ISSN: 2640-4133

Research Article

Advances in Bioengineering & Biomedical Science Research

ROC-Tree Algorithm for Stratification of Binary Classifier Sets with Varied Discrimination Threshold

Y.M.Ganushchak[1*], P.J.C.Barenburg[1], J.G.Maessen[1,2,3], P. Sardari Nia[1,3]

[1]*Department of Cardiothoracic Surgery, Maastricht University Medical Center+, Maastricht, the Netherlands*

[2]*Cardiovascular Research Institute Maastricht (CARIM), Maastricht University Medical Center+, Maastricht, the Netherlands*

[3]*Heart Team Academy, Stichting Heart Team Academy, Maastricht, the Netherlands*

[*]Corresponding author
Yuri M.Ganushchak, Department of Cardiothoracic Surgery, Maastricht University Medical Center+ P. Debyelaan 25, PO Box 5800, 6202 AZ Maastricht, the Netherlands

Submitted: 26 Apr 2022; **Accepted:** 05 May 2022; **Published:** 19 May 2022

Citation: *Y.M.Ganushchak, P.J.C.Barenburg, J.G.Maessen, P. Sardari Nia, .(2022). ROC-Tree Algorithm for Stratification of Binary Classifier Sets with Varied Discrimination Threshold. Adv Bioeng Biomed Sci Res, 5(2), 113-126.*

Abstract

Binary classifier systems are used in multiple practical situations. Evaluation of diagnostic ability of a binary classifier, as its discrimination threshold is varied, often requires data transformation by performing aggregation operations. One of the most used aggregation methods is division by percentiles which divides the data set at the equal by size subgroups blindly, independently from the structure of data. We developed a ROC-tree algorithm for selection of threshold values, which is a recursive downwards splitting of each group at the two subgroups (branches) by cut-off point of ROC curve. We showed that suggested ROC-tree algorithm allows to define optimal (natural) boundaries and number of groups.

Two methods of data aggregation (percentiles and ROC-tree algorithms) were tested using the dataset 'Credit Card Fraud Detection' (https://www.kaggle.com/mlg-ulb/creditcardfraud). The results of one-vs-one reduction for the assessment of the multiclass classifications were presented as macro-average of hybrid threshold performance metrics. The macro-averages of metrics like Youden index, accuracy, optimized precision, and geometric mean were significantly different between used aggregation algorithms. The differences between macro-average of metrics ROC-tree and quartiles algorithms of stratification were preserved during 10-fold stratified cross-validation procedure.

Using algorithm sensitive to the distribution patterns, e.g., ROC-tree algorithm showed adequate stratification at groups by natural cut-off points determined by the data set composition. This method provides effective aggregation for summarizing or analyzing data in a various field of sciences. In health care described algorithm allows effective evaluation of mortality causes and quality control specialized medical care by hospitals.

Keywords: Data Aggregation Algorithm, ROC Curve, C-Statistics, Hybrid Metrics

Introduction

Binary classifier systems where its elements are classified into two groups are used in multiple practical situations. These include: medical testing or prognostic (risk prediction) models, quality control, fraud detection, and machine learning and information retrial. However, evaluation of diagnostic ability of a binary classifier as its discrimination threshold is varied often requires data transformation by performing aggregation operations. Aggregating individual observations into groups is used in a various field of sciences as a form of categorization when the discrete groups (strata) of data are created. Grouped data serves as a convenient means of summarizing or analyzing the data. Identification of discrete groups is one of the most important and difficult tasks of data mining, that is why finding a good classifier and classification algorithm is an important component of data mining.

The selection of this threshold value (possibly subjective) can have dramatic effects on model accuracy [1]. One of the most used aggregation methods is division by percentiles (quartiles as a special case of percentiles division) which divides the data set at the equal by size subgroups independently from the structure of data. We developed a ROC-tree algorithm for selection of threshold values which is recursive downwards splitting of each group at the two subgroups (branches) by cut-off point of ROC curve.

We hypothesized that opposite to the percentiles division, the ROC-tree algorithm allows to define optimal (natural) boundaries and number of groups.

Materials and Methods

The 'Credit Card Fraud Detection' dataset downloaded from https://www.kaggle.com/mlg-ulb/creditcardfraud was used for the illustration of algorithm. The datasets contain transactions made by credit cards in September 2013 by European cardholders.

As a pre-processing step we used we used ROC based feature selection to handle class imbalance classification problem. The AUC for all possible classifiers variables are presented in appendix, Table 1B.

Table 1B: Selection of variable for the classification using ROC-tree algorithm.

variable	AUC	variable	AUC
V1	0,205094	V15	0,480252
V2	0,854955	V16	0,152869
V3	0,087927	V17	0,191805
V4	0,938258	V18	0,257586
V5	0,29043	V19	0,656731
V6	0,232995	V20	0,649972
V7	0,164188	V21	0,746375
V8	0,657842	V22	0,514477
V9	0,15591	V23	0,465125
V10	0,085943	V24	0,436131
V11	0,918083	V25	0,532548
V12	0,06296	V26	0,537999
V13	0,474604	V27	0,696805
V14	0,05084	V28	0,641929

The Youden Index (Bookmaker Informedness) was used for selection of cut-off points in recursive downwards dividing subgroup into two new subgroups (branches). An area under the curve less than 0.65 in at least one subgroup of iteration was considered as an exit condition while cut-off points and number of subgroups from the previous iteration were taken for further analysis (Figure 1).

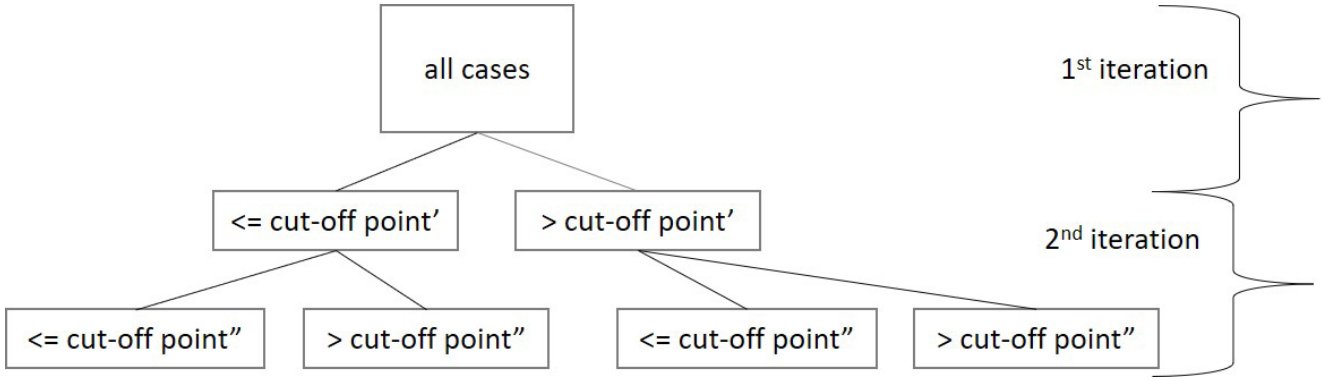

Figure 1: Flowchart shows ROC-tree algorithm of data aggregation. The division of all data at the two subgroups by Youden index followed by the second round of division at the two sequential subgroups.

The iterative usage of traditional default threshold of 0.5 as the cut-off generated four discrete groups (quartiles) with equal number of observations.

The comparison of classification algorithms was made using methods similar to the evaluation of multiclass classification. Similar to the assessment of the multiclass classification algorithms in machine learning, the one-vs-one reduction was used (Appendix A, Fig A1). Where applicable, the derivations of the 2*2 confusion matrix are presented as their macro-averages of post-hoc procedure results (one-vs-one pairwise comparison). A macro-average is the average of a metric computed independently for each class while treating all classes equally. The confusion matrix for binary classification is presented in Fig A2 (Appendix A).

The 10-fold stratified cross-validation procedure, where each fold has the same proportion of observations with the class outcome value, was used for internal validation of classification algorithm. The capabilities of algorithms were estimated as the average of performance metrics [2].

The list of variables used in the study and their equations are presented in Appendix A.

The R 3.6.3 for Windows with RStudio 1.2.5033 3 and standard packages with libraries 'lattice', 'readr' were used for the

classification of data and calculation of derivates of contingency tables for the comparison of classification algorithms.

Results

Credit Card Fraud Detection' data set presents transactions that occurred in two days, where was 492 case of frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The Imbalance ratio, which value lies within the [0, ∞] range, having a value IR = 1 in the balanced case was close to 0 (IR = 0.002) and imbalance coefficient (δ = - 0.997) with expected values within the [−1, 1] range, and 0 value for the perfectly balanced classes.

Iteration of ROC curve procedure through 28 discrimination parameters (Appendix B. Table 1B) uncover several variables with high AUC. Variable V4 (Figure 1, AUC 0,938) was used as a classifier for the ROC-tree downwards splitting. The density plot (Figure 2) shows that the distribution of cases with and without fraud by V4 parameter are different. Two-sample Kolmogorov-Smirnov test confirmed that V4 distribution is different in cases with and without fraud (D = 0.7664, p-value < 2.2e-16). However, Bhattacharyya distance for V4 case is 0.626 and Bhattacharyya coefficient (a measure of the amount of overlap between two statistical samples or populations) is 0,535.

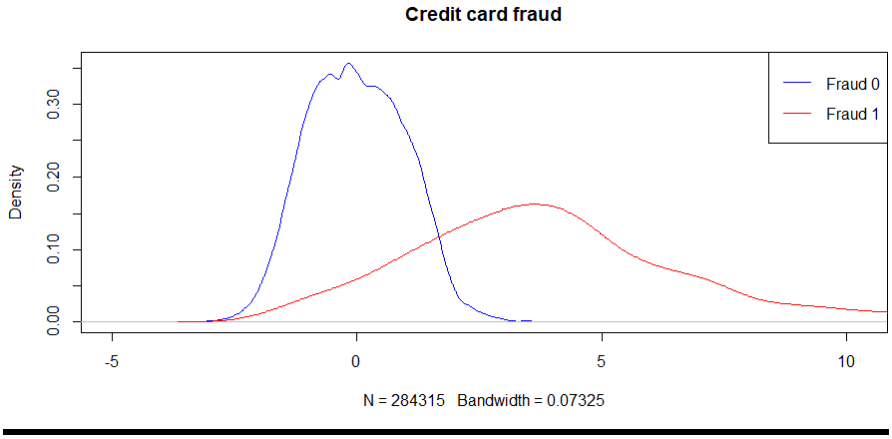

Figure 2: Kernel Density Estimation plot whole data set. V4 (mean ± std 8.32E-13 ± 1,42; minimum -5.68; maximum 16.88).

Both methods of stratification create 4 groups (Tables 1 and 2) with statistically significant differences with expected distributions. However, projection of cut-off point at the density chart (Fig.3) illustrates the fact that in case of quartiles algorithm most of fraud+ cases concentrated in group 4. ROC-tree algorithm provides more "fair" spreading cases with the highest density of fraud+ in the third group (Figure 3)

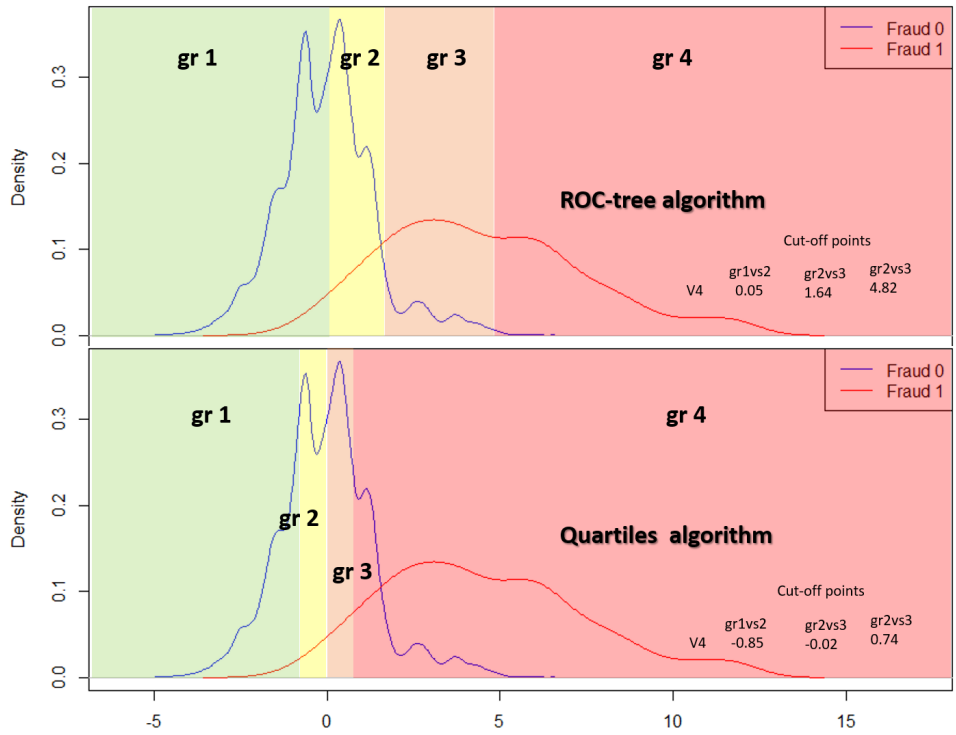

Figure 3: Cut-off points an groups areas projected at the density graph. ROC-tree algorithm (a) provides more "fair" spreading cases with the highest density of fraud+ in the third group

Table 1: Stratification at 4 groups using ROC-tree algorithm

	gr1	gr2	gr3	gr4	Total
Fraud+	14	60	204	214	492
Fraud-	148625	112038	22844	808	284315
Total	148639	112098	23048	1022	284807
X^2 (3, N= 284807) =26558, p < 0.0001					

Table 2: Stratification at 4 quartiles groups.

	gr1	gr2	gr3	gr4	Total
Fraud+	2	11	22	457	492
Fraud-	71200	71190	71180	70745	284315
Total	71202	71201	71202	71202	284807
X^2 (3, N= 284807) =1213, p < 0.0001					

Distribution of cases with and without fraud by V4 using ROC-tree algorithm or quartiles division presented at the figure 4. The results of one-vs-one reduction for the assessment of the multi-class classifications are presented as macro-average of performance metrics (Table 3).

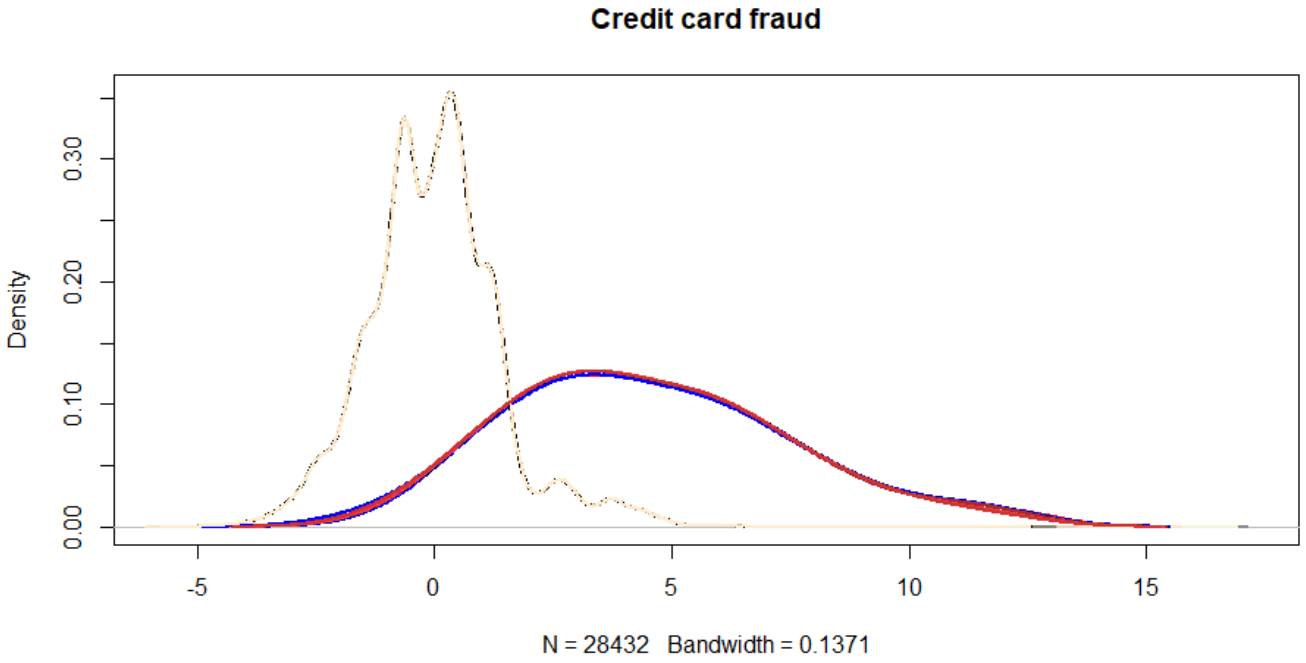

Figure 4: Density plots though 10 generated folds are similar.

Table 4: Number of cases and classifier* mean ± std per fold

	fraud -			fraud +		
fold	n	mean	std	n	mean	std
1	28432	-0,0076	1,4012	50	4,562	3,020
2	28432	-0,0077	1,4009	50	4,528	2,995
3	28432	-0,0077	1,4007	49	4,624	2,896
4	28432	-0,0078	1,4001	49	4,606	2,896
5	28432	-0,0080	1,3993	49	4,579	2,893
6	28431	-0,0078	1,3987	49	4,557	2,885
7	28431	-0,0079	1,3984	49	4,538	2,879
8	28431	-0,0080	1,3983	49	4,498	2,839
9	28431	-0,0080	1,3980	49	4,478	2,841
10	28431	-0,0081	1,3979	49	4,447	2,841
* V4 was selected as classifier						

Procedure of 10-fold cross-validation was done at the next way. Of the 10 folds, a single fold was reserved as the validation data for testing the model, and the remaining 9 subsamples were used as training sets of data. The cross-validation process is then repeated, with each of the 10 folds used exactly once as the validation data. The results from the folds were averaged to produce a single estimation.

The results of cross validation metrics included in the study are presented in the tables 2b ….16b, Appendix B. The differences between macro-average of metrics ROC-tree and quartiles algorithms of stratification were preserved during cross-validation procedure. However, the relative bias and mean square error of algorithm were statistically not different from 0 (One Sample t-test) and did not differ in ROC-tree vs Quartiles groups for all metrices included in the study. Additionally, computing confusion table metrics in control folds through cross-validation procedure of quartiles algorithm in 10% failed in comparison gr 3 vs 2, 3 vs 1 and 20% in comparison gr 2 vs 1 for GM, OP, and Youden index. This effect can be caused apparent to unsensitivity of quartiles algorithm to the distribution of fraud -/fraud + cases and concentration of most of fraud+ in fourth group (Figure.3).

Discussion

We evaluated the quality of two classification (aggregation) algorithms: ROC-tree and division at quartiles. The universal nature of the aggregation task allows to use for the demonstration of the algorithm 'Credit Card Fraud Detection' dataset downloaded from https://www.kaggle.com/mlg-ulb/creditcardfraud. This dataset contains much more cases than any available medical dataset. 'Credit Card Fraud Detection' preserves the imbalance structure inherent to the medical data. Furthermore, using dataset distant from the healthcare allows to avoid unnecessary discussion around acceptability of predictive scores (e.g. Euroscore, syntax score, CSA-AKI, Charlson comorbidity index, et cetera).

Classification methods are used in various fields of biological and medical sciences as a form of categorization when the discrete groups (strata) of data are created. Classification is one of the most important and difficult tasks of data mining, which is why finding a good classifier and classification algorithm is an important component of data mining. Classification into several tiers is the further step in the organization and understanding data. For example: division at high, medium, and low risk, based on scores of the patient cohort is an important step in the organization and understanding clinical contexts [4].

One of the most often used algorithm for division dataset into tiers is division at percentiles with creation of strata with similar number of cases or usage of early predefined cut-off points are traditional specially in medical investigations.

In the two-class classification task, the Receiver Operating Characteristic (ROC) curve is one of the most widely used tools to assess the performance of algorithms [5, 6]. The area under the receiver operating characteristic curve (AUC) (also referred to as the c statistic) is by far the most popular index of discrimination ability ROC curves have an attractive property: they are insensitive to changes in class distribution [7]. The ROC curves are independent of the proportion of positive to negative instances in a test set [8].

Several researchers have investigated the application ROC curves not only as a metrics of classification successes. Ferri et al. (2002) altered decision trees to use the AUC-ROC as their splitting criterion [9,10]. Another example of binary decision tree construction algorithm based at c-statistics is developed by Hossain et al. (2008). These authors used an AUC measure to select a node based on its classification performance and then used the misclassification rate to choose a split point [11]. In our study, we adapted the idea of ROC-tree as a form of tree which divides the classification process at a number of smaller steps which are intuitive and generally easily interpretable [12]. However, we used the Youden index (Bookmaker Informedness) for the determination of the optimal cut-off point. The misclassification rate as a complement of accuracy (one can be calculated from the other) can be misleading when the data are imbalanced, because of the dominating effect of the majority class [5, 13].

The Youden index, in contrast to the accuracy, directly includes a true positive and a true negative rate. This index is recognized as suitable performance metrics of the classification of imbalanced datasets [14].

The selection of performance metrics is another issue considered in this study. Accuracy and error rate, sensitivity and specificity are the most often used metrics for summarizing the performance of classification models. Comparing different classifiers using these measures is easy, but it has many problems such as the sensitivity to imbalanced data and ignoring the performance of some classes [13, 15-17]. Class imbalance is one of the significant issues which affect the performance of classifiers [18]. The determination of the most suitable performance metrics is a major issue in the classification of class imbalanced datasets [14]. In imbalanced datasets, not only is the class distribution skewed, the misclassification cost is often uneven too. The minority class examples are often more important than the majority class examples [5].

It is recommended to consider a combination of different measures instead of relying on only one measure when dealing with class-imbalance data [13]. Hybrid threshold metrics, such as the Geometric Mean or the Bookmaker Informedness showed to be useful as performance metrics for imbalance datasets [14, 19]. The F-measure (harmonic mean) is also recommended as the measure in this case.19 However, it still completely ignores true negatives which can vary freely without affecting the statistic [20]. The Matthews correlation coefficient (MCC) described as least influenced by imbalanced data [13].

In our study, we used hybrid measures for comparison of classification algorithms. The macro-average of the Youden index as a metric of discriminative power was significantly higher for the ROC-tree algorithm in the one-vs-one comparison (Table 3) [21]. Also, other hybrid threshold metrics such as optimized precision, geometric mean had difference with higher values of macro-averages for the ROC-tree algorithm in the one-vs-one comparison.

The "reproducibility" of cut-off points and metrices were tested by the 10-folds cross-validation which is more stable extension of split-sample validation [2, 22]. In this case cut-off points were determined in nine of the ten and testing in one of the ten, which is repeated ten times. In this way, all cases have served once to test the model. The performance is commonly estimated as the average of all assessments [2]. The cut-off points derived using the full dataset are accepted as unique and can be used for further evaluation [23].

Study limitations

Extending the number of studied datasets could increase the power of derived conclusions. The power of conclusions could also be increased by including more known confusion table derivates which could lead to the selection of most effective combination of classification performance metrics. We defined an optimal cut-off point in ROC analysis using the Youden index. However, a comparision of stabilty of cut-off points computed by other known methods could help in selecting optimal metrics for the determination of the splitting point.

The effects of sampling techniques such as down-sampling with reducing the number of samples in majority class and the assessment the differences in proportion of minority class in datasets were not evaluated in our study. However, these methods are known and recognized as effective in machine leaning fields. To some extent, the development of 'failure to rescue' as a quality indicator is an example of down-sampling in health care [24, 25].

In our study, the metrics in the one-vs-one comparison of classes were computed independently for each class and then their averages were compared. These macro-averages treated all classes equally. The combination of this approach with micro-average, which aggregates the contribution of all classes, to compute the average metric, could be effective in the evaluation of the effect of the individual classes.

Conclusion

Using algorithms sensitive to the distribution patterns, e.g. ROC-tree algorithm showed a better stratification at groups by natural cut-off points determined by the data set composition which is more convenient for summarizing or analyzing data in a various fields of sciences. In health care described algorithm allows effective evaluation of mortality causes and quality control specialized medical care by hospitals.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Kaggle repository, https://www.kaggle.com/mlg-ulb/creditcardfraud

Competing interests

The authors declare that they have no competing interests.

Funding

this work was not supported by any funding

References

1. Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological modelling, 217(1-2), 48-58.
2. Alonzo, T. A. (2009). Clinical prediction models: a practical approach to development, validation, and updating: by Ewout W. Steyerberg.
3. Team, R. C. (2018). R: A language and environment for statistical computing; 2018.
4. Wang, X., Wang, F., Hu, J., & Sorrentino, R. (2015). To-

wards actionable risk stratification: A bilinear approach. Journal of biomedical informatics, 53, 147-155.
5. Weng, C. G., & Poon, J. (2008, November). A new evaluation measure for imbalanced datasets. In Proceedings of the 7th Australasian Data Mining Conference-Volume 87 (pp. 27-32).
6. Swamidass, S. J., Azencott, C. A., Daily, K., & Baldi, P. (2010). A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics, 26(10), 1348-1356.
7. Wu, Y. C., & Lee, W. C. (2014). Alternative performance measures for prediction models. PloS one, 9(3), e91249.
8. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
9. Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning, 45(2), 171-186.
10. Ferri, C., Flach, P., & Hernández-Orallo, J. (2002, July). Learning decision trees using the area under the ROC curve. In Icml (Vol. 2, pp. 139-146).
11. Hossain, M. M., Hassan, M. R., & Bailey, J. (2008, April). ROC-tree: A novel decision tree induction algorithm based on receiver operating characteristics to classify gene expression data. In Proceedings of the 2008 SIAM International Conference on Data Mining (pp. 455-465). Society for Industrial and Applied Mathematics.
12. HAN, Jiawei, PEI, Jian, et KAMBER, Micheline. Data mining: concepts and techniques. Elsevier, 2011.
13. Akosa, J. (2017, April). Predictive accuracy: A misleading performance measure for highly imbalanced data. In Proceedings of the SAS Global Forum (Vol. 12).
14. LUQUE, Amalia, CARRASCO, Alejandro, MARTÍN, Alejandro, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 2019, vol. 91, p. 216-231.
15. THARWAT, A. Classification assessment methods. Appl Comput Inform 2018.
16. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.
17. Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4, 7940-7957.
18. POTOLEA, Rodica et LEMNARU, Camelia. A Comprehensive Study of the Effect of Class Imbalance on the Performance of Classifiers. In : ICEIS (1). 2011. p. 14-21.
19. HOSSIN, Mohammad et SULAIMAN, Md Nasir. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 2015, vol. 5, no 2, p. 1.
20. Powers D and Ailab (2011) Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol, 2, 2229-3981.
21. YOUDEN, William J. Index for rating diagnostic tests. Cancer, 1950, vol. 3, no 1, p. 32-35.
22. Berrar D .(2018). Cross-Validation. Reference Module in Life Sciences
23. FARAGGI, David et SIMON, Richard. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Statistics in medicine, 1996, vol. 15, no 20, p. 2203-2213.
24. FARJAH, Farhood, BACKHUS, Leah, CHENG, Aaron, et al. Failure to rescue and pulmonary resection for lung cancer. The Journal of Thoracic and Cardiovascular Surgery, 2015, vol. 149, no 5, p. 1365-1373. e3.
25. Johnston, M. J., Arora, S., King, D., Bouras, G., Almoudaris, A. M., Davis, R., & Darzi, A. (2015). A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. Surgery, 157(4), 752-763.

$$K = \frac{N * (N-1)}{2}$$

Where

K – number of possible combinations

N – number of groups (classes)

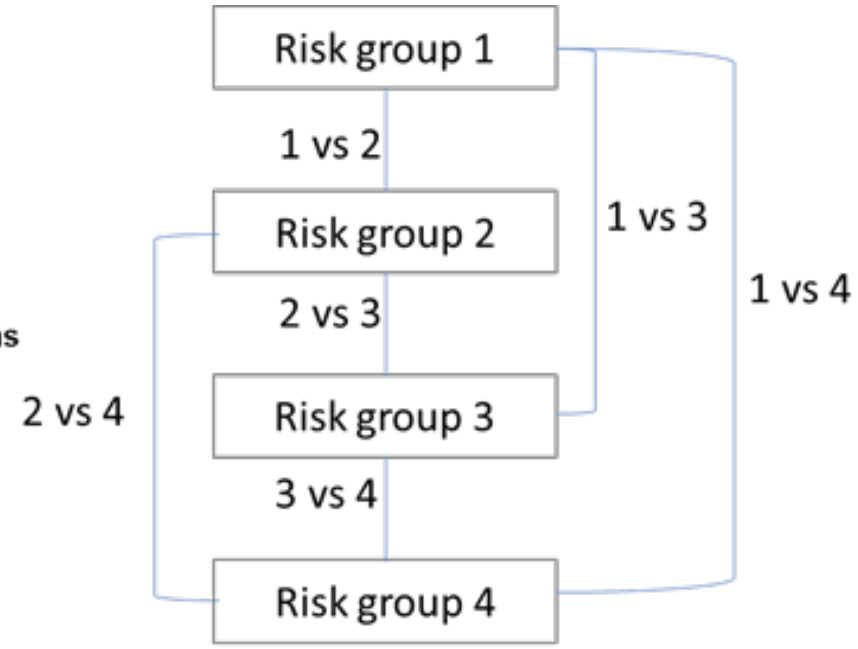

Figure A1: K one-vs-one combinations (6 in case of four groups).

		Predicted		
		P	N	
Observed	P	TP(a)	FN(b)	Op
	N	FP(c)	TN(d)	On
		Pp	Pn	N

Figure A2: Confusion matrix 2 x 2. P – positive; N – negative; TP or a - true positives; FN or b- false negatives; Op – total observed positives; FP or c – false positives; TN or d – true negatives; On – total observed negatives; PP – total predicted positives; Pn – total predicted negatives; N total number of observations.

Imbalance indices 1-3 Equation 1 and 2. The imbalance ratio ('skew') lies within the [0, ∞] and has value 1 in the balanced case.

$$IR = \frac{Op}{On} \quad (1)$$

where
IR – imbalance ratio;
Op – total observed positives;
On – total observed negatives.

Imbalance coefficient with a value [-1, 1] and 0 for the balanced data.

$$\delta = 2 * \frac{Op}{N} - 1 \quad (2)$$

where
δ – imbalance coefficient;
Op – total observed positives;
N – total number of observations.

In epidemiology, prevalence is the proportion of a population with a disease. In our study, prevalence was the proportion of deceased or mortality rate:

$$Prevalence = Op/N \quad (3)$$

where
Op – total observed positives;
N – total number of observations.
The area under the ROC curve (AUC) is a widely used measure of performance of classification algorithms 4. The AUC for each combination in one-vs-one pairwise comparisons was computed by the trapezoid rule and presented as a single scalar value together with its standard error. As an additional metric, the AUC was weighted by the prevalence of a class in the data set. In analogy to Provost and Domingos, an index equal to the one-vs-one AUCs weighted by the prevalence was computed.5, 6 However, the total weighted AUC was presented as the M measure7 in a form of macro-average of average of areas over all pairs of classes:

$$M = \frac{2}{C*(C-1)} \sum_{i=1}^{C} AUCc_i * PRc_i \quad (4)$$

where
M - multi-class generalization of the AUC
C – number of classes;
AUC – area under the curve over the pairs of classes;
PR – prevalence.

The next derivates of contingency tables were used in this study. The Youden index (J) 8-10 which is also known as the Bookmaker Informedness (BM)2 evaluates the discriminative power of a dichotomous diagnostic test. The formula of Youden's index combines the sensitivity and specificity. It ranged from 0 to 1. High values of J indicate good performance of a classifier.11, 12

$$\text{Youden index } (J) = \frac{ad - bc}{(a+b)(c+d)} \quad (5)$$

$$= \frac{TP * TN - FN * FP}{(TP + FN)(FP + TN)} = Se + Sp - 1$$

where
a, TP – true positives;
b, FN – false negatives;
c, FP – false positives;
d, TN – true negatives;
Se - sensitivity;
Sp – specificity.

The difference between two classification algorithms was evaluated by means of the standard errors of the Youden indexes. [8]

$$S.E._J = \sqrt{\frac{a*b}{(a+b)^3} + \frac{c*d}{(c+d)^3}} \quad (6)$$

$$= \sqrt{\frac{TP*FN}{(TP+FN)^3} + \frac{FP*TN}{(FP+TN)^3}}$$

where
S.E.J – standard error of Youden index
a, TP – true positives;
b, FN – false negatives;
c, FP – false positives;
d, TN – true negatives.

Pearson's approximating χ^2 test compares categorical information against what one would expect based on the Chi-Squared Distribution [9, 13, 14] The Chi-Squared statistic was used in the minimum P-value approach in finding the optimal cutoff point of the ROC curve. [10, 15]

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} + E_{ij})^2}{E_{ij}} \quad (7)$$

where
χ^2 - Pearson's Chi-Square;
Oij – observed number elements in cell (i,j);
Eij – expected number elements in cell (i,j);
r – number of rows in confusion table;
c – number of columns in confusion table.

Sensitivity (recall) assesses the effectiveness of the classifier on the positive/minority class while specificity assesses the classifier's effectiveness on the negative/majority class.11, 12

$$Se = \frac{TP}{TP + FN} = \frac{TP}{Op} \quad (8)$$

where
Se - sensitivity;
TP – true positives;
FN – false negatives;
Op – observed positives;

$$Sp = \frac{TN}{TN + FP} = \frac{TN}{On} \quad (9)$$

where
Sp - specificity;
TN – true negatives;
FP – false positives;
On – observed negatives.

The most commonly reported measure of a classifier is its accuracy. This measure evaluates the overall efficiency of an algorithm.[11, 12]

$$Accuracy = \frac{TP + TN}{N} \quad (10)$$

where
PPV - positive prediction value, precision;
TP – true positives;
TN – true negatives;
N – total number of cases.

Optimized Precision is a type of hybrid threshold metric. This metric is a combination of accuracy, sensitivity, and specificity metrics[16]

$$OP = acc - \frac{|sp - sn|}{sp + sn} \quad (11)$$

where
OP - optimized precision;
acc - the accuracy score;
sp - specificity score;
sn - sensitivity score.

Other most relevant measures of classification of imbalanced datasets are the following.
The geometric mean (GM) metric aggregates both sensitivity and specificity measures and can be used with imbalanced datasets.17 A low GM signifies a low performance in the classification of positive cases.[12]

$$GM = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}} \quad (12)$$

where
GM – geometric mean;
TP – true positives;
FN – false negatives;
FP – false positives;
TN – true negatives.

Matthews correlation coefficient (MCC): this metric represents the correlation between the observed and predicted classifications and is least influenced by imbalanced data. A coefficient of +1 indicates a perfect prediction, -1 represents incongruity between true values and prediction; and 0 means no better than random distribution.[17, 18]

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FP)}} \quad (13)$$

where
MCC – Matthews correlation coefficient;
TP – true positives;
FN – false negatives;
FP – false positives;
TN – true negatives.

References

1. Amin A, Anwar S, Adnan A, et al. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. . IEEE Access. 2016: 7940-57.
2. Luque A, Carrasco A, Martín A and de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019; 91: 216-31.
3. Sun Y, Wong A and Kamel MS. Classification of imbalanced data: a review. International Journal of Pattern Recognition and Artificial Intelligence. 2011; 23.
4. Davis J and Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. 2006, p. 233-40.
5. Provost F and Domingos P. Well-trained PETs: Improving probability estimation trees, CeDER Working Paper #IS-00-04. Stern School of Business, New York University, NY, NY 10012, 2001.
6. Van Calster B, Van Belle V, Condous G, Bourne T, Timmerman D and Huffel S. Multi-class AUC metrics and weighted alternatives. 2008, p.1390-6.
7. Hand DJ and Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning. 2001; 45: 171-86.
8. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3: 32-5.
9. Powers D and Ailab. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol. 2011; 2: 2229-3981.
10. Unal I. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. Computational and Mathematical Methods in Medicine. 2017; 2017: 14.
11. Sokolova M, Japkowicz N and Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. 2006, p.1015-21.
12. Akosa JS. Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data. Proceedings of the SAS Global Forum. 2017.
13. Jackson SL. Research methods and statistics : a critical thinking approach. 2016.

14. Balakrishnan N, Voinov V and Nikulin MS. Chi-Squared Goodness of Fit Tests with Applications. Boston: Academic Press, 2013, p.229.
15. Rota M and Antolini L. Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers. Computational Statistics & Data Analysis. 2014; 69: 1-14.
16. Hossin M and M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 2015; 5: 01-11.
17. Tharwat A. Classification assessment methods. Applied Computing and Informatics. 2018.
18. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975; 405: 442-51.

Table 2B: Youden index average value. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3T	90	0,496365	0,071093	90	0,456	0,027
gr4_2T	90	0,797632	0,04782	90	0,478	0,020
gr4_1T	90	0,948888	0,037448	90	0,497	0,009
gr3_2T	90	0,544237	0,087575	81	0,189	0,249
gr3_1T	90	0,816009	0,068496	81	0,389	0,209
gr2_1T	90	0,49889	0,167617	72	0,333	0,336
macro-avg ± std 0,684 ± 0,195 0,390 ± 0,116 T-test p = 0, 006423						

Table 3B: Youden index relative bias. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3B	90	-0,0053	0,1325	90	-0,0055	0,1473
gr4_2B	90	0,0155	0,0648	90	-0,0093	0,1397
gr4_1B	90	0,0225	0,0327	90	0,0010	0,0943
gr3_2B	90	-0,0059	0,1234	73	-0,1518	0,5830
gr3_1B	90	-0,0061	0,0630	73	-0,2412	0,8977
gr2_1B	89	-0,0469	0,3805	55	0,5878	1,4025
macro-avg ± std -0,0044 ± 0,0242 0,0302 ± 0,2903 T-test p = 0,794						

Table 4B: Youden index mean square error (MSE). 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3M	90	0,0109	0,0123	90	0,0015	0,0016
gr4_2M	90	0,0090	0,0106	90	0,0007	0,0009
gr4_1M	90	0,0064	0,0094	90	0,0002	0,0003
gr3_2M	90	0,0208	0,0282	73	0,1385	0,1575
gr3_1M	90	0,0150	0,0195	73	0,0984	0,1220
gr2_1M	90	0,0871	0,1432	56	0,2525	0,4015
macro-avg ± std 0,0249 ± 0,0309 0,0820 ± 0,1023 T-test p = 0,794						

Table 5B: Accuracy average value. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3T	90	0,8977	0,0827	90	0,5030	0,0002
gr4_2T	90	0,9676	0,0332	90	0,5031	0,0002
gr4_1T	90	0,9835	0,0136	90	0,5032	0,0001
gr3_2T	90	0,7827	0,0813	90	0,5001	0,0001
gr3_1T	90	0,8710	0,0449	90	0,5002	0,0001
gr2_1T	90	0,6389	0,1490	90	0,5001	0,0001
macro-avg ± std 0,857 ± 0,129 0,502 ± 0,002 T-test p = 0,001						

Table 6B: Accuracy relative bias. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3B	90	-0,0011	0,0074	90	0,0002	0,0142
gr4_2B	90	0,0006	0,0027	90	-0,0005	0,0160
gr4_1B	90	0,0004	0,0014	90	0,0002	0,0160
gr3_2B	90	0,0014	0,0085	90	-0,0007	0,0098
gr3_1B	90	0,0008	0,0060	90	0,0001	0,0111
gr2_1B	90	0,0006	0,0080	90	0,0008	0,0113
macro-avg ± std 0,0004 ± 0,0008 0,000005±0,00055 T-test p = 0,407						

Table 7B: Accuracy mean square error (MSE). 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3M	90	0,000042	0,000061	90	0,000051	0,000068
gr4_2M	90	0,000007	0,000013	90	0,000065	0,000075
gr4_1M	90	0,000002	0,000004	90	0,000064	0,000069
gr3_2M	90	0,000044	0,000062	90	0,000024	0,000038
gr3_1M	90	0,000027	0,000031	90	0,000031	0,000037
gr2_1M	90	0,000023	0,000028	90	0,000032	0,000036
macro-avg ± std 2,4102E-05± 1,73668E-05 0,000044 ± 1,79372E-05 T-test p = 0,190						

Table 8B: Specificity (Sp) average value. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3T	90	0,9035	0,0864	90	0,5015	0,0001
gr4_2T	90	0,9680	0,0335	90	0,5015	0,0001
gr4_1T	90	0,9835	0,0136	90	0,5016	0,0001
gr3_2T	90	0,7826	0,0816	90	0,5001	0,0001
gr3_1T	90	0,8709	0,0450	90	0,5001	0,0001
gr2_1T	90	0,6389	0,1491	90	0,5000	0,0000
macro-avg ± std 0,8579 ± 0,12940,5008 ± 0,0008 T-test p = 0,001						

Table 9B: Specificity (Sp) relative bias. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3B	90	-0,0011	0,0074	90	0,0002	0,0142
gr4_2B	90	0,0006	0,0027	90	-0,0005	0,0160
gr4_1B	90	0,0004	0,0014	90	0,0002	0,0160
gr3_2B	90	0,0014	0,0085	90	-0,0007	0,0098
gr3_1B	90	0,0008	0,0060	90	0,0001	0,0111
gr2_1B	90	0,0006	0,0080	90	0,0008	0,0113
macro-avg ± std 0,0004 ± 0,0008 0,000003 ± 0,0005 T-test p = 0,468						

Table 10B: Specificity (Sp) mean square error (MSE). 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3M	90	3,797E-05	5,987E-05	90	5,188E-05	7,014E-05
gr4_2M	90	6,885E-06	1,314E-05	90	6,608E-05	7,642E-05
gr4_1M	90	2,004E-06	3,568E-06	90	6,536E-05	6,941E-05
gr3_2M	90	4,340E-05	6,146E-05	90	2,397E-05	3,802E-05
gr3_1M	90	2,769E-05	3,106E-05	90	3,110E-05	3,732E-05
gr2_1M	90	2,345E-05	2,803E-05	90	3,176E-05	3,560E-05
macro-avg ± std 2,357E-05 ± 1,650E-05 4,503E-05 ± 1,853E-05 T-test p = 0,171						

Table 11B: Optimized precision (OP) average value. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3T	90	0,6691	0,0728	90	0,1922	0,0127
gr4_2T	90	0,8861	0,0304	90	0,1819	0,0090
gr4_1T	90	0,9638	0,0167	90	0,1731	0,0038
gr3_2T	90	0,6938	0,1080	81	0,2674	0,0826
gr3_1T	90	0,8067	0,0775	81	0,2410	0,1393
gr2_1T	90	0,4239	0,2679	72	0,1073	0,2395
macro-avg ± std 0,7406 ± 0,1913 0,1938 ± 0,0560 T-test p = 0,0006						

Table 12B: Optimized precision (OP) relative bias. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3B	90	-0,0053	0,1325	90	-0,0055	0,1473
gr4_2B	90	0,0155	0,0648	90	-0,0093	0,1397
gr4_1B	90	0,0225	0,0327	90	0,0010	0,0943
gr3_2B	90	-0,0059	0,1234	73	-0,1518	0,5830
gr3_1B	90	-0,0061	0,0630	73	-0,2412	0,8977
gr2_1B	90	-0,0469	0,3805	55	0,5878	1,4025
macro-avg ± std -0,0044 ± 0,0242 0,0302 ± 0,2903 T-test p = 0,794						

Table 13B: Optimized precision (OP) mean square error (MSE). 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3M	90	0,0076	0,0091	90	0,0008	0,0010
gr4_2M	90	0,0034	0,0046	90	0,0006	0,0007
gr4_1M	90	0,0015	0,0028	90	0,0003	0,0003
gr3_2M	90	0,0076	0,0147	73	0,0170	0,0203
gr3_1M	90	0,0025	0,0046	73	0,0447	0,0553
gr2_1M	90	0,0278	0,1023	56	0,1293	0,2133
macro-avg ± std 0,0084 ± 0,0098 0,0321 ± 0,0506 T-test p = 0,226						

Table 14B: Geometric mean (GM) average value. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3T	90	0,724	0,052	90	0,692	0,010
gr4_2T	90	0,895	0,027	90	0,700	0,007
gr4_1T	90	0,974	0,019	90	0,707	0,003
gr3_2T	90	0,768	0,045	81	0,575	0,116
gr3_1T	90	0,905	0,036	81	0,661	0,087
gr2_1T	90	0,726	0,098	72	0,603	0,233
macro-avg ± std 0,832 ± 0,106 0,656 ± 0,055 T-test p = 0,004						

Table 15B: Geometric mean (GM) relative bias. 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3B	90	0,0056	0,0930	90	0,0000	0,0197
gr4_2B	90	0,0131	0,0585	90	0,0002	0,0139
gr4_1B	90	0,0271	0,0328	90	0,0001	0,0102
gr3_2B	90	0,0022	0,1017	73	-0,0523	0,3666
gr3_1B	90	0,0371	0,0585	73	-0,0284	0,2254
gr2_1B	90	0,1034	0,2041	49	0,1384	0,3692
macro-avg ± std 0,0314 ± 0,0376 0,0097 ± 0,0666 T-test p = 0,202						

Table 16B: Optimized precision (OP) mean square error (MSE). 10-folds cross-validation procedure

	N	Mean	Std. deviation	N	Mean	Std. Deviation
gr4_3M	90	0,0044	0,0052	90	0,0002	0,0002
gr4_2M	90	0,0028	0,0034	90	0,0001	0,0001
gr4_1M	90	0,0017	0,0026	90	0,0001	0,0001
gr3_2M	90	0,0058	0,0087	73	0,0303	0,0358
gr3_1M	90	0,0039	0,0051	73	0,0168	0,0208
gr2_1M	90	0,0310	0,0901	56	0,1222	0,2077
macro-avg ± std 0,0083 ± 0,0112 0,0283 ± 0,0476 T-test p = 0,239						