <https://doi.org/10.1016/j.combiomed.2024.107953>
14. The World Health Organization: The top 10 causes of death. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. Accessed 2 May 2024
15. Valente, F., Henriques, J., Paredes, S., et al: A new approach for interpretability and reliability in clinical risk prediction: acute coronary syndrome scenario. *Artif. Intell. Med.* **117** (2021). <https://doi.org/10.1016/j.artmed.2021.102113>

Author Queries

Chapter 20

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	