Evaluating Performance and Interpretability of Machine Learning Methods for Predicting Delirium in Gerontopsychiatric Patients

Michael NETZER[a,1], Werner O. HACKL[a], Michael SCHALLER[a], Lisa ALBER[b], Josef MARKSTEINER[c] and Elske AMMENWERTH[a]

[a] *Institute of Medical Informatics, Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria*

[b] *Nursing Management Board, Tyrolean Federal Hospitals, Innsbruck, Austria*

[c] *Department of Psychiatry and Psychotherapy A, State Hospital Hall, Hall in Tirol, Austria*

Abstract. Delirium is an acute mental disturbance that particularly occurs during hospital stay. Current clinical assessment instruments include the Delirium Observation Screening Scale (DOSS) or the Confusion Assessment Method (CAM). The aim of this work is to analyze the performance of machine learning approaches to detect delirium based on DOSS and CAM information obtained from two geropsychiatric wards in Tyrol. From a machine learning perspective, the questions of these two assessment instruments represent the features and the ICD 10 diagnoses of delirium (yes/no) is the corresponding class variable. We compare seven popular classification methods and analyze the performance and interpretability of the learning models. As our dataset is highly imbalanced, we also evaluate the effect of common sampling methods including down- and up-sampling methods as well as hybrid methods. Our results indicate a high predictive ability of advanced methods such as Random Forest that can handle even unbalanced datasets. Overall, combining a good performance of a prediction model with the ability of users to understand the prediction is challenging. However, for clinical application in fully electronic settings, a good performance seems to be more important than an easy interpretation of the prediction by the user. On the other hand, explanations of decisions are often needed to assess other criteria such as safety.

Keywords. delirium, machine learning, statistics, computerized medical record

1. Introduction

Delirium is characterized by a temporary change (usually hours to days) in mental status and is often confused with depression, dementia and psychosis [1]. In a clinical setting delirium is an important patient safety indicator [2].

Popular instruments for classifying delirium include Delirium Observation Screening Scale (DOSS) or the Confusion Assessment Method (CAM). These

[1] Corresponding Author: Michael Netzer, Private University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria, Address, E-Mail: michael.netzer@umit.at

instruments are commonly accepted in clinical practice [3]. Recently, machine learning methods have been used to predict delirium [4, 5]. In these approaches, researchers commonly compare the predictive performance of different learning models such as logistic regression, support vector machines or neural networks. Very recently, Corradi et al. [6] used a Random Forest machine learning algorithm to predict delirium based on Confusion Assessment Method (CAM) and Electronic Health Record (EHR) data. The authors also used an under-sampling strategy to address the class imbalance. Also, Jauk et al. [7] very recently used a machine learning approach to predict delirium evaluating the impact of missing data.

The aim of our work is to analyze the predictive ability of DOSS and CAM compared to popular machine learning approaches. In contrast to previous works we only use CAM and DOSS data and we use different machine learning models and methods for dealing with imbalanced data, with the aim to improve the performance of methods. The analysis of the interpretability of the models also is a further new aspect of our approach.

2. Material and Methods

2.1. Dataset

CAM and DOSS data were scored and entered into the Electronic Health Record (EHR) by qualified nurses (DOSS) and physicians (CAM) in two geropsychiatric wards in Tyrol between 03.2019 and 10.2019. The questions of CAM and DOSS represent features and the ICD 10 diagnoses of delirium (yes/no) are the corresponding class variables of our two datasets. For instance, the first two CAM features represent yes/no questions evaluating acute onset and fluctuating course: 1a) “Is there evidence of an acute change in mental status from the patient’s baseline?” and 1b) “Did the (abnormal) behavior fluctuate during the day, that is tend to come and go or increase and decrease in severity?” [8].

The documentation guideline stipulated to do the first delirium assessments at admission (within 48 hours at latest) and to repeat them on a weekly basis. In addition to the delirium assessments, information on the diagnosis of delirium (according to ICD 10 criteria) was documented by physicians for all included patients as gold standard (delirium YES/NO).

The total number of included patients was 318 with 370 distinct stays. All CAM, DOSS and ICD data were retrospectively extracted, combined to CAM-DOSS-ICD triples according to documentation date (all three had to be documented in a time window of maximum 24 hours) and anonymized. The total number of triples was 1,320, comprising 1,271 controls (without delirium) and 49 cases (resulting from a sample of n=20 distinct patients with diagnosed delirium) according to the gold standard. Table 1 summarizes the considered datasets including number of features and instances.

Ethical approval of this study was granted by the Ethical Committee of the Medical University of Innsbruck (EK Nr: 1032/2019).

Table 1. Description of the considered datasets including number of features (without class variable) and samples.

Dataset	Description	Number of features	Number of instances
DS_{DOSS}	Collection of completed questionnaires using DOSS method	14	1320
DS_{CAM}	Collection of completed questionnaires using CAM method	5	1320

2.2. Machine Learning Approach

We implemented the machine learning methods using the programming language R and *caret* package [9]. We used seven popular classification approaches that are summarized in Table 2, using a categorization based on Witten et al. [10]. The optimization of model parameters was based on caret default tuning parameter grid (see Table 3).

We used a 10-fold cross-validation strategy to estimate the generalization performance of the learning models. More specifically, we divided the dataset into 10 partitions using 9 parts for training the model and using the remaining part for testing.

We evaluated the final performance for predicting delirium using averaged Kappa statistic as evaluation measure. In contrast to accuracy, this measure is also a suitable measure if a skew class distribution is present [9]. Kappa is defined by

$$\kappa = \frac{O-E}{1-E} \tag{1}$$

where O is the observed agreement and E is the expected agreement under independence [11]. Thus, in simple words, Kappa shows whether a chosen classification approach was able to predict delirium. According to McHugh [12] values between 0.21–0.40 are considered as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect.

In summary, model tuning and evaluation is depicted in the following pseudocode (see also *caret* package documentation):

```
for each parameter set do
    for each fold do
            determine hold-out samples
            fit model on the remainder
            predict hold-out samples
    end
    calculate average kappa across hold-out predictions
end
determine optimal parameter set
```

Imbalanced class distribution can also have a negative impact on model fitting. Sampling methods address this problem by either selecting a subset of the majority class (down-sampling), replicate the minority class (up-sampling) or using a hybrid procedure. Consequently, we also implemented down- and up-sampling and SMOTE which is a hybrid procedure of up- and down-sampling and is described in [13].

Table 2. Categorization of classification approaches used in our study

Method	Category	Description
k-Nearest-Neighbor	Lazy	Classify instances based on the label of k-nearest neighbors
Naïve Bayes	Bayes	Probabilistic classifier assuming independence between features
RPART	Trees	Recursive Partitioning and Regression Trees
Logistic Regression	Function	Generalized linear model using a logistic link function
Support Vector Machine	Function	Establishes a maximum-margin hyperplane
Multi-Layer Perceptron	Function	Artificial neural network
Random Forest	Ensemble	Creates a set of decision trees in each iteration of a bagging algorithm

Table 3. Tuning parameters for the considered classifier using caret R package [9].

Method	Tuning Parameters
k-Nearest-Neighbor	Number of neighbors
Naïve Bayes	Laplace correction, distribution type, bandwidth adjustment
RPART	Complexity parameter
Logistic Regression	No tuning parameters
Support Vector Machine	Cost (using a linear kernel)
Multi-Layer Perceptron	Number of hidden units, weight decay
Random Forest	Number of randomly selected predictors

3. Results

The Kappa values and corresponding sensitivity and specificity values for the considered datasets are depicted in Table 4.

Table 4. Kappa and corresponding sensitivity and specificity values for the current DOSS and CAM classification scheme.

Dataset	Kappa	Sensitivity	Specificity
$\mathbf{DS_{DOSS}}$	0.098	0.878	0.648
$\mathbf{DS_{CAM}}$	0.612	0.959	0.958

Table 5 summarizes the Kappa values for all considered classifiers and sampling strategies.

The highest Kappa values for the DOSS dataset was obtained using Random Forest without sampling (Kappa = 0.29). The highest Kappa values for the CAM dataset were obtained using k-Nearest-Neighbor, Logistic Regression and Random Forest without sampling (Kappa = 0.65). The ROC curves for the two Random Forest models and the corresponding area under the curve (AUC) values are depicted in Figure 1. The resulting RPART decision tree for predicting delirium using DS_{CAM} dataset is depicted in Figure 2. The decisions are built on variables “cam_1a” and “cam_1b”.

Table 5. Kappa values for the applied classifiers for the considered sampling strategy (i.e., none, up, down and SMOTE sampling).

Method	Dataset	Kappa for sampling strategy			
		None	**Up**	**Down**	**SMOTE**
k-Nearest-Neighbor	DS_{DOSS}	0.051	0.108	0.115	0.102
	DS_{CAM}	**0.648**	0.525	0.471	0.528
Naïve Bayes	DS_{DOSS}	0.000	0.150	0.118	0.188
	DS_{CAM}	0.446	0.408	0.374	0.486
RPART	DS_{DOSS}	0.021	0.128	0.083	0.131
	DS_{CAM}	0.534	0.373	0.327	0.404
Logistic Regression	DS_{DOSS}	0.063	0.126	0.082	0.097
	DS_{CAM}	**0.648**	0.417	0.359	0.535
Support Vector Machine	DS_{DOSS}	0.000	0.116	0.085	0.076
	DS_{CAM}	0.616	0.539	0.433	0.531
Multi-Layer Perceptron	DS_{DOSS}	0.174	0.151	0.088	0.098
	DS_{CAM}	0.581	0.542	0.542	0.547
Random Forest	DS_{DOSS}	**0.289**	0.228	0.098	0.157
	DS_{CAM}	**0.645**	0.569	0.513	0.544

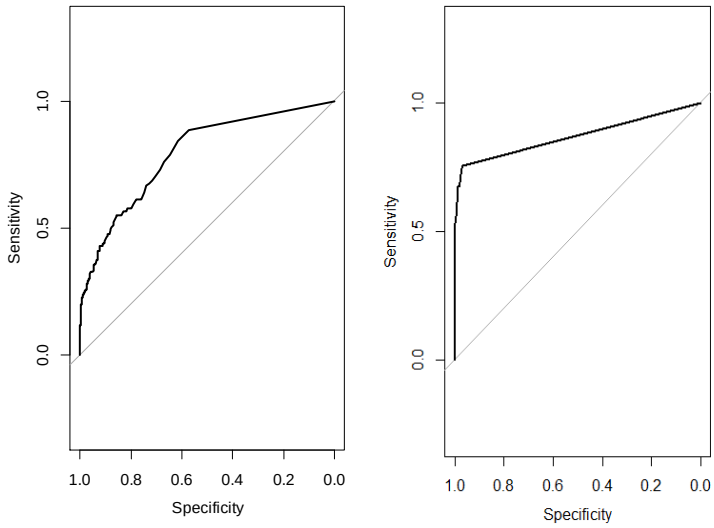

Figure 1. ROC curves for the Random Forest classifiers with highest kappa for DS_{DOSS} (left, AUC= 0.794) and DS_{CAM} (right, AUC= 0.869) dataset.

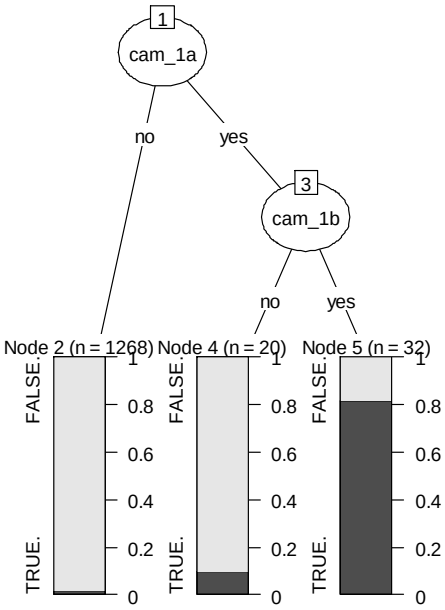

Figure 2. RPART decision tree for the CAM dataset. The decision is based solely on variables "cam_1a" (first decision) and "cam_1b" (second decision if cam_1a is TRUE). The corresponding questions 1a and 1b can be found in section 2.1.

4. Discussion

We applied popular machine learning approaches to predict the clinical diagnosis of delirium using DOSS and CAM data only and considering different sampling strategies. Our machine learning approach using DS_{DOSS} clearly outperformed the current DOSS classification procedure in terms of predictive ability (Kappa of 0.289 vs. 0.098).
In contrast, the predictive ability using DS_{CAM} was only slightly improved using a machine learning approach (Kappa of 0.648 vs. 0.612).

The results indicate a good performance of the Random Forest classifier (highest Kappa values for DS_{DOSS} and DS_{CAM} respectively) even without using a sampling strategy. This can be explained due to the robustness of this classifier to imbalanced data [14]. In contrast, methods such as Naïve Bayes require sampling methods to generate a valid model from the imbalanced data.

One general challenge of machine learning is whether users (for example physicians) are able to understand and interpret the predictions of a given model. Typically, we consider that decision trees can be easier understood and interpreted by physicians. From the approaches we tested, Recursive Partitioning and Regression Trees (RPART) use decision trees (see Table 2). RPART and comparable approaches are an extension of rule-based methods. The class of a newly unlabeled sample can be easily determined by traversing the tree from the root to the particular leaf that represents the class. Scientists can also easily interpret the importance of variables of a tree model

because in general variables with higher importance are located in the upper part of the tree.

Random Forest also uses decision trees and showed a high performance in our study. However, due their architecture using bootstrapping the resulting models are highly complex and interpretation and manual application by physicians is difficult. Overall, combining a good performance of a prediction model with the ability of users to understand the prediction is challenging.

However, we may argue that understanding a prediction may not be that necessary anymore. The prediction algorithms implemented in the original CAM and DOSS instruments may also not be known in detail by the users. For clinical application in fully electronic settings with automated prediction approaches, a good performance seems to be more important than an easy interpretation of the prediction by the user. In this case, a physician may only use the prediction result (e.g., delirium risk or not) obtained from a model with a high predictive ability (such as Random Forest in our case). In contrast, clear explanations of decisions are often needed to assess other criteria such as safety [15].

Currently we used only features based on DOSS and CAM, which have to be documented manually. The usage of clinical routine data such as laboratory data and other available EHR data including nursing process data is part of our ongoing work. Here we want to test if it is possible to reach similar predictive performance by only using available data without DOSS and CAM information. But also combining DOSS and CAM information together may improve the predictive performance which has to be proven in further studies. For that we can rely on the solid gold standard which had been established in this work.

Disclosure

This work forms part of the PATIS Project (Patient Safety Intelligence System and framework for the secondary use of multimodal clinical data to assess and improve patient safety) funded by the Austrian Science Fund (FWF – P 29076).

Acknowledgement

We thank all participating physicians, nurses and IT department staff members for their help and especially Manuel Kern and Stephanos Rauchegger from the nursing management board for their help in data collection and extraction.

References

[1] O. C. Gleason, “Delirium,” *Am. Fam. Physician*, vol. 67, no. 5, pp. 1027–1034, 2003.

[2] S. K. Inouye, R. G. J. Westendorp, and J. S. Saczynski, “Delirium in elderly people,” *Lancet*, vol. 383, no. 9920, pp. 911–922, 2014, doi: 10.1016/S0140-6736(13)60688-1.

[3] W. Hasemann, R. W. Kressig, D. Ermini-Fünfschilling, M. Pretto, and R. Spirig, “Screening, Assessment und Diagnostik von Delirien,” *Pflege*, 2007, doi: 10.1024/1012-5302.20.4.191.

[4] A. Wong, A. T. Young, A. S. Liang, R. Gonzales, V. C. Douglas, and D. Hadley, “Development and Validation of an Electronic Health Record-Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment,” *JAMA Netw. open*, vol. 1, no. 4, p.

e181018, Aug. 2018, doi: 10.1001/jamanetworkopen.2018.1018.
[5] S. P. K. Veeranki *et al.*, "An improvised classification model for predicting delirium," in *Studies in Health Technology and Informatics*, 2019, doi: 10.3233/SHTI190537.
[6] J. P. Corradi, S. Thompson, J. F. Mather, C. M. Waszynski, and R. S. Dicks, "Prediction of Incident Delirium Using a Random Forest classifier," *J. Med. Syst.*, 2018, doi: 10.1007/s10916-018-1109-0.
[7] S. Jauk *et al.*, "Information adapted machine learning models for prediction in clinical workflow," in *Studies in Health Technology and Informatics*, 2019, vol. 260, pp. 65–72, doi: 10.3233/978-1-61499-971-3-65.
[8] S. K. Inouye, "The Confusion Assessment Method (CAM): Training Manual and Coding Guide," *Bost. Hosp. Elder Life Progr.*, 2003.
[9] M. Kuhn, "Building Predictive Models in R Using the caret Package," *J. Stat. Softw.*, 2008.
[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. 2016.
[11] N. Moradzadeh, M. Ganjali, and T. Baghfalaki, "Weighted kappa as a function of unweighted kappas," *Commun. Stat. Simul. Comput.*, 2017, doi: 10.1080/03610918.2015.1105975.
[12] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012, doi: 10.11613/bm.2012.031.
[13] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, 2013, doi: 10.1186/1471-2105-14-106.
[14] D. J. Dittman, T. M. Khoshgoftaar, and A. Napolitano, "Is data sampling required when using random forest for classification on imbalanced bioinformatics data?," in *Advances in Intelligent Systems and Computing*, 2016.
[15] F. Doshi-Velez and B. Kim, "Considerations for Evaluation and Generalization in Interpretable Machine Learning," 2018.